# Proyecto BI Analítica de Textos

Juan David Cardona

Juan Sebastian Villamil

Nicolás Calero

* + **Comprensión del negocio y enfoque analítico**.

Los objetivos del proyecto son entonces lograr clasificar los comentarios de las publicaciones de un sector de mercado, con el fin de conocer los sentimientos del usuario frente a un producto. El criterio de éxito será un umbral de clasificación que supere el 70% de precisión. Nótese que se utiliza la precisión como métrica debido a que se concluye que esta es más valiosa en el contexto del negocio, porque el costo de un falso positivo es mayor.

Las tareas de analítica de texto necesarias para alcanzar el objetivo de predicción serán la limpieza de datos, el entrenamiento del modelo y finalmente la prueba de la efectividad del modelo. La limpieza de datos va a consistir en eliminar duplicados y nulos, en cuanto a la analítica de texto se va a tener que reconocer las palabras, eliminar pronombres y distinguir las palabras claves para lograr procesar el comentario y tener una predicción exitosa.

|  |  |  |
| --- | --- | --- |
| Oportunidad/problema  Negocio | La oportunidad de negocio es filtrar los comentarios por publicaciones referentes a cierto comercio o sector de negocios es posible hacerse a una idea de cómo se encuentra la opinión publica al respecto, lo cual puede ser significativamente útil para el negocio. | |
| Descripción del requerimiento desde el punto de vista de aprendizaje de máquina | La máquina deberá predecir a que tipo de sentimiento pertenece un comentario, así se podrá saber que opina el cliente o el usuario sobre la publicación de interés para el negocio | |
| **Detalles de la actividad de minería de datos** | | |
| Tarea | Técnica | Algoritmo e hiper-parámetros utilizados (con la justificación respectiva) |
| Predecir el sentimiento del usuario de redes sociales según el comentario publicado | Analítica de Textos con Machine Learning | Naive Bayes |

* + **Comprensión de los datos y preparación de los datos**.

Para el caso a desarrollar, se tienen dos conceptos importantes. El primero es el comentario, en donde este es una secuencia de palabras que el cliente publica refiriéndose a algún producto. El segundo concepto es la emoción, esta es el sentimiento al que se resume la persona al momento de realizar el comentario. Ahora bien, habiendo entendido las variables relevantes en el caso, se procede con la preparación de los datos. Primero, se verifica la no existencia de datos nulos:

Imagen que contiene Tabla

Descripción generada automáticamente

Una vez hecho esto, se procede a asignarle un valor numérico a cada emoción, para mejor tratabilidad.

Texto

Descripción generada automáticamente

Como se puede evidenciar, a la felicidad se le asigna el valor 1, a la tristeza el valor 2, al enojo el valor 3, y así hasta llegar a un sexto valor. Se crea entonces una tabla para visualizar la frecuencia de cada una de las emociones.

Gráfico, Gráfico de barras

Descripción generada automáticamente

A continuación, se emplea la asistencia del paquete *sentiment package* de la plataforma *Natural Language Toolkit* (NLTK) con el fin de obtener más información acerca de los comentarios. NLTK permite el procesamiento de información de lenguaje humano, y ofrece herramientas para clasificación, tokenizacion, derivación de información, entre otros. En el caso actual, el paquete de *sentiment package* permite el análisis de los sentimientos que se encuentran en cadenas de texto. Esto lo permite mediante varios tipos de analizadores, y los que se usaran en este proyecto son el analizador de sentimientos positivos, el analizador de sentimientos negativos, el analizador de sentimientos neutrales y finalmente el analizador compuesto de sentimientos. Se aplican estos analizadores a los datos provistos, como se evidencia a continuación:

Tabla

Descripción generada automáticamente

Como se puede observar, estos analizadores retornan un puntaje para cada comentario de los datos en el dataframe. Además de esto, se añade información acerca de la polaridad y la subjetividad de cada uno de los comentarios.

Tabla

Descripción generada automáticamente

Se procede a verificar que las columnas de información que se agregaron sean de los tipos de datos requeridos.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Finalmente, se procede con la normalización de los datos, con el fin de evitar errores o perturbaciones en el modelo por causa de diferencias de dominios.

Tabla

Descripción generada automáticamente

Como se evidencia en la tabla, los datos son normalizados exitosamente. Con esto se concluye la preparación de los datos.

* + **Modelado y evaluación**.

**MODELO MULTI CLASS DE CLASIFICACIÓN ONE VS ONE BASADO EN SUPPORT VECTOR MACHINE (SVM)**

Para comenzar, se definen los conjuntos X y Y del modelo, en donde el conjunto y será la columna emoción, al ser la variable objetivo, y x será el dataframe sin el valor de la variable objetivo. Se particionan ahora estos conjuntos con el fin de tener una muestra para medir el desempeño.

Imagen que contiene Forma

Descripción generada automáticamente

Los hiperparametros a buscar son gamma y kernel. El gamma es el valor de la distancia a la cual se encuentran los puntos de una posible linea divisoria. Cuando Gamma es alto, se tendrá en cuenta la posición de todos los puntos y se puede incurrir en Overfitting porque el modelo se puede acostumbrar a estos valores para definir el hiperplano. Si Gamma es pequeño, solo se tendrá en cuenta la posición de los puntos cercanos a la línea divisoria, por lo cual el modelo será más flexible. En cuanto al kernel, este busca transformar los datos que se le pasan al algoritmo. De tal forma que si están distribuidos de una manera no lineal pueden ser pasados a un plano de más dimensiones lo cual ayudaría a encontrar una ecuación lineal que ayude al algoritmo SVM a clasificar de mejor manera los datos. Se efectúa la búsqueda de hiperparametros para este modelo, en donde se encuentra lo siguiente:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Como se evidencia en la imagen anterior, se encuentra que los hiperparametros adecuados para el conjunto de datos es utilizar un Gamma de escala, y una transformación kernel de RBF. Se efectua el entrenamiento para el modelo one vs one (ovo) y se obtienen una exactitud sobre el entrenamiento de 0.534 y una exactitud sobre los datos de prueba de 0.525. A continuación, se muestra la matriz de confusión para el modelo:

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Las métricas se encuentran en el siguiente reporte de clasificación:

Tabla

Descripción generada automáticamente

Se obtiene una precisión del 63%, sensibilidad del 28% y un puntaje F1 de 0.22.

**MODELO DE CLASIFICACIÓN GAUSSIAN NAIVE BAYES POR SENTIMIENTO (6)**

Ahora, se utilizará el modelo Gaussian Naive Bayes, pero esta vez se empleará el modelo para cada uno de los sentimientos. Para empezar, se crean 6 dataframes distintos, uno por cada sentimiento. Se entrena entonces el modelo para cada uno de los dataframes, en donde se obtienen los siguientes resultados:

*Sentimiento Joy:*

Para el sentimiento Joy, se obtiene una exactitud del 0.794 sobre el conjunto de entrenamiento, y una exactitud de 0.778 sobre el conjunto de prueba. La matriz de confusión es la siguiente:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Las métricas se encuentran en el siguiente reporte de clasificación:

**Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza media**

Se obtiene una precisión del 75%, sensibilidad del 77% y un puntaje F1 de 0.76. Esto ya representa una mejora frente al modelo anterior.

*Sentimiento Sadness:*

Para el sentimiento Sadness, se obtiene una exactitud del 0.723 sobre el conjunto de entrenamiento, y una exactitud de 0.721 sobre el conjunto de prueba. La matriz de confusión es la siguiente:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Las métricas se encuentran en el siguiente reporte de clasificación:

**Tabla

Descripción generada automáticamente**

Se obtiene una precisión del 65%, sensibilidad del 62% y un puntaje F1 de 0.63.

*Sentimiento Anger:*

Para el sentimiento Anger, se obtiene una exactitud del 0.867 sobre el conjunto de entrenamiento, y una exactitud de 0.859 sobre el conjunto de prueba. La matriz de confusión es la siguiente:

Gráfico

Descripción generada automáticamente

Las métricas se encuentran en el siguiente reporte de clasificación:

**Tabla

Descripción generada automáticamente**

Se obtiene una precisión del 93%, sensibilidad del 50% y un puntaje F1 de 0.46.

*Sentimiento Fear:*

Para el sentimiento Fear, se obtiene una exactitud del 0.877 sobre el conjunto de entrenamiento, y una exactitud de 0.887 sobre el conjunto de prueba. La matriz de confusión es la siguiente:

Gráfico

Descripción generada automáticamente

Las métricas se encuentran en el siguiente reporte de clasificación:

**Tabla

Descripción generada automáticamente**

Se obtiene una precisión del 94%, sensibilidad del 50% y un puntaje F1 de 0.47.

*Sentimiento Love:*

Para el sentimiento Love, se obtiene una exactitud del 0.921 sobre el conjunto de entrenamiento, y una exactitud de 0.91 sobre el conjunto de prueba. La matriz de confusión es la siguiente:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Las métricas se encuentran en el siguiente reporte de clasificación:

**Tabla

Descripción generada automáticamente**

Se obtiene una precisión del 96%, sensibilidad del 50% y un puntaje F1 de 0.48.

*Sentimiento Surprise:*

Para el sentimiento Surprise, se obtiene una exactitud del 0.964 sobre el conjunto de entrenamiento, y una exactitud de 0.964 sobre el conjunto de prueba. La matriz de confusión es la siguiente:

Gráfico

Descripción generada automáticamente

Las métricas se encuentran en el siguiente reporte de clasificación:

**Tabla

Descripción generada automáticamente**

Se obtiene una precisión del 98%, sensibilidad del 50% y un puntaje F1 de 0.49.

**MODELO MULTI CLASS DE CLASIFICACIÓN ONE VS REST BASADO EN ÁRBOLES**

Para el modelo basado en árboles, una vez más se particiona el conjunto en conjuntos para entreno y conjuntos para pruebas. También se efectúa una búsqueda de hiperparametros, y para esto se genera el siguiente espacio de búsqueda:



En donde se encuentra que los mejores valores de hiperparametros para los datos son utilizar el criterio de entropía, con una máxima profundidad de 7 y un numero mínimo de muestras de división de 5.

Utilizando este modelo con los hiperparametros obtenidos, se obtiene una exactitud de 0.572 sobre el conjunto de entrenamiento, y una de 0.516 sobre el conjunto de prueba. La matriz de confusión es la siguiente: Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Las métricas se encuentran en el siguiente reporte de clasificación:

**Tabla

Descripción generada automáticamente**

Se obtiene una precisión del 39%, sensibilidad del 30% y un puntaje F1 de 0.27 .

**MODELO DE CLASIFICACIÓN NAIVE BAYES NLTK**

Para el modelo de Naive Bayes utilizado mediante el paquete de NLTK, se obtiene una precisión del 75.69%.

Para el modelo, se obtienen las siguientes palabras como las más relevantes a la hora de clasificar el comentario:

Tabla

Descripción generada automáticamente

* + **Resultados**.

El tablero de control que se realiza para le proyecto es el siguiente:

Gráfico, Gráfico de barras

Descripción generada automáticamente

En la gráfica superior izquierda, se puede ver la cantidad de comentarios por sentimiento. En la inferior izquierda, una gráfica para ver el total de comentarios por sentimiento, así como la distribución de estos. Por último, en la gráfica de la derecha, se puede ver la existencia de los comentarios duplicados y cuales son.

Con el fin de ampliar el conocimiento que la organización tiene de sus clientes, así como cumplir el objetivo de lograr clasificar los comentarios de las publicaciones de un sector de mercado, y así conocer los sentimientos del usuario frente a un producto, las posibles estrategias que la organización debe plantear son la implementación de alguno de los modelos realizados en el proyecto. Específicamente, de los modelos analizados, se puede concluir que la mejor opción para la organización será implementar el modelo de clasificación Gaussian Naive Bayes por sentimiento. Esto se debe, en primer lugar, a que este fue el modelo del cual se obtuvieron las mejores métricas, en donde la mayoría de los sentimientos superan el 70% de precisión. Además de esto, este modelo permite asignarle a un comentario varias emociones, lo cual puede resultar de mayor utilidad a asignar una única emoción a un comentario.