Universidad Autónoma Metropolitana
División de Ciencias de la Comunicación y Diseño
Departamento De Tecnologías de la Información

Licenciatura en Tecnologías y Sistemas de Información

Prediciendo el impacto de una publicación de Facebook

Por:

Erika Sarai Rosas Quezada

Asesores:

Mtra. Adriana Gabriela Ramírez de la Rosa

Dr. Esaú Villatoro Tello

Noviembre 2019

Resumen

Hoy en día, Facebook se ha convertido para las empresas y figuras públicas el canal de comunicación y divulgación más importante para difundir sus productos y servicios. Sin embargo, las empresas también asumen que esta nueva forma de compartir información trae consigo nuevos retos, puesto que ahora las empresas deben tener cuidado con lo que publican en redes sociales. Es por ello, la necesidad de plantear estrategias precisas para anticipar el impacto de una publicación es fundamental para cualquier empresa o figura pública.

Por lo tanto, este trabajo tiene como objetivo crear una herramienta web que permita predecir el impacto que producirá una publicación de Facebook, a partir de emplear técnicas de aprendizaje computacional. Para lograr nuestro objetivo planteamos un modelo predictivo, el cual lo entrenamos con más de 37,000 publicaciones recopiladas de Facebook de 10 páginas públicas y de 14 políticos mexicanos. Realizamos experimentos y encontramos que el contenido y los atributos de comportamiento, estilo, interacción, popularidad y tiempo de las publicaciones proporcionan información relevante para nuestro modelo de predicción.

Palabras claves: análisis de impacto, extracción de características, clasificación de texto y aprendizaje computacional.

Índice

1 Introducción **5**

1.1 Objetivos 7

1.1.1 Objetivo General 7

1.1.2 Objetivos Específicos 7

1.2 Organización del documento 7

2 Marco Teórico **9**

2.1 Clasificación automática de Textos 9

2.1.1 Extracción de características 10

2.1.2 Etapa de clasificación 12

2.1.3 Etapa de evaluación 13

3 Trabajo relacionado **17**

3.1 Investigaciones existentes con sistemas predictivos 18

4 Método propuesto **23**

4.1 Colección de datos 24

4.1.1 Estadísticas de los datos 26

4.1.2 Metodología del etiquetado 31

4.2 Preprocesamiento 34

4.3 Extracción de atributos 35

4.4 Representación 36

4.5 Clasificación 37

5 Experimentos y resultados **39**

5.1 Resultados de predicción de impacto en las empresas 40
5.2 Resultados de predicción de impacto en los políticos 46
5.3 Conclusiones generales de las predicciones de las empresas y políticos . . 48

6 Desarrollo del sistema **53**
6.1 Módulos de carga 54
6.2 Módulo de extracción de características 55
6.3 Módulo de predicciones 55
6.4 Módulo de visualización 55
6.5 Vista del sistema 58
6.5.1 Predecir el impacto de una publicación paso a paso 59

7 Conclusiones y trabajo futuro **63**

8. Referencias **67**

Appendix A Anexos **71**
A.1 Resultados al predecir las métricas de las empresas 72
A.1.1 Resultados con el conjunto de experimentos **ExpA** 72
A.1.2 Resultados con el conjunto de experimentos **ExpB** 74
A.2 Resultados al predecir las métricas de los políticos 75
A.2.1 Resultados con el conjunto de experimentos **ExpA** 75
A.2.2 Resultados con el conjunto de experimentos **ExpB** 76

1

Introducción

Actualmente hay 7.6 mil millones de habitantes en el mundo, de los cuales 4.2 mil millones tienen acceso a Internet. De éstos últimos, 3.03 mil millones son usuarios activos de las redes sociales [1].

La principal red social a la que se conectan los usuarios es a Facebook con 2.41 mil millones de usuarios activos mensuales a partir del segundo trimestre de 2019 [1]. En consecuencia Facebook se ha declarado la red social más grande del mundo según las estadísticas de la figura 1.1

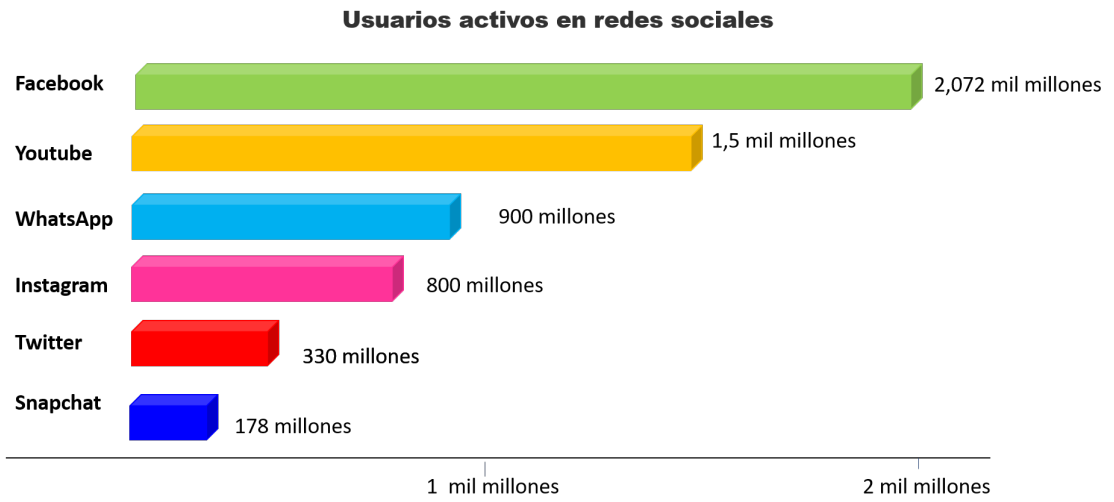

Figura 1.1: Usuarios activos en redes sociales. [1]

Dada la cantidad de usuarios que contiene Facebook, según los resultados de la

figura 1.1 . Facebook se ha convertido para las empresas y figuras públicas en el canal de comunicación y divulgación más importante para difundir sus productos y servicios. Sin embargo, las empresas también asumen que esta nueva forma de compartir información trae consigo nuevos retos, puesto que ahora las empresas deben tener cuidado con lo que publican en redes sociales. Ya que hoy, los clientes tienden a convertirse en potenciales influyentes a través de sus recomendaciones, opiniones o quejas.

Por ende, el marketing de redes sociales requiere incorporar constantes estrategias que permitan gestionar de manera eficaz las metas y objetivos de cada empresa [2]. Especialmente los objetivos relacionados con la participación, relación y comunicación de los clientes [2]. Por lo tanto, medir el impacto que produce una publicación es un tema importante que las empresas deben de incluir como parte de sus estrategias de marketing [3].

De acuerdo con investigaciones anteriores, el impacto de una publicación se mide a través de varias métricas disponibles, por ejemplo: las reacciones (me gusta, me encanta, me divierte, me entristece, me asombra, me enoja), los comentarios y las interacciones que genera el cliente en una publicación son algunas de las métricas que se consideran útiles. Por otro lado, se dice que aumentar el impacto de las publicaciones trae consigo relaciones más sólidas entre las empresas y los clientes, esto permite que el contenido que se genera sea más valioso en las redes sociales [4].

Recientemente, diferentes compañías de comercio electrónico, investigaciones de negocios e investigaciones de marketing digital, se han tomado a la tarea de estudiar las relaciones entre las publicaciones de las redes sociales y el impacto que generan. Las investigaciones que estas compañías realizan contienen un carácter de análisis a posteriori [2, 5, 6, 7, 8, 4]. En consiguiente, éstas se centran en encontrar aquellas características que permitan diseñar estrategias de marketing. Sin embargo, a pesar de todo el conocimiento que estos estudios poseen, no es suficiente para predecir el impacto que tendrá una publicación. Por lo tanto, un sistema capaz de anticipar el impacto de una publicación; puede traer grandes ventajas a la hora de decidir publicar algo en las plataformas de redes sociales [4].

A partir de lo anterior, el presente trabajo tiene como objetivo crear una aplicación que permita predecir el impacto (alto o bajo) que tendrá una publicación de Facebook. En contraste con los enfoques tradicionales, nuestro método propuesto está basado en una metodología de aprendizaje supervisado, que incorpora las características del contenido, el estilo y los atributos de comportamiento para representar una publicación. Adicionalmente, en este trabajo se presenta una validación del método de predicción

desarrollado. Para esto, se recopiló un conjunto de publicaciones de Facebook, de diez cuentas públicas de empresas y catorce cuentas públicas de políticos mexicanos que se postularon o tenían un cargo público en el año 2018.

Así pues, los experimentos realizados para predecir seis problemas de clasificación diferentes, indican que la combinación de las características propuestas con algunos atributos basados en metadatos, permiten que nuestra aplicación obtenga resultados de rendimiento aceptables.

1.1 | Objetivos

1.1.1 | Objetivo General

- Construir una herramienta automática que permita predecir el impacto que producirá una publicación en Facebook a partir de emplear técnicas de aprendizaje computacional.

1.1.2 | Objetivos Específicos

1. Identificar qué características del lenguaje escrito ayudan a detectar el impacto de una publicación de Facebook.

2. Definir el problema de clasificación basado en el análisis de los datos para generar un modelo que ayude a la predicción.

3. Evaluar mediante un esquema de aprendizaje supervisado la pertinencia de los atributos identificados para predecir el impacto de una publicación de Facebook.

4. Crear una herramienta que permita a un usuario predecir el impacto que tendrá una publicación de Facebook.

1.2 | Organización del documento

Este documento está formado por los siguientes capítulos:

Capítulo 2: Este capítulo describe brevemente el marco teórico, el cual está compuesto por conceptos esenciales para entender el resto del trabajo. Algunos de los

conceptos que se abordan en este capítulo son: aprendizaje automático, clasificación automática de textos, representación de textos y métricas de evaluación.

Capítulo 3: Este capítulo muestra algunos trabajos relacionados, los cuales describen el problema de las redes sociales y la gestión de las relaciones con los clientes. Así mismo, se abordan otros trabajos que han predicho el impacto de una publicación de alguna red social.

Capítulo 4: Este capítulo explica la metodología propuesta basado en un enfoque de aprendizaje supervisado. Sucesivamente en este capitulo se describe la recopilación del corpus, así también cómo se realizó el etiquetado, y algunas estadísticas sobre su composición.

Capítulo 5: Este capítulo describe la configuración experimental y los resultados obtenidos para todos los experimentos realizados.

Capítulo 6: Este capítulo describe el desarrollo de la aplicación web y la integración del sistema y finalmente, el capítulo 7: describe algunas conclusiones e ideas del trabajo futuro.

2

Marco Teórico

El presente capítulo tiene como objetivo dar a conocer los conceptos esenciales y los elementos teóricos que fundamentan a este trabajo. Es por ello, que esta sección muestra conceptos relacionados con el aprendizaje automático, especialmente clasificación de textos que es la base de este proyecto.

2.1 | Clasificación automática de Textos

El Aprendizaje Automático o *Machine Learning* (ML) es una rama de la Inteligencia Artificial, el cual tiene por objetivo desarrollar técnicas que permitan a los programas de computadoras aprender. Según [9], se dice que un programa aprende a realizar una tarea T, si después de obtener una experiencia E con una medida de desempeño P, el desempeño en la tarea T evaluada por P, mejora con la experiencia E.

Del aprendizaje automático se derivan varias categorías de tareas, una de estas tareas es la *clasificación automática de textos*, donde se entiende que un *texto* es una unidad de datos textuales que pertenece a escritos del mundo real como lo son reportes, noticias, artículos científicos, libros, publicaciones de redes sociales, etc [10]. Un texto puede ser parte de más de una colección y una colección de textos comúnmente es denominada corpus, el cual representa uno de los recursos léxicos principales en la mayoría de textos [10].

Dicho lo anterior, *la clasificación de textos* es una tarea general del aprendizaje automático que consiste en asignarle a un texto un rótulo de una clase predefinida. Existen dos tipos de rótulos de clases: multi-clase y binaria, la clasificación multi-clase consiste en asignarle un texto más de una clase, mientras que la clasificación binaria consiste

en asignarle a un texto una de dos clases.En este trabajo utilizaremos la clasificación binaria).

Formalmente, dada una colección de $D = \{d_1, ..., d_n\}$ y un conjunto de clases predefinidas $C = \{c_1, ..., c_m\}$, la tarea de clasificación puede verse como la asignación de un valor booleano a cada par (d_j, c_i), en donde $d_j \epsilon D$ y $c_i \epsilon C$. Si el valor es verdadero, entonces quiere decir que el documento d_j pertenece a la clase c_i; caso contrario, no pertenece a la clase. Sea $f : DxC \rightarrow \{verdadero, falso\}$ una función desconocida que clasifica correctamente todos los documentos, la tarea consiste en aproximar dicha función f mediante otra función h denominada $clasificador$ de tal forma que ambas coincidan tanto como sea posible [10].

Para generar un clasificador h existen dos enfoques de entrenamiento en el que se realiza el proceso de aprendizaje [10]:

- Enfoque de aprendizaje supervisado: Es una técnica donde el clasificador trata de inferir la f a través de un conjunto de datos etiquetados. Cabe mencionar que el siguiente trabajo estará basado en este tipo de aprendizaje.

- Enfoque de aprendizaje no supervisado: Consiste en una técnica que no tiene datos de entrenamiento. Por lo que la inferencia de la f se realiza a partir de datos no etiquetados.

Con el fin de crear un $clasificador$ h se requiere seguir una metodología que a su vez contiene una serie de etapas como se muestran en la Figura 2.1. Cada etapa de esta figura se explica en secciones siguientes.

2.1.1 | Extracción de características

La extracción de características generalmente consiste en dos etapas: Preprocesamiento y Representación de documentos. A continuación se describen ambas:

1. Preprocesamiento

   El *preprocesamiento* de los datos consiste en eliminar signos, palabras y caracteres sobrantes que generalmente no contienen información relevante para el clasificador. Las principales tareas que se realizan en el preprocesamiento según [11] son:

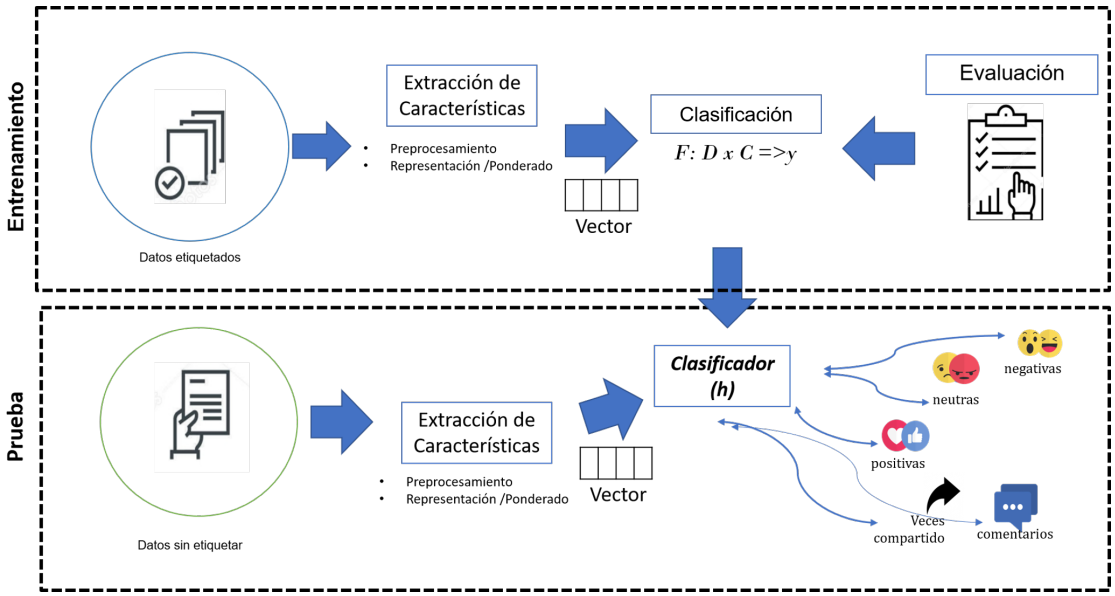

Figura 2.1: Metodología de la clasificación de textos

- Eliminación de palabras vacías: Las palabras vacías son palabras que son muy frecuentes y que por lo general no contienen información. Por ejemplo: pronombres, preposiciones, conjunciones, artículos, etc.
- Lematización de palabras. Por lematización nos referimos al proceso de remover los sufijos para reducir una palabra a su lema o raíz. Por ejemplo, dije, diré y dijéramos tiene el lema decir.
- Transformación de palabras en etiquetas: Consiste en transformar palabras o caracteres específicos en etiquetas determinadas, con el fin de identificar aspectos repetitivos o relevantes de un texto. Por ejemplo: *http: //www.google.com* es igual a **<Link>**

Cada una de las tareas mencionadas anteriormente se realiza de acuerdo con las necesidades que se tengan consideradas en cada problema de clasificación.

2. Representación de documentos

Para llevar a cabo la clasificación automática de textos, es necesario trasformar los textos de entrenamiento escritos en lenguaje natural a representaciones [10]. Existen varias maneras de representar un texto, pero la más usada es el *modelo vectorial*.

El *modelo vectorial* consiste en representar la colección de documentos o textos como una matriz de palabras o términos por documento. Es decir, cada texto será

representado por un vector de palabras en un espacio de n dimensiones, siendo n el número de palabras de los documentos. De esta manera, los documentos quedan representados como un vector $d = \{w_1, ..., w_n\}$, donde cada término indexado corresponde a una palabra en el texto y tiene un peso asociado a él, que refleja la importancia del término ya sea para el documento o para la colección completa de documentos [12].

En todo modelo vectorial los términos tienen un peso asociado, este peso representa la importancia del término en el documento o en el conjunto de documentos. Existen varios esquemas con los que se puede calcular el peso, pero a continuación solo se listan los esquemas de pesado más comunes:

- Ponderado booleano: Consiste en asignarle a cada término el peso con valor de 1 si el término ocurre en el texto y 0 si no ocurre [13].
- Ponderado por frecuencia (*TF*): Consiste en asignar a cada término un valor igual a la cantidad de veces que aparece en el documento[14].

Para crear un modelo vectorial existen varios métodos, pero el mas utilizado es el:

- Método de bolsa de palabras (*BOW* [1]): Es un método para representar textos, en donde no se toma en cuenta la gramática ni el orden de los términos. Este modelo consiste en ver a cada texto cómo una bolsa de palabras representada por medio de un vector numérico que contiene en cada instancia el peso de un término dentro del texto [13].

2.1.2 | Etapa de clasificación

Consiste en alimentar a un algoritmo de aprendizaje con los textos de entrenamiento representados en vectores, con el fin de generar como salida un *clasificador*.

Los algoritmos de aprendizaje más comunes son:

- Bayes Ingenuo (Naïve Bayes):

  Es un método probabilístico que tiene sus bases en el teorema de Bayes y recibe el apelativo de ingenuo dadas algunas simplificaciones adicionales que determinan la hipótesis de independencia de las variables predictoras [15].

[1] Del inglés, Bag of Words

- Arboles de decisión:

  Es una estructura de árbol similar a un diagrama de flujo donde un nodo interno representa una característica (o atributo), la rama representa una regla de decisión y cada nodo hoja representa el resultado o la etiqueta de la clase [16].

- Máquina de Vectores de Soporte (SVM [2]):

  Es un algoritmo que toma distintos términos de los elementos que se quieren clasificar. Dichos términos los lleva a un espacio vectorial multidimensional. En este espacio, el algoritmo identifica de forma óptima un hiperplano que separa a los vectores de una clase del resto [15].

- Clasificador K-vecinos (k-Nearest Neighbors):

  Es un algoritmo que simplemente busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de datos que le rodean [17].

2.1.3 | Etapa de evaluación

Una vez creado el modelo de clasificación, es importante conocer el desempeño de este, por lo que es necesario elegir una o más métricas de evaluación, que se acoplen a los objetivos que se quieren evaluar.

Para entender las diferentes métricas de evaluación que existen es indispensable conocer qué es una matriz de confusión, ya que de ella se calculan las métricas más utilizadas.

Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo. Cada columna de la matriz representa el número de predicciones de cada fila representa el número de instancias etiquetadas en cada clase [18]. Un ejemplo de una matriz se puede observar en la figura 2.2.

Algunas de las métricas de evaluación derivadas de una matriz de confusión son las siguientes:

- Precisión: Es el número de resultados positivos correctos dividido por el número de resultados positivos predichos por el clasificador [18].

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

[2] Del inglés, Support Vector Machines

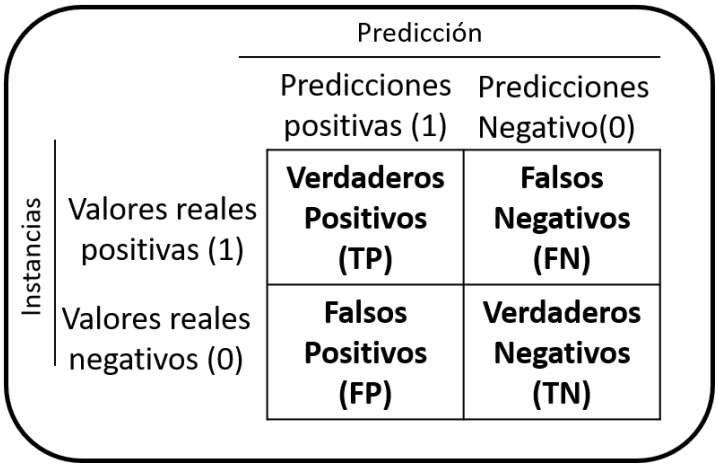

Figura 2.2: Matriz de confusión

- Exhaustividad (Recall): Es el número de resultados positivos correctos dividido por el número de todas las muestras relevantes (todas las muestras que deberían haberse identificado como positivas) [18].

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

- Exactitud (Accuracy): Es el porcentaje total de elementos clasificados correctamente [18].

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{2.3}$$

- Medida F1 (F1 Score): Es una puntuación de la media armónica entre la presión y exhaustividad. El rango para la puntuación F1 es [0, 1] [18].

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{2.4}$$

Cada una de las métricas mostradas puede ser obtenidas con una precisión macro-promedio o micro-promedio:

- Macro-promedio: Se obtiene calculando la *Precisión* y *Exhaustividad*. Luego se promedian estas medidas para obtener las medias globales que determinan a la macro [19].
- Micro-promedio. Primero se calculan los totales TP, FP,FN para todas las clases. Después se usan estos totales para calcular las medidas de rendimiento *Precisión* y *Exhaustividad* [19].

Para evaluar un clasificador es indispensable tener un conjunto de datos independiente del conjunto de entrenamiento, a este conjunto de datos se le llama conjunto prueba. Sin embargo, muchas veces dado las dimensiones del corpus no se tiene este conjunto prueba, por lo que es necesario utilizar la validación cruzada.

La validación cruzada es uno de los métodos para estimar errores de predicción. Consiste en dividir por n veces, el conjunto de datos en k partes iguales, donde una parte k es el conjunto prueba y $(k-1)$ es el conjunto entrenamiento. En cada n divisiones, k prueba es diferente de los datos de entrenamiento [11].

Al final de las n rondas los valores de las métricas de evaluación se promedian obteniendo una sola medida de evaluación global del desempeño del clasificador.

3

Trabajo relacionado

En este capítulo, se describen los trabajos centrales que se han llevado a cabo con relación a nuestros objetivos. Por lo tanto, en esta sección se resumen las literaturas que emplean métodos de aprendizaje automático, para predecir el impacto que tendrá una publicación en redes sociales.

Actualmente en el marketing digital es muy común escuchar el término *Consumer engagement* o *Compromiso del cliente* [1] . El *Consumer engagement* es la conexión emocional entre un cliente y una empresa o marca. Donde se entiende que los clientes altamente comprometidos compran más, promueven más y demuestran más lealtad [20].

El *Consumer engagement* se mide por el número de interacciones que realiza un usuario en alguna plataforma de redes sociales. Normalmente, estas interacciones varían de una plataforma a otra, sin embargo, en Facebook un conjunto típico de métricas que ayudan a evaluar el nivel de interacción son: el número de reacciones generadas (reacciones positivas, negativas y neutras), número de comentarios y el número de veces que se comparte una publicación [4]. Por lo tanto, las publicaciones que contengan números elevados en las métricas anteriores son ejemplos de publicaciones que generan un alto impacto y una relación de calidad ante el cliente, mientras que las publicaciones que contienen un número bajo en las métricas, son ejemplos de publicaciones que contienen un bajo impacto y una relación poco saludable.

Como consecuencia, la mayoría de los trabajos encontrados establecen que brindar una experiencia de alta calidad, es un componente importante para que las marcas o empresas construyan y mantengan lazos emocionales y sociales entre sus clientes. Por esta razón, muchas de estas investigaciones [2, 5, 6, 7, 8, 4] han abordado el problema

[1] Traducción al español

de cómo contribuir a que los clientes generen interacciones y relaciones con una marca mediante las diferentes plataformas de redes sociales.

La mayoría de los trabajos antes citados han enfrentado el problema como una técnica de extracción de conocimientos para diseñar potentes estrategias de marketing. En otras palabras, este tipo de investigaciones proponen analizar las relaciones entre varias variables y el nivel de compromiso que tienen los clientes. Por consiguiente, con el conocimiento de estos estudios es posible encontrar cuáles son las principales características que generan las interacciones (reacciones positivas, negativas, neutras, etc) de los clientes en una red social. Sin embargo, este tipo de análisis tiene un inconveniente y es que no consideran, el uso de este conocimiento como parte de un método automático para anticipar el impacto de una publicación.

Por otro lado, los siguientes trabajos de investigación [3, 21, 22, 23] han propuesto y evaluado metodologías distintas para la implementación de sistemas predictivos. Cada uno de estos trabajos citados será resumido de forma independiente en las siguientes subsecciones.

3.1 | Investigaciones existentes con sistemas predictivos

En el siguiente apartado se muestran los trabajos [3, 21, 22, 23], que tienen como característica general, emplear técnicas de aprendizaje computacional para anticipar el impacto que tendrá una publicación en redes sociales. A continuación, se resume cada uno de estos trabajos:

El primer trabajo a mencionar es el trabajo de Moro, Rita y Vala [3], el cual tiene como principal objetivo generar un modelo que prediga el impacto de una publicación de Facebook. Los autores consideran que el impacto de una publicación se mide a través de varias métricas disponibles relacionadas con la visualización e iteración. Donde las métricas de visualización son todos aquellos meta-atributos que cuentan el número de veces que una publicación se carga en el navegador de un usuario. Mientras que las métricas de iteración son todos aquellos meta atributos que son incrementados por un clic, por ejemplo, los comentarios, el número de veces que se comparte una publicación, etc. Viendo esta división de métricas, los autores proponen predecir 12 métricas, algunas de visualización y otras de iteración.

Dado que la arquitectura del trabajo [3] utiliza un método de aprendizaje, es nece-

sario tener un conjunto de datos, por lo que los autores se dieron a la tarea de recopilar de Enero a Diciembre del 2014, un conjunto de publicaciones de Facebook de una marca de cosméticos de renombre mundial. Como resultado, el conjunto de datos contiene un total de 790 publicaciones, de estas solo 751 son utilizadas para entrenar el modelo.

Para generar el modelo predictivo de cada una de las doce métricas de rendimiento, los autores proponen entrenar modelos con técnicas de regresión y con el algoritmo *Support Vector Machine Regression (SVR)*. Por otro lado, los modelos son alimentados con una representación vectorial compuesta por 6 características: hora, día, mes, contenido de la publicación (texto, foto o vídeo), numero de me gusta de la página y una categoría que indica si la publicación es o no pagada).

Los resultados de los experimentos de este trabajo demuestran que: predecir las métricas de interacción resultó una tarea más fácil, a diferencia de los resultados de las métricas de visualización, que obtuvieron resultados contradictorios. Según los autores el rendimiento de los modelos se debe a que las métricas de visualización están más asociadas a la aleatoriedad, pues por muchas razones cualquier usuario puede cargar el contenido de una publicación en su navegador.

Otro trabajo que emplea un enfoque parecido al anterior es la investigación de Silva, Moro, Rita [22]. Solo que este trabajo tiene como objetivo predecir el éxito que tendrán los vendedores de teléfonos inteligentes eBay. Al igual que la investigación antes descrita, este trabajo está basado en técnicas de aprendizaje automático. Por ende, se requiere de un conjunto de datos para llevar a cabo su metodología. El conjunto de datos que emplean los autores está compuesto por el historial de 623 clientes italianos de eBay de 2010 a 2012.

La metodología que proponen los autores en este trabajo para generar sus modelos predicativos consiste en: representar los textos en un vector de atributos con más de 20 características entre estas: el precio, la variedad de productos, la accesibilidad, los comentarios de los clientes, la información del cliente (nombre, país, etc). Sucesivamente también se propone entrenar los modelos con el algoritmo *Support Vector Machine*. Los resultados de este trabajo se concluyen, en que los atributos (subasta, precio y variedad de productos), son los que más influyen en la predicción del número de ventas.

El tercer trabajo perteneciente a Sabate, Ferran, Berbegal-Mirabent y Jasmina [21], también es muy parecido a los antes mencionados, pero a diferencia de estos, este trabajó tiene la finalidad de predecir mediante un modelo de regresión la cantidad de "me gusta" y la cantidad de comentarios que genera una publicación de Facebook. Para

poder realizar esto los autores se dieron a la tarea de recopilar un conjunto de 164 publicaciones en español de cinco agencias de viajes de España. Este trabajo al contar con un conjunto de datos en español, marca una gran diferencia respecto a los trabajos anteriores que contienen un conjunto de datos en el idioma inglés.

Por otra parte, los autores proponen un conjunto de características que consideran sobresalientes ante las predicciones. Estas características son agrupadas de la siguiente manera: Características de riqueza; son definidas por el número de imágenes, vídeos y links. Otra característica es el tiempo; el cual está conformado, por el día de la semana de la publicación y la hora, más dos variables de control (el número de caracteres de la publicación y el número de seguidores de la cuenta. Todas estas características son representadas en una representación vectorial y entrenadas con el algoritmo de regresión lineal.

Las conclusiones de los experimentos y el entrenamiento realizado en el trabajo [21] se centran en expresar que: incluir imágenes en una publicación está correlacionado tanto con el número de *me gusta*, como con el número de *comentarios* de un post. Así mismo, que los vídeos ayudan sólo para predecir los me gusta, pero no los comentarios. La hora del día es relevante para los comentarios solamente y por ultimo, que los links tienen un efecto negativo para predecir los comentarios.

Finalmente, en el último trabajo los autores Yano y Smith y Noah A [23] tienen como objetivo; crear un modelo que relacione el texto de una publicación de un blog de un político, con la cantidad de comentarios que este tendrá. En consecuencia, los autores desarrollaron dos metodologías; una basada en un problema de regresión, y otra en un problema de clasificación. La metodología de regresión tiene como fin predecir el valor absoluto de comentarios de una publicación de un blog, mientras que la metodología de clasificación tiene como tarea predecir el impacto de comentarios que tendrá una publicación.

En este trabajo los autores se encargaron de recopilar un conjunto de datos; compuesto por las publicaciones de dos cuentas de blogs. Una de *Matthew Yglesias* y otra de *RedState*.

Por otro lado, en esta misma investigación [23] un aspecto importante a resaltar; es que los autores no pretenden utilizar un conjunto de características para alimentar sus representaciones, como se había visto en las tres investigaciones mencionadas anteriormente. En esta investigación se utilizan técnicas de LDA[2] en lugar de características

[2] Por sus cifras en inglés, Latent Dirichlet Allocation

basadas en meta-datos. Así mismo, cada uno de los modelos es entrenado con algoritmos de regresión lineal, mientras que los modelos de clasificación son entrenados con el algoritmo Naïve Bayes.

Entre los resultados de este último trabajo se encuentra que: evaluar con dos modelos diferentes (regresión y clasificación) y utilizando las técnicas de LDA, se pueden obtener resultados aceptables, al predecir el impacto y el número absoluto de comentarios que tendrá una publicación de un blog.

Una vez concluido la descripción de cada uno de los trabajos, cabe destacar que en la tabla 3.1, se muestra un esquema comparativo de los cuatro trabajos mencionados anteriormente, más nuestra propuesta planteada. Entre los aspectos más sobresalientes a comparar de la tabla 3.1; es que solo en el trabajo [23] se incorpora el contenido del texto, para la representación de características.

Por lo tanto, a diferencia de las investigaciones anteriores; nuestra propuesta pretende utilizar el texto como parte esencial de la representación de características: basadas en el contenido, estilo y el comportamiento. Esto con el fin de validar nuestra hipótesis principal que establece: que el contenido de una publicación (lo que dice), así como el estilo en cómo se escribe (cómo se dice), en combinación con la forma en que se diseñó la publicación para interactuar con los clientes. Son elementos importantes para predecir el impacto de una publicación en Facebook.

En secciones posteriores se explica más a detalle la validación de nuestra hipótesis y los meta-atributos del texto que se consideran importantes.

Tabla 3.1: Tabla comparativa de investigaciones existentes con sistemas predictivos

Trabajo	Selección de características	Metodología	Tamaño del Data	Objetivos de predicción
[3]	6 características (Hora, día, mes contenido, No.Me gusta,etc)	- Regresión - Algoritmo SVR	790 publicaciones en *ingles* de una compañía de cosméticos	12 métricas (Visualización e interacciones) de publicaciones de Facebook
[22]	20 características (precio,accesibilidad, información del cliente, etc)	- Regresión - Algoritmo SVR	623 historiales en *ingles* de los clientes de eBay	Éxito de ventas de la compañía eBay
[21]	Características de riqueza, Caracteristicas de tiempo No.Seguidores, No.Caracteres	- Regresión - Algoritmo: Regresion Lineal	164 publicaciones en *español* de 5 agencias de viaje	No. de comentarios y No. de me gusta de publicación de Facebook
[23]	LDA	- Regresión y - Clasificación	Publicaciones en *ingles* de blogs de políticos	No. de comentarios e impacto de comentarios
Propuesta	Características de estilo, comportamiento, contenido, etc.	- Clasificación	30,000 publicaciones en *español* de empresas y políticos mexicanos	6 métricas de publicaciones de Facebook

4

Método propuesto

El presente capítulo describe el método propuesto para predecir seis problemas de clasificación, que permiten conocer el impacto general de una publicación en Facebook. Los seis problemas de clasificación son: el impacto del numero total de comentarios, el impacto del número de veces compartidos, el impacto de las reacciones totales, así como el impacto de las reacciones positivas, el impacto de las reacciones negativas y el impacto de las reacciones neutras. Por lo tanto, nuestro método tiene como objetivo aprender seis funciones $F(x)$, una para cada métrica.

El método utilizado en todos los problemas de clasificación se muestra en la figura 4.1. Dentro del recuadro punteado de esta figura se muestra que, dado un conjunto de publicaciones de Facebook (corpus), cada publicación pasa por un preprocesamiento,

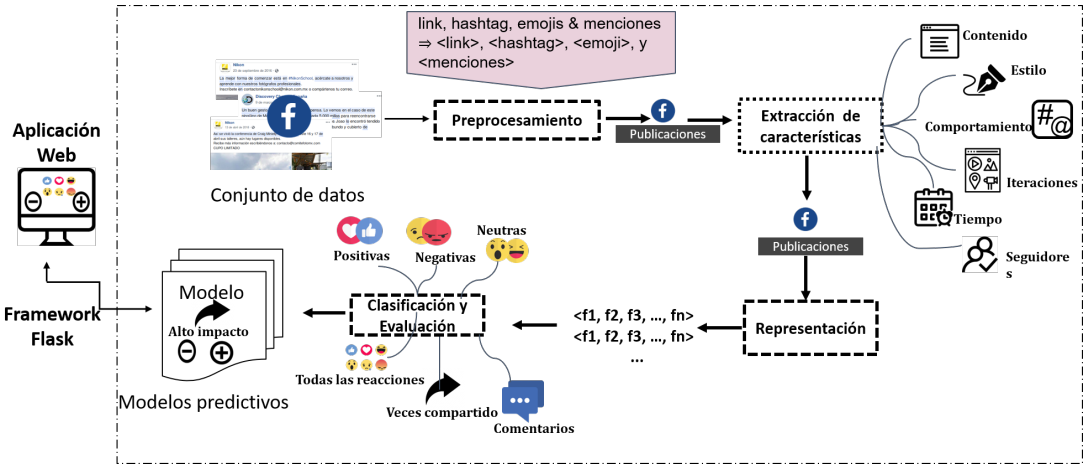

Figura 4.1: Esquema general del método propuesto

el cual tiene como objetivo remplazar del contenido de la publicación, los links, emojis, hashtags y menciones por las etiquetas <links>, <emojis>, <hastang> y <menciones>. Sucesivamente, cada publicación pasa por el proceso de extracción de características, este consiste en obtener los atributos que se describen en la sección 4.3. Hecho el paso anterior se procede a la etapa de representación, donde las publicaciones se transforman en un vector multidimensional. Posteriormente, se procede a la clasificación y evaluación, en esta etapa se entrena con la representación a más de un clasificador, una vez entrenados los clasificadores se evalúan y se seleccionan los modelos predictivos óptimos de cada métrica. Finalmente, cada modelo es incorporado a la aplicación web. La cual tiene como función capturar una publicación de Facebook y predecir el impacto que tendrá.

A continuación, en las siguientes secciones se describe más a detalle el conjunto de datos utilizado, así como cada proceso de la figura 4.1.

4.1 | Colección de datos

La primera fase de este método requiere contar con un conjunto de publicaciones en español de Facebook, que evalúen la predicción del impacto de las publicaciones. Sin embargo, dada la falta de un corpus en español, asumimos la tarea de recopilar un conjunto de publicaciones de Facebook; de catorce cuentas públicas de políticos mexicanos que se postularon o tenían un cargo público en el año 2018 y diez cuentas públicas de empresas reconocidas en México. El conjunto de datos se recolectó en los meses de octubre a diciembre de 2018, con apoyo de la Red Temática de Tecnologías del Lenguaje (RedTTL) [1].

El corpus recolectado contiene 37,585 publicaciones, de estas 14,522 pertenecen a las publicaciones de las empresas y 23,063 a los políticos. Cada publicación contiene una colección de meta-datos como lo son: el día, la hora, el mes y el año en el que se publica, el número de imágenes, vídeos, álbum y links, el texto de la publicación, el número absoluto de cada reacción (me gusta, me encanta, me asombra, me divierte, me enoja y me entristece), el numero absoluto de las veces que se comparte y el numero absoluto de comentarios de la publicación.

Para conocer más a detalle el corpus en la tabla 4.1 se muestra el resumen del tamaño del corpus. Mientras, que la volumen de meta atributos y otros elementos importantes que contiene el conjunto de datos se detallan en la subsección 4.1.1.

[1] http://www.redttl.mx/

En la tabla 4.1 bajo la columna *Num de publicaciones*, se muestran dos columnas, (OG) representa el número absoluto de publicaciones recopiladas y (FL) representa la cantidad resultante después de haber descartado algunas publicaciones. En particular se eliminaron todas aquellas que cumplían con alguno de los siguientes puntos:

- No contienen texto
- No contienen ninguna reacción
- La única reacción que contienen es "Me gusta"

Al final se eliminaron 2,863 publicaciones del corpus, de estas 871 pertenecen a las empresas y 1,992 a los políticos. Así mismo, si se observa la tabla 4.1 podemos destacar que de las cuentas de las empresas *Muy interesante México* es la que más genera publicaciones, mientras que de las cuentas de los políticos, es la cuenta de la política *Alejandra Barrales*, quien fue candidata a jefa de gobierno en la Ciudad de México en las elecciones del 2018.

Tabla 4.1:

Giro la cuenta	**Nombre de la cuenta**	**Num. de publicaciones** (OG)	(FL)
Empresas	Clash Royale ES *(CR)*	561	558
	Canon Mexicana *(CM)*	1157	1114
	Muy Interesante México *(MIM)*	2175	2168
	Cinépolis *(CP)*	1985	1949
	Discovery Channel *(CH)*	1712	1675
	National Geographic *(NG)*	2076	1744
	Fisher-Price *(FP)*	848	842
	Xbox México *(XM)*	1737	1654
	Nikon *(NK)*	1357	1232
	Lacoste *(LC)*	914	715
Total empresas:		**14522**	**13651**
Políticos	Alejandra Barrales *(AB)*	2134	1853
	Alfredo Del Mazo Maza *(AMM)*	1813	1691
	Andrés Manuel López Obrador *(AMLO)*	1150	1128
	Claudia Sheinbaum *(CS)*	1620	1492
	Delfina Gómez Álvarez *(DGA)*	1797	1415
	Enrique Peña Nieto *(EPN)*	1711	1673
	Gerardo Fernández Noroña *(GFN)*	2096	1772
	Jaime Rodriguez Calderon *(JRC)*	2005	1927
	José Antonio Meade *(JAM)*	1883	1783
	Manuel Velasco Coello *(MVC)*	1901	1827
	Mariana Boy *(MB)*	277	270
	Martha Erika Alonso *(MEA)*	1912	1806
	Rafael Moreno Valle *(RMV)*	1483	1213
	Ricardo Anaya Cortés *(RAC)*	1281	1221
Total políticos:		**23063**	**21071**
Total de publicaciones:		**37585**	**34722**

4.1.1 | Estadísticas de los datos

El siguiente apartado, muestra las estadísticas más relevantes de los meta-datos del corpus; tanto del conjunto de las empresas como de los políticos. Las estadísticas se calcularon con la versión FL de cada conjunto de datos (empresas y políticos) que se observa en la tabla 4.1.

Por consiguiente, las tablas 4.2 y 4.3 muestran las estadísticas básicas del número de reacciones, veces compartidos y comentarios que tienen las publicaciones, así también, la popularidad de cada una de las cuentas. Los valores mostrados en la tabla 4.2 son calculados con el conjunto de datos de las empresas, mientras, que los valores de la tabla 4.3 son calculados con el conjunto de datos de los políticos. En ambas tablas, la primera columna posee el nombre de la cuenta abreviado (ver abreviaciones en la tabla 4.1), sucesivamente en las 3 columnas siguientes se observan los valores de las reacciones (R), comentarios (C) y veces compartidos (VC). De bajo de cada una de estas columnas se muestra $|R|$, $|C|$, $|VC|$ que representan el número absoluto de reacciones, comentarios y veces compartidos de todas publicación de cada empresa al momento en el que se recopiló este corpus. Consecutivamente, los valores x y σ indican el promedio y la desviación estándar. Finalmente, las ultima dos columnas: Seguidores (S) y Me gusta (MG) muestran el número absoluto de seguidores $|S|$ y de me gusta $|MG|$ que contiene cada cuenta, al momento de extraer el conjunto de datos.

Tabla 4.2: Tabla que muestra el número absoluto, la desviación estándar y el promedio, que tienen los comentarios, veces compartidos y reacciones de las publicaciones de las *empresas*.

Nombre de la cuenta	**Reacciones (R)**			**Comentarios (C)**			**Veces Compartidos (VC)**			**Seguidores (S)**	**Me gustas (MG)**
	$\|R\|$	x	σ	$\|C\|$	x	σ	$\|VC\|$	x	σ	$\|S\|$	$\|MG\|$
CR	3,464,687	6209.12	11457.63	561,540	1006.34	2341.34	258,904	463.99	1495.29	4,807,148	4,679,280
CM	2,316,406	2079.36	6626.56	112,280	100.79	265.89	426,502	382.88	1030.55	4,453,565	4,450,690
MIM	14,900,267	6872.82	11314.93	258,214	119.10	296.41	4,801,967	2214.93	8436.72	8,108,208	8,234,731
CP	28,074,276	14404.45	28790.81	2,784,917	1428.90	4179.64	8,165,961	4189.82	18652.82	18,315,418	18,271,513
DC	2,422,143	1446.06	2164.16	44,258	26.42	63.43	392,068	234.07	474.24	1,324,191	1,326,106
NG	2,700,761	1548.60	3680.85	107,884	61.86	247.85	1,034,515	593.19	4643.73	63,704,330	63,463,491
FP	3,550,516	4216.76	6053.21	155,811	185.05	400.89	249,981	296.89	709.63	6,587,392	6,577,083
XM	4,697,179	2839.89	6360.91	479,033	289.62	749.36	513,323	310.35	860.45	3,882,848	3,899,524
NK	829,560	673.34	1306.00	36,308	29.47	81.29	145,370	118.00	262.84	14,944,460	14,943,253
LC	1,057,118	1478.49	2559.50	10,005	13.99	36.79	37,037	51.80	266.68	15,235,016	15,240,405
Total:	*64,012,913*	-	-	*4,550,250*	-	-	*16,025,628*	-	-	*147,949,968*	*147,663,159*

Observe que en la tabla 4.2, la empresa que más produce reacciones, comentarios y veces compartidos es *Cinepolis*. Esta empresa es muy conocida en México, por dedicarse

Tabla 4.3: Tabla que muestra el número absoluto, la desviación estándar y el promedio, que tienen los comentarios, veces compartidos y reacciones de las publicaciones de los *políticos*.

Nombre de la cuenta	**Reacciones (R)**			**Comentarios (C)**			**Veces Compartidos (VC)**			**Seguidores (S)**	**Me gustas (MG)**
	$\|R\|$	x	σ	$\|C\|$	x	σ	$\|VC\|$	x	σ	$\|S\|$	$\|MG\|$
AB	1,264,546	152.99	448.33	237,212	128.02	272.09	283,483	153	448	303,009	305,453
AMM	3,952,803	542.1	706.63	497,458	294.18	765.31	916,698	542	707	993,471	982,391
AMLO	26,985,815	10,796.74	17,798.28	2,604,937	2,309.34	3,722.53	12,178,725	10,797	17,798	6,430,878	5,982,214
CS	1,491,677	257.56	1,190.29	129,440	86.76	393.04	384,276	258	1,190	290,041	280,752
DGÁ	2,195,972	614.59	3,183.07	156,341	110.49	594.69	869,640	615	3,183	344,514	338,860
EPN	9,963,773	1,118.54	4,265.31	1,395,358	834.05	5,002.93	1,871,322	1,119	4,265	5,722,724	5,799,270
GFN	4,959,826	2,356.99	17,012.74	670,350	378.3	777.97	4,176,579	2,357	17,013	1,011,839	1,136,245
JRC	7,183,832	1,373.59	9,120.54	1,291,465	670.19	2,252.04	2,646,909	1,374	9,121	2,914,278	2,925,950
JAM	5,056,513	839.79	1,688.75	1,251,043	701.65	2,831.43	1,497,340	840	1,689	839,905	874,423
MVC	12,931,851	813.72	2,100.50	549,481	300.76	473.67	1,486,662	814	2,100	3,485,577	3,457,439
MB	31,509	14.97	16.88	1,714	6.35	17.54	4,041	15	17	6,153	6,421
MEA	1,678,867	171.09	289.77	197,216	109.2	340.26	308,982	171	290	174,372	167,405
RMV	3,594,736	693.37	1,389.06	249,803	205.94	321.99	841,053	693	1,389	2,246,939	2,263,533
RAC	21,170,806	3,010.76	9,328.24	3,474,221	2,845.39	7,854.98	3,676,141	3,011	9,328	2,103,742	2,052,749
Total:	*102,462,526*	-	-	*12,706,039*	-	-	*31,141,851*	-	-	*26,867,442*	*26,573,105*

a la exhibición de películas en el cine. En segundo lugar, la empresa que más produce reacciones y veces compartidos es *Muy interesante México*. Esta marca esta dedicada principalmente a la difusión de la ciencia y la tecnología. Así mismo, es interesante notar que a pesar de que *Cinepolis* provoca la mayor cantidad de interacciones en sus usuarios, no es la empresa que más produce publicaciones, ni mucho menos contiene el mayor números de seguidores, como es el caso de *Muy interesante México*.

Al observar la tabla de los políticos (Tabla 4.3), el político que más produce re-/acciones y veces compartidos es *Andrés Manuel López Obrador*. Quien es el presidente actual de México a partir del 1 de diciembre de 2018. Mientras, que el político *Ricardo Anaya Cortés* es el que más genera comentarios y seguido de *Andrés Manuel López Obrador* es el que más produce reacciones. Podemos asumir, que esto se debe a que ambos políticos fueron los candidatos a la presidencia de México con mayor auge en las elecciones del 2018. Del igual forma, es sobresaliente notar que a pesar de que el político *Ricardo Anaya Cortés* es uno de los que más genera manifestaciones en sus usuarios, no es el que más publica, ni tiene el mayor número de seguidores, como lo es el caso del político *Andrés Manuel López Obrador* quien tiene el mayor número de seguidores o la política *Alejandra Barrales* que tuvo mayor número de publicaciones en el periodo analizado.

Como se ha mencionado, parte de la hipótesis de este trabajo consiste en considerar al texto de las publicaciones, como un elemento indispensable para predecir el impacto

Tabla 4.4: Esta tabla muestra el tamaño de palabras, vocabulario, riqueza léxica, el número promedio de palabras y caracteres de las publicaciones de las *empresas*

Nombre de la cuenta	**Número total de:**			**Número promedio por publicación:**	
	palabras	*vocabulario*	*riqueza léxica*	*palabras* (σ)	*caracteres* (σ)
Clash Royale ES (CR)	14,531	4,046	0.27	26.04 ($\pm$27.53)	163.21 ($\pm$165.07)
Canon Mexicana (CM)	21,885	5,006	0.22	19.65 ($\pm$12.63)	128.02 ($\pm$81.06)
Muy Interesante México (MIM)	42,321	8,916	0.21	19.52 ($\pm$14.74)	117.70 ($\pm$88.79)
Cinépolis (CP)	44,071	7,536	0.17	22.61 ($\pm$10.78)	133.95 ($\pm$64.36)
Discovery Channel (DC)	44,659	10,862	0.24	26.66 ($\pm$12.94)	158.09 ($\pm$75.28)
National Geographic (NG)	46,039	6,988	0.15	26.40 ($\pm$13.33)	153.16 ($\pm$73.32)
Fisher-Price (FP)	19,863	4,788	0.24	23.59 ($\pm$78.70)	149.89 ($\pm$517.25)
Xbox México (XM)	27,639	5,283	0.19	16.71 ($\pm$7.21)	112.27 ($\pm$52.98)
Nikon (NK)	28,251	6,044	0.21	22.93 ($\pm$42.99)	147.28 ($\pm$278.2)
Lacoste (LC)	13,455	3,570	0.26	18.82 ($\pm$24.75)	128.06 ($\pm$155.12)

de una publicación. Es por ello, que las tablas 4.4 y 4.5 muestran algunas estadísticas básicas sobre el tamaño del texto. La tabla 4.4 muestra las estadísticas calculadas con el conjunto de datos de las empresas y la tabla 4.5 con el conjunto de datos de los políticos. Ambas tablas contienen, en la primera columna el nombre de la cuenta, sucesivamente en las siguientes tres columnas muestran el tamaño del texto de las publicaciones de cada cuenta, en términos del tamaño del número de palabras, tamaño del vocabulario y la riqueza léxica de las publicaciones. Posteriormente las últimas dos columnas muestran el numero promedio de las palabras y el numero promedio de los caracteres que contienen las publicaciones.

A partir de la tabla 4.4, podemos expresar que las empresas que tienen el mayor número de palabras en sus publicaciones son: National Geographic y Discovery Channel, ambas están dedicadas a la difusión de programas relacionados con la naturaleza, ecología, vida salvaje, ciencia, entre otros temas. Por lo tanto, tener un gran número de palabras indica que en general, las publicaciones de estas empresas tienden a ser grandes en términos de palabras. Esto lo podemos ver en la columna cinco de la tabla 4.4 , donde es posible observar el numero promedio de palabras que tienen las publicaciones.

Ahora bien, también es relevante expresar que la riqueza léxica (RL) es el valor que indica como se utilizan los términos del vocabulario dentro de un texto. Formalmente se dice que la riqueza léxica es la división entre el tamaño del vocabulario y la cantidad de palabras de un texto ($LR = |V|/|T|$). Por lo tanto, si se obtiene un (LR) con un valor cercano a 1 significa que los términos del vocabulario se usan solo una vez, mientras que si se obtienen valores cercanos a 0 representan un número mayor de palabras, que se usan con más frecuencia (es decir, más repetitivos).

Tabla 4.5: Esta tabla muestra el tamaño de palabras, vocabulario, riqueza léxica, el número promedio de palabras y caracteres de las publicaciones de los *políticos*

Nombre de la cuenta	**Número total de:**			**Número promedio por publicación:**	
	palabras	*vocabulario*	*riqueza léxica*	*palabras* (σ)	*caracteres* (σ)
Alejandra Barrales(AB)	64,777	9,284	0.14	34.96 ($\pm$36.34)	213.88 ($\pm$ 34.96)
Alfredo Del Mazo Maza (AMM)	46,471	8,796	0.19	27.48 ($\pm$ 18.55)	117.79 ($\pm$ 27.48)
Andrés Manuel López (AMLO)	59,491	13,045	0.22	52.74 ($\pm$ 173.14)	1038.1 ($\pm$ 52.74)
Claudia Sheinbaum (CS)	48,421	9,375	0.19	32.45 ($\pm$ 44.49)	279.34 ($\pm$ 32.45)
Delfina Gómez Álvarez (DGA)	43,799	7,743	0.18	30.95 ($\pm$ 20.93)	130.88 ($\pm$ 30.95)
Enrique Peña Nieto (EPN)	64,227	10,121	0.16	38.39 ($\pm$ 33.04)	205.26 ($\pm$ 38.39)
Gerardo Fernández (GFN)	50,081	12,268	0.24	28.26 ($\pm$ 115.07)	692.71 ($\pm$ 28.26)
Jaime Rodriguez Calderon (jJRC)	72,828	11,633	0.16	37.79 ($\pm$ 31.84)	187.94 ($\pm$ 37.79)
José Antonio Meade (JAM)	46,076	8,207	0.18	25.84 ($\pm$ 22.02)	139.02 ($\pm$ 25.84)
Manuel Velasco Coello (MVC)	65,270	9,189	0.14	35.73 ($\pm$ 18.18)	112.79 ($\pm$ 35.73)
Mariana Boy (MB)	8,650	2,599	0.3	32.04 ($\pm$ 18.43)	110.28 ($\pm$ 35.73)
Martha Erika Alonso (MEA)	55,468	8,899	0.16	30.71 ($\pm$ 27.16)	165.49 ($\pm$ 30.71)
Rafael Moreno Valle (RMV)	39,701	7,529	0.19	32.73 ($\pm$ 15.65)	98.42 ($\pm$ 32.73)
Ricardo Anaya Cortés (RAC)	41,593	7,026	0.17	34.06 ($\pm$15.92)	96.21 ($\pm$ 34.06)

Por lo tanto, observe que en la tabla 4.4 las empresas con los valores más bajos de LR son: National Geographic y Cinépolis, lo que significa que sus publicaciones emplean un lenguaje similar. Nosotros asumimos, que esto podría ser una estrategia de marketing ya que, para el caso de Cinépolis escribir publicaciones con un estilo repetitivo le ha permitido tener un impacto alto.

En contraste, ahora observe que en la tabla 4.5 los políticos que tienen el mayor número de palabras en sus publicaciones son: Jaime Rodriguez Calderon, Manuel Velasco Coello y Enrique Peña Nieto. Por lo tanto, esto indica que en general, estos políticos tienden a escribir publicaciones grandes en términos de palabras. Esto se puede observar en la columna cinco de la tabla 4.5, donde se visualiza el número promedio de palabras que tienen las publicaciones. Sucesivamente, observe que en la columna tres de la tabla 4.5, los valores más bajos en LR pertenecen a las publicaciones de Jaime Rodríguez Calderón y Manuel Velasco Coello. Esto representa que ambos políticos emplean publicaciones con un lenguaje similar.

Desde otra perspectiva, es significativo recalcar que las publicaciones de Jaime Rodríguez Calderón y Manuel Velasco Coello, tienden a ser un poco diferentes, con respecto a las de Andrés Manuel López Obrador y Ricardo Anaya Cortés, quienes sus publicaciones son las que generan la mayor cantidad de interacciones en sus usuarios. Las diferencias más notables son: que Andrés Manuel López Obrador y Ricardo Anaya Cortés escriben publicaciones cortas, así también, estas publicaciones no son tan repetitivas, esto se puede observar en la columna tres de la tabla 4.5. Donde es posible ver

que la *LR* de las publicaciones de estos políticos es más alto en comparación, con la *LR* de Jaime Rodríguez Calderón y Manuel Velasco Coello.

Por último, las tablas 4.6 y 4.7 muestran algunos otros promedios de los meta- datos. En la tabla 4.7 se observan los resultados obtenidos con el conjunto de datos de las empresas y en la tabla 4.7 con el conjunto de datos de los políticos. Ambas tablas, contienen en la primera columna el nombre de la cuenta, mientras que el resto de las columnas contiene el promedio y entre paréntesis la desviación estándar de dígitos numéricos, número de hashtags, menciones, emojis y links que contienen las publicaciones de cada cuenta.

Tabla 4.6: Esta tabla muestra el número promedio de números, hashtags, menciones, emojis y links que contienen las publicaciones de las *empresas*. Entre paréntesis se indica la desviación estándar.

Nombre de la cuenta	**Número promedio de :**				
	números	*hashtags*	*menciones*	*emojis*	*links*
Clash Royale ES	1.14 (±2.35)	0.06 (±0.25)	0.13 (±0.46)	1.42(±1.12)	1.93 (±1.07)
Canon Mexicana	0.89 (±2.66)	0.70 (±0.63)	0.06 (±0.26)	1.10 (±1.43)	1.64 (±1.34)
Muy Interesante México	0.63 (±1.70)	0.09 (±0.33)	0.08 (±0.28)	0.09 (±0.40)	0.82 (±0.78)
Cinépolis	0.47 (±1.22)	1.09 (±1.02)	0.29 (±0.74)	0.02 (±0.25)	1.19 (±0.82)
Discovery Channel España	2.12 (±2.72)	0.06 (±0.30)	0.02 (±0.17)	0.23 (±0.49)	0.86 (±1.05)
National Geographic	3.25 (±3.71)	0.84 (±0.70)	0.09 (±0.30)	0.38 (±0.96)	1.05 (±0.62)
Fisher-Price	0.54 (±5.01)	0.08 (±0.35)	0.03 (±0.19)	1.44 (±1.31)	0.57 (±0.64)
Xbox México	1.14 (±3.17)	0.73 (±0.69)	0.07 (±0.30)	0.39 (±0.85)	1.52 (±0.84)
Nikon	1.52 (±3.33)	0.47 (±0.57)	0.18 (±0.39)	0.19 (±0.46)	1.22 (±1.06)
Lacoste	1.18 (±2.86)	0.58 (±0.81)	0.07 (±0.29)	0.06 (±0.29)	1.15 (±1.01)

Como se puede observar en la tabla 4.6, las publicaciones de la empresa National Geographic tienen la mayor cantidad de números, hastags y links. Así mismo, la empresa Clash Royale reconocida por la comercialización de videojuegos, es la que más destaca en tener emojis y links en sus publicaciones.

Para concluir, en la tabla 4.7 se puede visualizar que las publicaciones que tienen la mayor cantidad de menciones y links pertenecen al político Alfredo Del Mazo, mientras que las publicaciones de Gerardo Fernández contienen el mayor número de números, en tanto que en las publicaciones de Manuel Velasco los *hashtags* se utilizan con mayor frecuencia y por último en las publicaciones de Martha Erika Alonso el

Tabla 4.7: Esta tabla muestra el número promedio de números, hashtags, menciones, emojis y links que contienen las publicaciones de los *políticos*. Entre paréntesis se indica la desviación estándar.

Nombre de la cuenta	**Número promedio de :**				
	números	*hashtags*	*menciones*	*emojis*	*links*
AB	0.46 (±2.6)	0.59 (±0.8)	0.13 (±0.46)	0.13 (±0.56)	1.07 (±0.43)
AMM	0.7 (±1.75)	0.77 (±1.03)	0.28 (±0.62)	0.03 (±0.29)	1.3 (±1.08)
AMLO	1.91 (±13.2)	0.02 (±0.31)	0.02 (±0.63)	0 (±0.06)	1.16 (±0.96)
CS	1.14 (±4.69)	0.37 (±0.69)	0.26 (±0.75)	0.02 (±0.18)	2.25 (±1.74)
DGA	0.54 (±1.64)	1.36 (±1.55)	0.16 (±0.57)	0.02 (±0.19)	0.96 (±1.26)
EPN	1.05 (±2.39)	0.08 (±0.33)	0.03 (±0.22)	0.03 (±0.35)	1.02 (±0.19)
GFN	1.73 (±6.14)	0.09 (±0.59)	0.18 (±0.58)	0.01 (±0.11)	0.97 (±0.78)
JRC	1.03 (±3.33)	0.11 (±0.45)	0.14 (±0.51)	0.03 (±0.34	1.08 (±0.63)
JAM	0.64 (±2.03)	0.47 (±0.73)	0.21 (±0.54)	0.01 (±0.14)	1.48 (±1.04)
MVC	1.45 (±02.6)	1.48 (±1.23)	0.15 (±0.42)	0 (±0.02)	1.27 (±1.08)
MB	0.55 (±1.51)	1.34 (±0.87)	0.25 (±0.88)	0.08 (±0.46)	2.34 (±1.58)
MEA	0.34 (±1.58)	0.48 (±0.62)	0.22 (±0.54)	0.27 (±0.98)	0.98 (±0.23)
RMV	0.89 (±2.27)	0.35 (±0.59)	0.23 (±0.61)	0.1 (±0.4)	1.02 (±0.16)
RAC	0.66 (±1.81)	0.65 (±0.75)	0.02 (±0.2)	0.01 (±0.12)	1.17 (±1.3)

uso de emojis es notorio. Por otro lado, es relevante destacar que las publicaciones de Andrés Manuel López Obrador y Ricardo Anaya Cortés a pesar de que son las que más producen interacciones en sus usuarios, no son las que más contienen números, hashtags, menciones, emojis y links.

4.1.2 | Metodología del etiquetado

Como se mencionó anteriormente el objetivo de tener un conjunto de datos de Facebook es para entrenar y evaluar el rendimiento de los métodos de clasificación automática, para determinar el impacto de una publicación de Facebook. Por lo tanto, e inspirados en el trabajo de los autores Yano y Smith [23], definimos la tarea de predecir el impacto de una publicación de Facebook:

- Como el proceso de clasificar si una publicación tendrá un volumen de impacto *alto* o *bajo* con respecto al promedio observado en los datos recolectados.

Otro ejemplo de cómo predecir el impacto de una publicación es: Anticipar el

número absoluto de reacciones, de comentarios o veces en las que se comparte una publicación.

En cambio, en este trabajo solo se definió anticipar el impacto de una publicación, como un problema de clasificación binaria, es decir, solo se desea aprender a predecir si una publicación tendrá alto o bajo impacto, en las siguientes métricas:

- Comentarios $|C|$
- Reacciones totales $|R|$
- Reacciones negativas $|R-|$
- Veces compartidos $|V|$
- Reacciones positivas $|R+|$
- Reacciones neutras $|R\odot|$

Dado que el corpus tiene dos conjuntos de publicaciones uno perteneciente a cuentas de empresas y otro a cuentas de políticos, la metodología que se siguió para asignarle a cada publicación el valor (*alto* o *bajo*) en las seis métricas, se aplicó de manera idéntica en ambos conjuntos.

La metodología que se siguió para etiquetar las distintas métricas de cada publicación es:

1. Por cada problema de clasificación (es decir las métricas: $|C|$, $|VC|$, $|R|$, $|R+|$, $|R-|$ y $|R\odot|$) calculamos el valor promedio de la métrica K, entre todas las publicaciones del conjunto de datos ya sea de las empresas o políticos. Al valor promedio se le conoce como $\bar{x}_k$. Para conocer el valor promedio de las métricas de las empresas y políticos ver la tabla 4.8.

2. Posteriormente, para cada publicación p contenida en la cuenta i, se revisa el valor de p_i en la métrica k. Así entonces, si $p_{i,k} > \bar{x}_k$, el valor de la clase que se asigna a p_i será *alto impacto*, de lo contrario *bajo impacto* .

El proceso de etiquetado representa un enfoque muy sencillo para asignar el valor *alto* o *bajo* en las clasificaciones de las *reacciones totales*, los *comentarios* y las *veces compartidas*. Mientras que, para etiquetar las *reacciones positivas, negativas y neutrales*, actuamos

de la siguiente manera; agrupamos como reacciones positivas: las reacciones *me gusta* y *me encanta*, como reacciones negativas las reacciones *me entristece* y me *enoja* y como reacciones neutrales las reacciones *me asombran* y *me divierte*.

En la tabla 4.9 y 4.10 se muestra el número de instancias de cada categoría después del proceso de etiquetado de las publicaciones. Cabe señalar que la tabla 4.9 pertenece al etiquetado de las publicaciones de las empresas y la tabla 4.10 al etiquetado de las publicaciones de los políticos. Para concluir, la gran mayoría de las publicaciones son de bajo impacto, resultando en un conjunto datos altamente des balanceado, por lo que la tarea de clasificación es más compleja

Tabla 4.8: Número de promedio ($\bar{x}_k$) de k métricas de *empresas* y *políticos*

Métrica k	$\bar{x}_k$ Promedio de:	
	Participación de empresas	Participación de los políticos
Reacciones |R|	4689.24	4862.72
Reacciones positivas $\|R+\|$	4054.47	4340.35
Reacciones negativas $\|R-\|$	80.16	139.42
Reacciones neutras $\|R\odot\|$	554.6	382.94
Comentarios $\|C\|$	333	603.01
Veces Compartidos $\|V\|$	1173.95	1477.94

Tabla 4.9: Número de instancias de *alto* y *bajo*- impacto para cada problema de clasificación con las publicaciones de las *empresas*

Nombre de la cuenta	$\|R\|$		$\|R+\|$		$\|R-\|$		$\|R\odot\|$		$\|C\|$		$\|V\|$	
	alto	*bajo*	*alto*	*bajo*	*alto*	*bajo*	*alto*	*bajo*	*alto*	*bajo*	*alto*	*bajo*
CR	189	369	165	393	209	349	217	341	264	294	39	519
CM	100	1014	94	1020	43	1071	90	1024	78	1036	96	1018
MIM	775	1393	790	1378	136	2032	374	1794	153	2015	824	1344
CI	991	958	966	983	353	1596	729	1190	883	1067	745	1204
DC	109	1566	112	1563	110	1565	85	1590	9	1666	60	1615
NG	124	1620	132	1612	110	1634	53	1691	50	1694	115	1629
FP	248	594	266	576	15	827	31	811	119	723	43	799
XM	230	1424	230	1424	168	1486	155	1499	299	1355	92	1562
NK	24	1208	33	1199	16	1216	6	1226	12	1220	10	1222
LC	46	669	57	658	0	715	2	713	2	713	4	711

Tabla 4.10: Número de instancias de *alto* y *bajo*- impacto para cada problema de clasificación con las publicaciones de los *políticos*

Nombre de la cuenta	$\|R\|$		$\|R+\|$		$\|R-\|$		$\|R\odot\|$		$\|C\|$		$\|S\|$	
	alto	*bajo*	*alto*	*bajo*	*alto*	*bajo*	*alto*	*bajo*	*alto*	*bajo*	*alto*	*bajo*
AB	50	1803	29	1824	70	1783	103	1750	79	1774	18	1835
AMM	204	1487	184	1507	282	1409	155	1536	190	1501	116	1575
AMLO	1117	11	1118	10	211	917	461	667	853	275	1052	76
CS	39	1453	34	1458	104	1388	70	1422	29	1463	41	1451
DGA	85	1330	94	1321	72	1343	21	1394	26	1389	78	1337
EPN	540	1133	527	1146	579	1094	247	1426	371	1302	204	1469
GFN	179	1593	158	1614	265	1507	152	1620	280	1492	223	1549
JRC	326	1601	344	1583	122	1805	137	1790	391	1536	288	1639
JAM	272	1511	264	1519	199	1584	222	1561	267	1516	284	1499
MVC	560	1267	582	1245	236	1591	277	1550	167	1660	200	1627
MB	0	270	0	270	2	268	1	269	0	270	0	270
MEA	36	1770	28	1778	61	1745	79	1727	61	1745	15	1791
RMV	187	1026	196	1017	78	1135	81	1132	68	1145	121	1092
RAC	588	633	610	611	338	883	450	771	482	739	415	806

4.2 | Preprocesamiento

Una vez teniendo nuestro corpus etiquetado, la siguiente etapa según la figura 4.1 consiste en aplicarle a cada publicación del corpus un preprocesamiento. La idea general de esta etapa reside en estandarizar el contenido de las publicaciones con el fin de identificar aspectos repetitivos o relevantes de un texto. Por ejemplo, para fines de este proyecto no nos preocupamos por saber sobre los diferentes *hashtags* o *emojis* que tiene el contenido de una publicación; si no por solo saber si tiene *emojis* o *hashtags*. En consecuencia, el procesamiento que aplicamos en cada una de las publicaciones consiste en:

- Remplazar todas las *URL* del contenido de una publicación por la etiqueta <*URL*>.
- Remplazar todos los *hashtags* [2] de una publicación por la etiqueta <*hashtag*>
- Remplazar todos los *emojis* de una publicación por la etiqueta <*emoji*>

[2] Son palabras u oraciones (sin espacios) que van precedidas del símbolo #.

- Y por último remplazar todas las *menciones* [3] de una publicación por la etiqueta *<mención>*

Se puede observar un ejemplo de una publicación con contenido preprocesado en la figura 4.2.

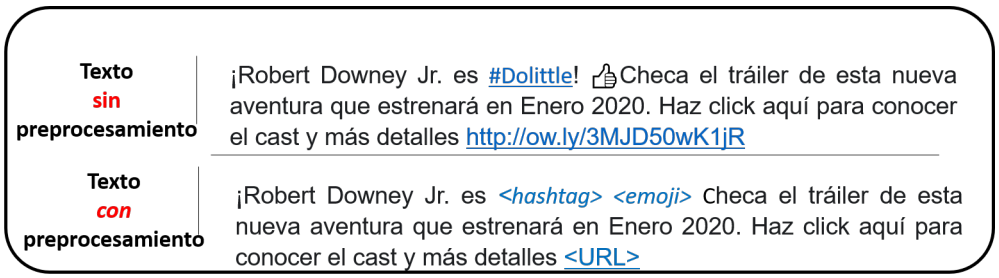

Figura 4.2: Ejemplo de texto preprocesamiento

El preprocesamiento antes mencionado se empleó con cada una de las publicaciones del conjunto de datos de las empresas y de los políticos.

4.3 | Extracción de atributos

La siguiente etapa que indica la figura 4.1 es el proceso de extracción de atributos. Esta fase consiste principalmente en identificar aquellos elementos de una publicación que se consideren importantes para predecir su impacto en Facebook. Como hemos establecido en la hipótesis de este trabajo, consideramos que utilizar el texto es parte esencial para representar una publicación, pues, es a partir del contenido que podemos conocer el qué y el cómo se publica.

Nuestra metodología para seleccionar características consistió en extraer tres tipos de atributos a partir del contenido: el primer tipo de atributo es llamado *Contenido* y captura el *qué* se publica, el segundo tipo de atributo es llamada *Estilo* y representa el *cómo* y el último tipo de atributo es llamada *Comportamiento* e incorpora el *cómo se diseña una publicación*. Por otro lado, también se consideró sobresaliente incluir otros tres tipos de atributos, basados en los meta-datos de las publicaciones: Estos tipos son: *Interacciones, popularidad y tiempo*. Los cuales también se han considerado útiles en los trabajos relacionados [3], [21].

[3]Son enlaces a un perfil o una página concretos dentro de la red social.

Para entrar en detalle, cada uno de los tipos de atributos se conforma por un subconjunto de características que a continuación se mencionan:

1. Características de contenido (t): Representa el texto de una publicación, donde consideramos las palabras individuales como atributos. Para obtener los atributos de contenido se utilizaron bolsas n-gramas de palabras, específicamente con n tamaño 1, 2, y 3.
2. Características de estilo (e): Esta definido por cinco variables es como: el número de palabras, el número de mayúsculas, minúsculas, números y signos de una publicación.
3. Características de comportamiento (c): Definido por cuatro características: el número de emojis, hashtags, menciones y links que tiene una publicación.
4. Características de interacción (i): Definido por cinco características el número de links a imágenes, a álbum, a vídeos, o a otros, que contiene una publicación.
5. Características de tiempo (T): Representado por el día, la hora, el mes y el año de una publicación.
6. Características de popularidad (p): Representa dos atributos, el número de seguidores y el número de me gusta que contiene la página de la empresa en revisión.

4.4 | Representación

Una vez que hemos definido los tipos de características, el siguiente paso según la figura 4.1 consiste en seleccionar el modelo de representación para incorporar los atributos de una publicación. En este trabajo representamos cada publicación en un vector multidimensional basado en la metodología de bolsa de palabras *(Bow)* y n-gramas con pesado *booleano* y *TF*, donde el número de dimensiones d corresponden al número total de características de una representación dada. Por ejemplo: si deseamos representar las publicaciones por medio de las características de popularidad (p), los vectores tendrán dos dimensiones una que representará el número de seguidores y otra el número de me gusta de una página. Así mismo, a cada uno de estos vectores v se le aplica un proceso de normalización; este proceso consiste en estandarizar todos los valores del vector v en un rango de 0 y 1. Esto con el fin de reducir el impacto de las diferencias entre rangos de distintos tipos de características.

4.5 | Clasificación

La última fase de nuestra metodología es la *clasificación* según la figura 4.1. En esta fase se entrenan con las representaciones un algoritmo de aprendizaje para cada problema de clasificación. Para esta etapa, utilizamos cuatro de los algoritmos más utilizados para la clasificación de textos. Los algoritmos que seleccionamos fueron Naïve Bayes, Árboles de Decisión (Decisions Trees), Máquina de soportes (SVM) con funciones de kernel y por último Vecinos más cercanos (k-NN).

Para predecir el impacto general de una publicación, generamos un modelo de clasificación por cada problema de predicción. En la siguiente sección describimos los experimentos realizados para obtener estos modelos, así como también la metodología que seguimos para incorporar estos a una herramienta de visualización web.

5

Experimentos y resultados

En esta sección se describen los principales experimentos y los resultados que se obtuvieron al predecir el impacto de una publicación de Facebook. Los experimentos se realizaron de forma independiente con dos conjuntos de datos filtrados (*FL*: ver la tabla 4.1): el primer conjunto contiene 13,561 publicaciones de 10 empresas y el segundo 21,071 publicaciones de 14 políticos mexicanos. En consecuencia, al final se obtuvieron dos series de algoritmos entrenados: una serie exclusivamente para predecir las seis métricas (reacciones totales, veces compartidas, comentarios, etc.) de las publicaciones de las empresas y otra únicamente para anticipar las métricas de las publicaciones de los políticos.

Para evaluar el rendimiento de cada una de las clasificaciones se utilizó la métrica de evaluación *F-macro* y para todos los experimentos se empleó una técnica de validación cruzada de 10 pliegues para calcular el rendimiento. Es importante destacar que, en los experimentos realizados tanto con las cuentas de las empresas y de los políticos, no se hizo ninguna distinción entre el nombre de una cuenta, ya que una de las metas de este proyecto es construir predicciones generales más que una predicción por cuenta.

A lo largo de este trabajo se ha considerado el contenido de una publicación como un atributo indispensable para predecir su impacto. Razón por la cual, para comprobar lo anterior, se realizaron los siguientes dos conjuntos de experimentos:

1. En el primer conjunto **(ExpA)**: Se utilizan en los experimentos los tipos de atributos (comportamiento, estilo, interacción, tiempo y popularidad) de forma individual, para representar a cada publicación. Esta serie de experimentos en general tiene como objetivo mostrar la importancia que tiene entrenar un algoritmo, solo con un tipo de atributos sin incluir el contenido del texto.

2. El segundo conjunto de experimentos **(ExpB)**: mezcla los tipos de atributos y el contenido de la publicación. Es decir, el conjunto de estos experimentos representa a cada publicación con los atributos de contenido mas algún otro atributo ya sea (estilo, comportamiento, popularidad, etc.). Esto con el fin de verificar si el contenido de una publicación mas algún otro atributo es sobresaliente, para predecir el impacto.

En la tabla 5.1 se muestra la configuración de cada uno de los experimentos realizados en ambos conjuntos, *(ExpaA, ExpB)*.

5.1 | Resultados de predicción de impacto en las empresas

En esta sección se visualizan y se analizan los resultados de los experimentos que se obtuvieron al predecir el impacto de las publicaciones de las diez empresas.

A continuación, se muestra en la figura 5.1 los resultados que se obtuvieron de cada uno de los experimentos del conjunto *ExpA* con los datos de las empresas. Para conocer mas a detalle los resultados de cada uno de los experimentos del conjunto *ExpA* (Ver tablas del anexo: A.1.1).

En la figura 5.1 se observa que el mejor algoritmo de aprendizaje es Árboles de decisión. Esto tiene sentido ya que el tamaño del vector de las representaciones para cada uno de los experimentos es muy pequeño (entre 4 y 5 características). Otro aspecto sobresaliente, es que el experimento *Exp2a(e)* que se entrena con atributos de estilo, es el segundo mejor para predecir las métricas de impacto, excepto para predecir las reacciones totales, en este caso el experimento *Exp5a(p)* que se entrena con atributos de popularidad es el segundo mejor.

Sin embargo, los mejores resultados del conjunto de experimentos *ExpA* para predecir las métricas de las empresas, ocurre cuando se usa una combinación de los 5 atributos *(c + e + i +t +p)*, en otras palabras los mejores resultados se obtienen con el *Exp6A(c + e + i +t +p)*. Mientras que los peores resultados se obtienen cuando se entrena con los atributos de interacción, comportamiento y tiempo.

En las figuras 5.2 y 5.3 se pueden ver los resultados que se obtuvieron al predecir las métricas con el segundo conjunto de experimentos *ExpB*. Cada uno de los experimentos incluye atributos extraídos del contenido de una publicación. Por lo tanto, para incluir

Tabla 5.1: Configuración experimental del conjunto *ExpA, ExpB*

Conjunto de experimentos	Experimentos:	
	Nombre del experimento	Descripción
ExpA	Exp 1a . Comportamiento(c)	Consiste en entrenar los algoritmos con los atributos *comportamiento* .
	Exp 2a. Estilo *(e)*	Consiste en entrenar los algoritmos con los atributos de *estilo* .
	Exp 3a. Interacción *(i)*	Consiste en entrenar los algoritmos con los atributos de *interacciones* .
	Exp 4a. Tiempo *(T)*	Consiste en entrenar los algoritmos con los atributos de *tiempo* .
	Exp 5a. Popularidad *(p)*	Consiste en entrenar los algoritmos con los atributos de *popularidad* .
	Exp 6a.- *(c + e + i + T + p)*	Consiste en entrenar los algoritmos con todos los atributos excepto el contenido del texto.
ExpB	Exp 1b. Contenido *(t)*	Consiste en entrenar los algoritmos empleando solo el *texto de las publicaciones* .
	Exp 2b (c + t)	Consiste en entrenar los algoritmos con los atributos de *comportamiento más el atributo de contenido* .
	Exp 3b. *(e + t)*	Consiste en entrenar los algoritmos con los atributos de *estilo más el atributo de contenido* .
	Exp 4b *(i + t)*.	Consiste en entrenar los algoritmos con los atributos de *interacciones más el atributo de contenido* .
	Exp 5b.*(T + t)*	Consiste en entrenar los algoritmos con los atributos de *tiempo más el atributo de contenido* .
	Exp 6b,*(p + t)*	Consiste en entrenar los algoritmos con los atributos de *popularidad más el atributo de contenido* .
	Exp 7b *(c + e + i + T + p + t)*	Consiste en entrenar los algoritmos *todos los atributos*.

Figura 5.1: Resultados del conjunto de los experimentos *ExpA* con los datos de las *empresas*

el texto en los experimentos, se utilizó un enfoque que representa el texto como una bolsa de palabras con 10,000 dimensiones ya sea de bi-gramas, tri-gramas o de palabras.

En cada gráfica de las figuras 5.2, 5.3 se incluye una línea continua, esta indica el mejor rendimiento del conjunto anterior de experimentos *ExpA* (es decir de la figura 5.1). Para conocer más a detalle los resultados de los experimentos del conjunto (ver las tablas en el anexo A.1.2).

Como análisis de los resultados de las figuras 5.2, 5.3 se nota que todos los experimentos que usan el atributo de *contenido* en combinación con el atributo *popularidad* superan los mejores resultados de los experimentos *ExpA*. Sin embargo, al utilizar solo

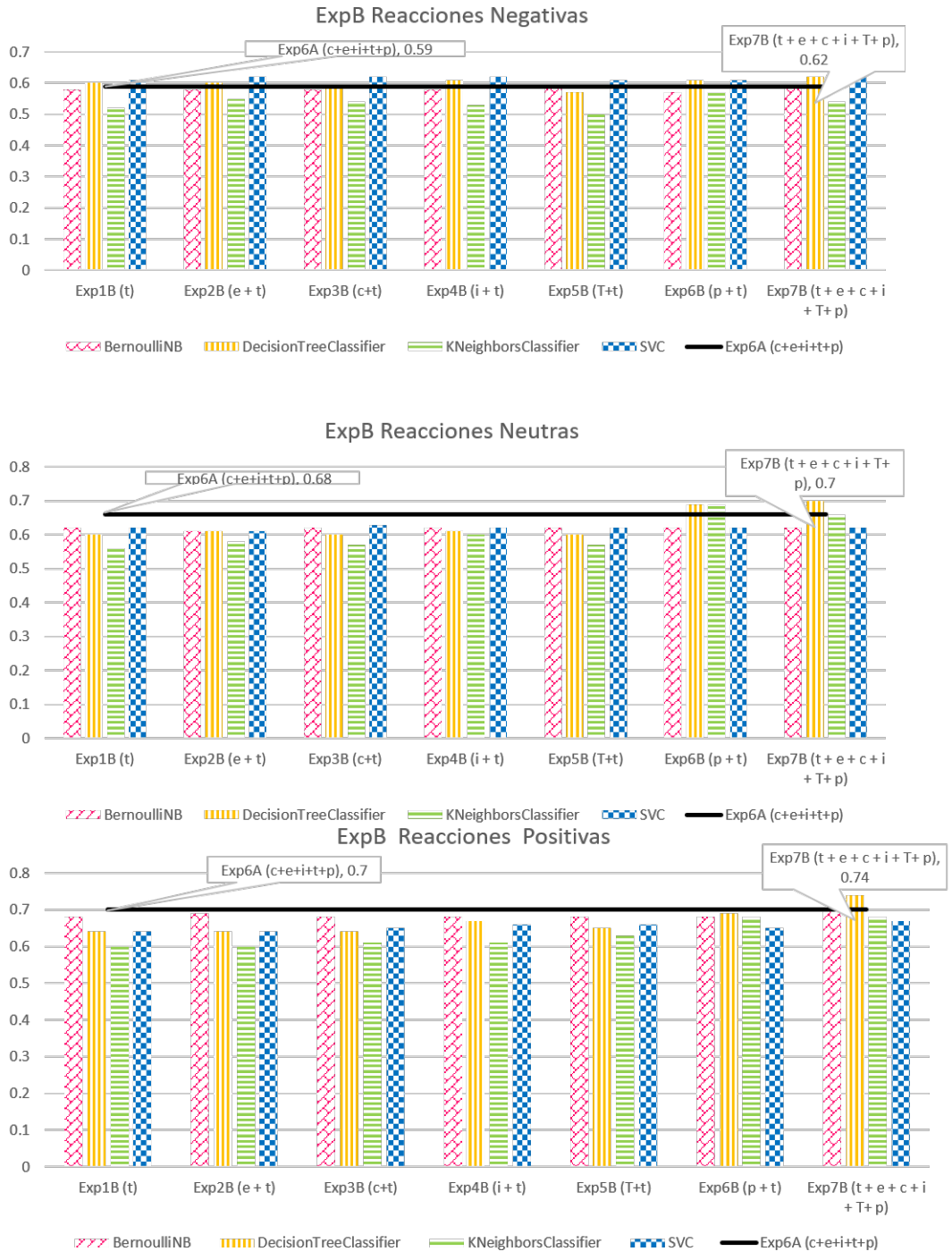

Figura 5.2: Resultados del conjunto de los experimentos *ExpB* con los datos de las *empresas*

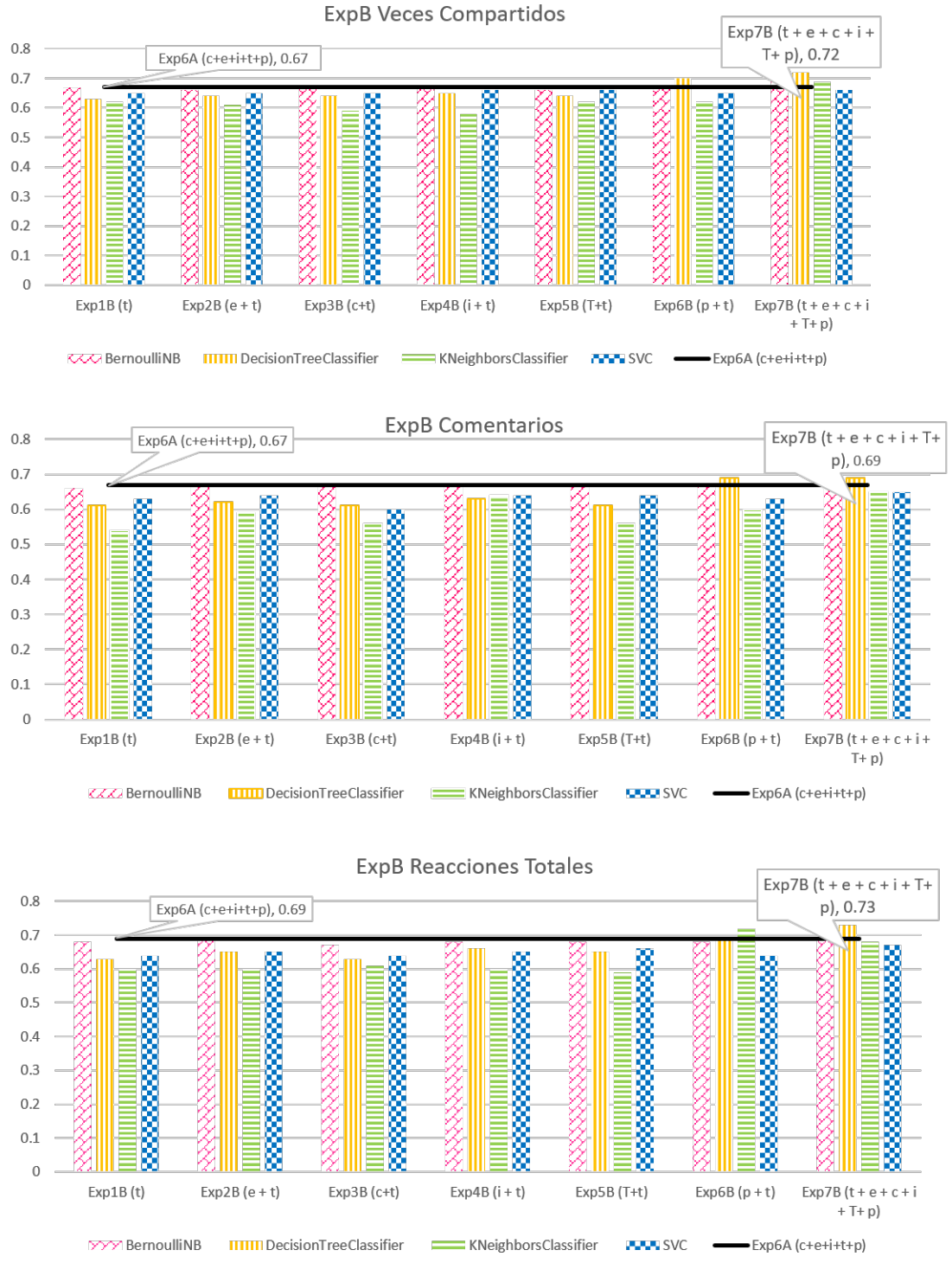

Figura 5.3: Resultados del conjunto de los experimentos *ExpB* con los datos de las *empresas*

el *contenido* no es posible superar a dichos resultados de la figura *ExpA*. Por otro lado, el mejor algoritmo que predice las seis métricas es *Árboles de Decisiones*, cuando este se entrena con la combinación de todos los atributos propuestos, dicho de otra manera el mejor experimento es *Exp7B(t,c,e,i,T,p)* con *Arboles de Decisiones*.

Desde otro punto, si comparamos los mejores resultados del conjunto *ExpA* y *ExpB* (ver figura 5.4), decimos que en ambos casos la predicción más baja se obtuvo al predecir las reacciones negativas. Esto quiere decir que predecir las reacciones negativas es una tarea difícil. Por el contrario, los mejores resultados se obtuvieron al predecir las reacciones positivas, el total de reacciones y veces compartidos, mientras que con menor eficiencia los comentarios. Para finalizar, en la figura 5.4 podemos observar en particular que para predecir las reacciones negativas, veces compartidos y reacciones totales. El hecho de incluir el atributo texto tiende a mejorar el rendimiento de las predicciones de estas métricas. Esto significa que usar el contenido de una publicación es relevante para predecir las distintas métricas de impacto de una publicación.

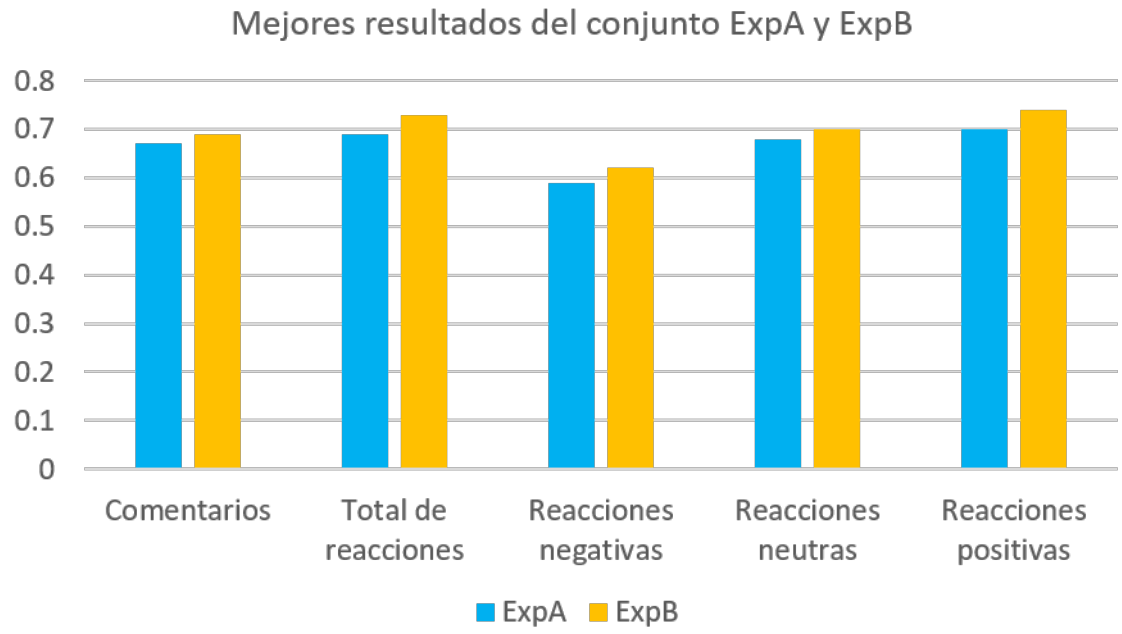

Figura 5.4: Mejores resultados del conjunto de *ExpA* y *ExpB* para predecir las métricas de las *empresas*

5.2 | Resultados de predicción de impacto en los políticos

En esta sección se muestran los resultados y el análisis correspondiente a los experimentos realizados del conjunto *ExpA* y *ExpB* para predecir las métricas de las publicaciones de los políticos.

En la figura 5.5 se muestran los resultados obtenidos de cada uno de los experimentos del conjunto *ExpA* para predecir las seis métricas de los *políticos* . Para conocer más a detalle los resultados de cada uno de los experimentos del conjunto *ExpA* de los políticos (Ver tablas del anexo: A.2.1).

Dada la figura 5.5 se observa que el mejor algoritmo de clasificación en todos los experimentos y en cada una de las métricas de los políticos es *Arboles de Decisiones*. Otra cosa sobresaliente, es que el *Exp5b(p)* que consiste en entrenar a un algoritmo con representaciones basadas en atributos de *popularidad* es el segundo mejor experimento para predecir las métricas de veces compartidos, comentarios, reacciones positivas y total de reacciones. Mientras que el experimento *Exp2(e)* (entrena solo con atributos de estilo) y el *Exp2(t)* (entrena solo con atributos de tiempo) son los segundos mejores resultados que predicen las reacciones negativas y neutras. Dicho lo anterior, podemos inferir que los atributos de popularidad, estilo y tiempo ayudan en parte a las predicciones del las métricas de los políticos. A diferencia de los atributos de interacción y comportamiento que muestran no ser tan relevantes, al igual que en las predicciones de las métricas de las empresas.

Sin embargo, los mejores resultados para predecir las seis métricas de los políticos con el conjunto de experimentos *ExpA*, se consiguieron cuando se entrena al algoritmo *Árboles de Decisión* con la combinación de los cinco atributos: comportamiento, estilo, interacción, tiempo y popularidad (es decir con el *Exp6a(c,e,i,t,p)*. Otro aspecto importante por mencionar es que, los atributos de popularidad han mostrado ser ventajosos para predecir las diferentes métricas de los políticos. Mientras que para predecir las métricas de las empresas no lo son, ver en figura 5.1.

A continuación, en las figuras 5.6 y 5.7 se muestran los mejores resultados que se obtuvieron con el segundo conjunto de experimentos *ExpB* al predecir las distintas métricas (Ver tablas del anexo: A.1.2). Cabe mencionar, que cada uno de los experimentos incluye los atributos de contenido de una publicación. Así mismo, con el fin de comparar el rendimiento de los resultados del conjunto *ExpB* con respecto al conjunto *ExpA*;

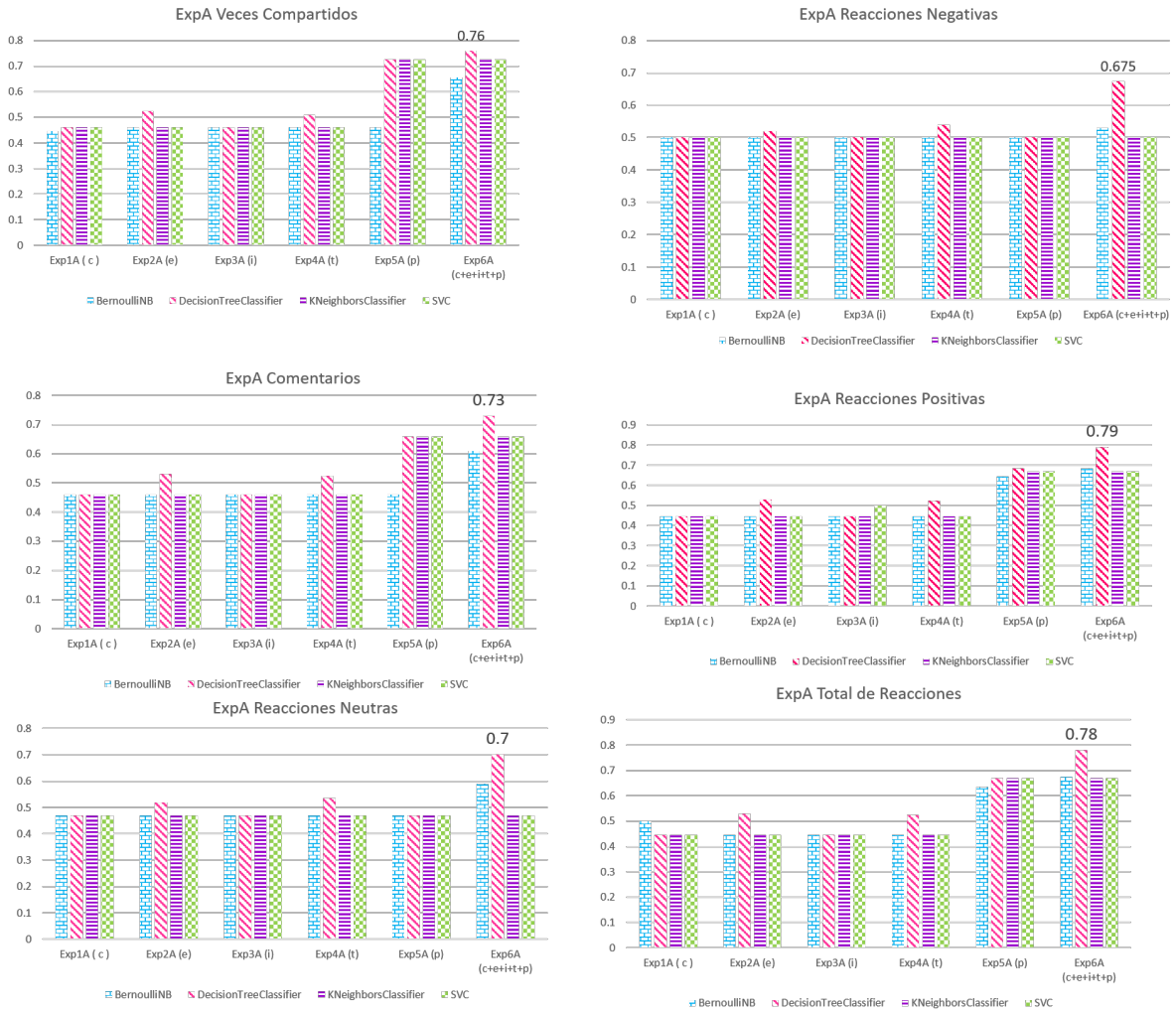

Figura 5.5: Resultados del conjunto de los experimentos *ExpA* con los datos de los *políticos*

Se incluye en cada gráfica de las figuras 5.6 y 5.7 una línea continua, esta indica el mejor rendimiento del conjunto anterior de experimentos *ExpA* (es decir, de la figura 5.5).

Como resultado de las figuras 5.6 y 5.7 se observa que al igual que en la figura 5.5, el mejor algoritmo en cada experimento y para cada métrica es *Árboles de Decisión*. Así mismo, se nota que el experimento *ExpB7(t,c,e,i,T,p)* que usa el atributo de contenido en combinación con todos los atributos planteados (contenido, comportamiento, estilo, iteración, tiempo y popularidad), es el único que supera los mejores resultados de los experimentos *ExpA*. Esto indica que los mejores resultados para predecir las métricas de los políticos se alcanzan con el experimento *ExpB7(t,c,e,i,T,p)*.

Por otro lado, el *Exp6B(t,p)* que representa el texto en combinación con los atributos de popularidad, es el segundo mejor experimento para todas las métricas. Esto quiere decir que, el atributo popularidad en combinación con el contenido son favorables para predecir gran parte de las métricas. Al contrario de los atributos de comportamiento, estilo, tiempo e interacción que al combinarse con el atributo contenido o incluso el mismo atributo contenido por sí solo, no muestra ser relevantes ante las predicciones de las métricas de los políticos.

Desde otra perspectiva si comparamos los mejores resultados del conjunto de experimentos *ExpA* y *ExpB* (ver figura: 5.2) podemos decir que en ambos casos las predicciones más baja se obtuvieron al predecir las reacciones negativas y neutras. Esto quiere decir que predecir estas dos métricas es una tarea compleja. Por el contrario, los mejores resultados se obtuvieron al predecir las reacciones positivas, el total de reacciones, veces compartidos y con un menor resultado los comentarios. Para concluir, podemos expresar que el hecho de incluir el atributo contenido al predecir las métricas de los políticos no resulta ser un atributo con gran relevancia. Como lo es para predecir las métricas de las empresas.

5.3 | Conclusiones generales de las predicciones de las empresas y políticos

Como se ha mencionado, en las dos secciones anteriores se realizaron dos conjuntos de experimentos (*ExpA* y *ExpB*) para predecir las métricas de las empresas y de los políticos. En ambos casos los mejores resultados para predecir las métricas se obtuvieron al entrenar al algoritmo *Arboles de Decisiones* con todos los atributos propuestos (contenido, comportamiento, estilo, interacciones, tiempo y popularidad) es decir, con el

Figura 5.6: Resultados del conjunto de los experimentos *ExpB* con los datos de los *políticos*

Figura 5.7: Resultados del conjunto de los experimentos *ExpB* con los datos de los *políticos*

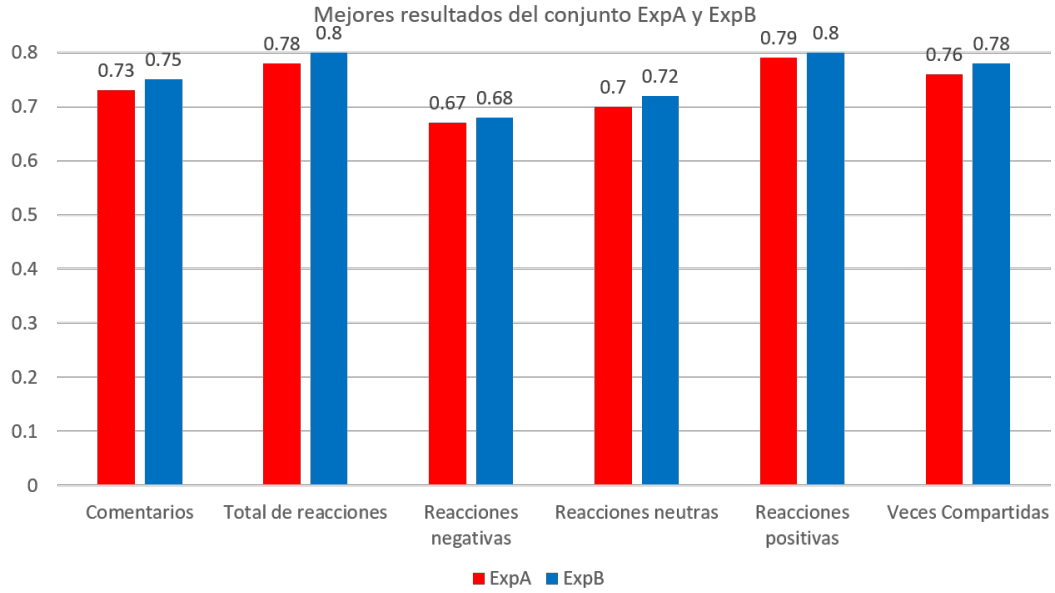

Figura 5.8: Mejores resultados del conjunto de los experimentos *ExpA* y *ExpB* con los datos de los *políticos*

experimento *Exp7b(t+c+e+i+T+p)*. Sin embargo, a pesar de que los mejores resultados de las métricas de las empresas y políticos se obtuvieron con el mismo experimento, los resultados de las métricas de los políticos tienden a ser mucho mejores, que los de las empresas (ver la figura 5.9). Esto se debe a que el conjunto de datos de los políticos es más grande que el de las empresas, por lo tanto, los clasificadores de los experimentos de los políticos tuvieron más publicaciones de las cuales aprendieron.

En cambio, también al observar la figura 5.9 se dice que la predicción más baja en las empresas y en los políticos se obtuvo al predecir las reacciones negativas. Por el contrario, los mejores resultados se obtuvieron al predecir las reacciones positivas, el total de reacciones y veces compartidos.

Por otra parte, una vez que se identificó que el experimento *Exp7b(t+c+e+i+T+p)* da los mejores resultados para predecir las métricas, se crearon los modelos de clasificación de cada métrica de los políticos y de las empresas, con todos los atributos propuestos. Así mismo, en la siguiente sección describimos la metodología que seguimos para incorporar estos modelos a una herramienta de visualización web, que predice el impacto que tendrá una publicación de Facebook.

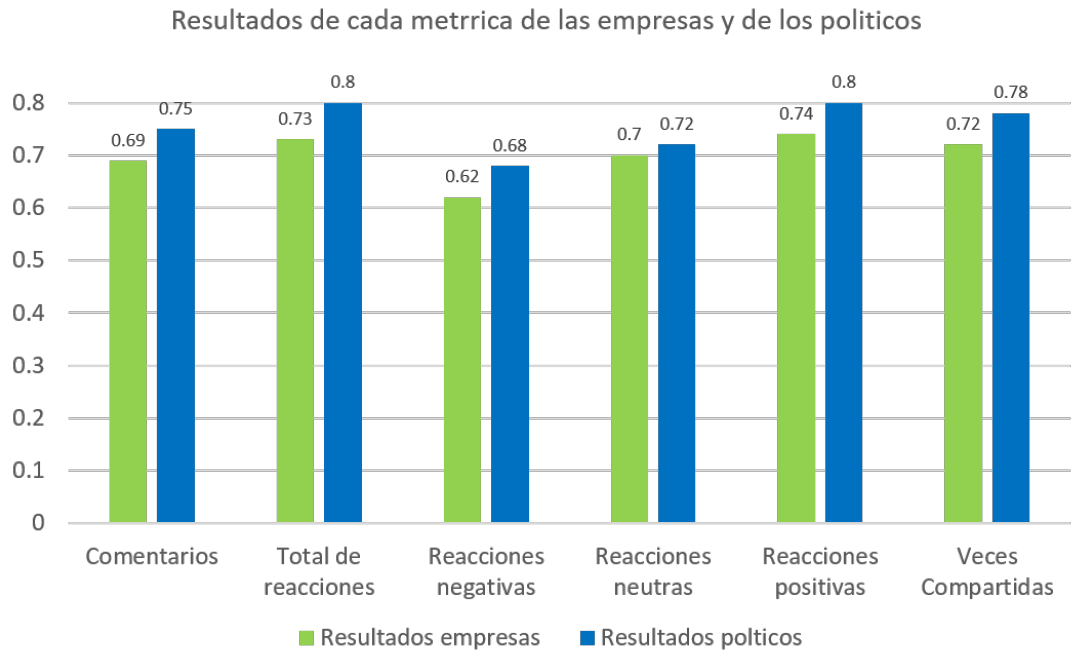

Figura 5.9: Mejores resultados del conjunto de las métricas de las empresas y políticos

6

Desarrollo del sistema

En este capítulo se describe el el desarrollo y el funcionamiento de la aplicación web. La cual tiene como principales funciones: permitir redactar una publicación de Facebook dada por un usuario, predecir el impacto que tendrá la publicación y mostrar de una manera visual dicha predicción. Así mismo, en este apartado se relata detalladamente los cuatro módulos (carga, extracción de características, predicciones y visualización) que integran la aplicación web.

A continuación, en la figura 6.1 se muestra la arquitectura general y cada uno de los módulos que completan a la aplicación web. Los módulos se describen en secciones posteriores.

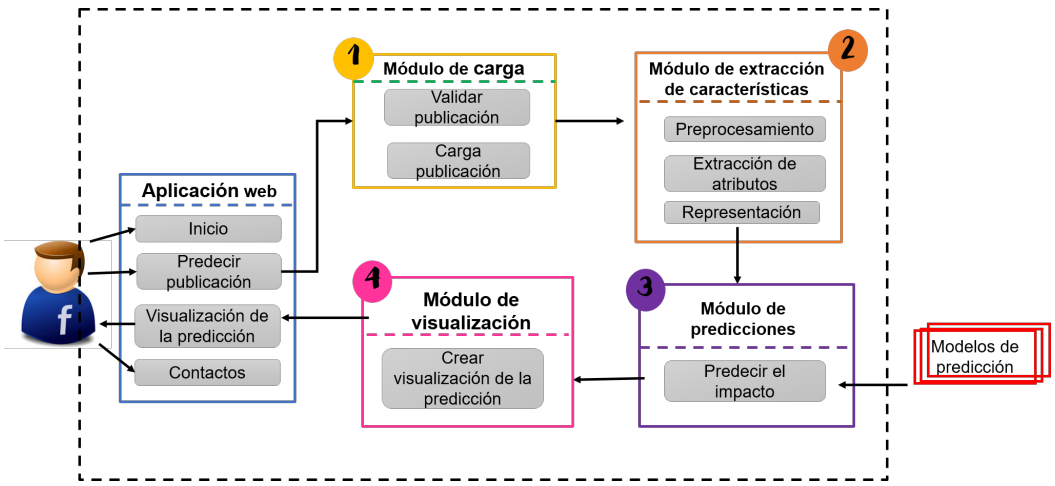

Figura 6.1: Arquitectura general del sistema web

En la figura 6.1, se observa un rectángulo color azul que tiene como título *Aplicación*

web. En este rectángulo se muestran las cuatro páginas que tiene el usuario para interactuar con la aplicación. Las funciones de las páginas son:

- Inicio (Página principal): El usuario puede observar la información e instrucciones relevantes de como interactuar con la aplicación.
- Predecir publicación: Permite al usuario capturar la publicación que desea predecir. Así mismo, está página se asocia con los cuatro módulos que componen a la aplicación, con el objetivo de predecir el impacto que tendrá una publicación.
- Visualización: Esta página muestra de forma visual el impacto que tendrá una publicación. Por lo tanto, esta página es visible cada vez que el usuario solicita ver el impacto de una publicación.
- Contactos: Esta página muestra la información de cada uno de los participantes en el proyecto.

6.1 | Módulos de carga

Este módulo recibe como entrada una cadena de texto (que contiene el texto de la publicación) y un conjunto de variables (que contienen atributos de la publicación como el día, la hora, el año, etc.). Una vez que este módulo recibe dichos datos, se encarga de validar que realmente éstos vengan en el formato adecuado. Si los datos son válidos, se cargan en memoria para después ser utilizados por los demás módulos. El proceso de este módulo se puede observar en la figura 6.2.

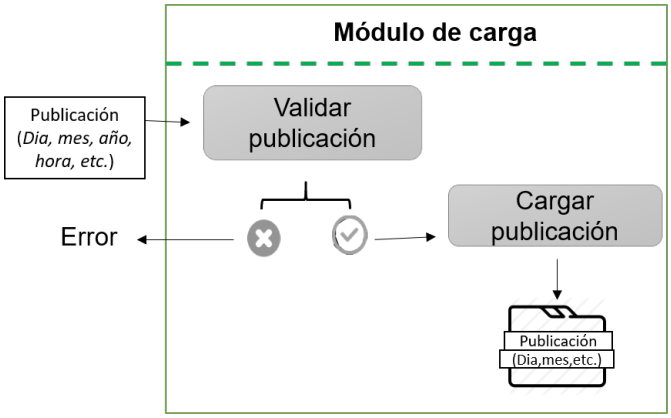

Figura 6.2: Módulo de carga

6.2 | Módulo de extracción de características

Una vez que los datos de la publicación a predecir están cargados en memoria, se realizan las siguientes tres tareas de este módulo:

1. Preprocesamiento: Se sustituyen las URL, menciones, hashtag y emojis que contenga el texto de la publicación, por las etiquetas <url>, <menciones>, <hashtag> y <emojis>.

2. Extracción de atributos: Realizado el preprocesamiento se extraen de los datos de la publicación, los atributos de estilo, comportamiento, iteración, popularidad, contenido y tiempo. Al final, por cada conjunto de atributos, los cuales se describen en la sección 4.3 tenemos un vector numérico.

3. Representación: Se hace la unión de todos los vectores numéricos pertenecientes a cada conjunto de atributos, con el fin de generar un solo vector.

En la figura 6.3 se muestran gráficamente las tareas de este módulo.

6.3 | Módulo de predicciones

Este módulo tiene como tarea predecir el impacto (alto o bajo) de cada una de las seis métricas ($|R|$, $|VC|$, $|C|$, $|R-|$, $|R+|$ y $|R\odot|$). Para lograr esto, este módulo recibe la representación vectorial de la publicación a predecir, sucesivamente esta representación es usada para alimentar a los seis modelos ya generados. Al concluir, se obtuvieron seis predicciones, que pertenecen al impacto de alguna de las seis métricas usadas en este proyecto. Estos pronósticos son almacenados en memoria, para ser utilizados en el siguiente módulo. En la figura 6.4 se muestra gráficamente lo antes mencionado.

6.4 | Módulo de visualización

La tarea de este módulo es recibir los resultados de las predicciones de cada una de las métricas. Sucesivamente trasformar los resultados en una visualización, específicamente en una gráfica de barras, como la que se muestra en la figura 6.5. El funcionamiento de este módulo se encuentra gráficamente expresada en la figura 6.6.

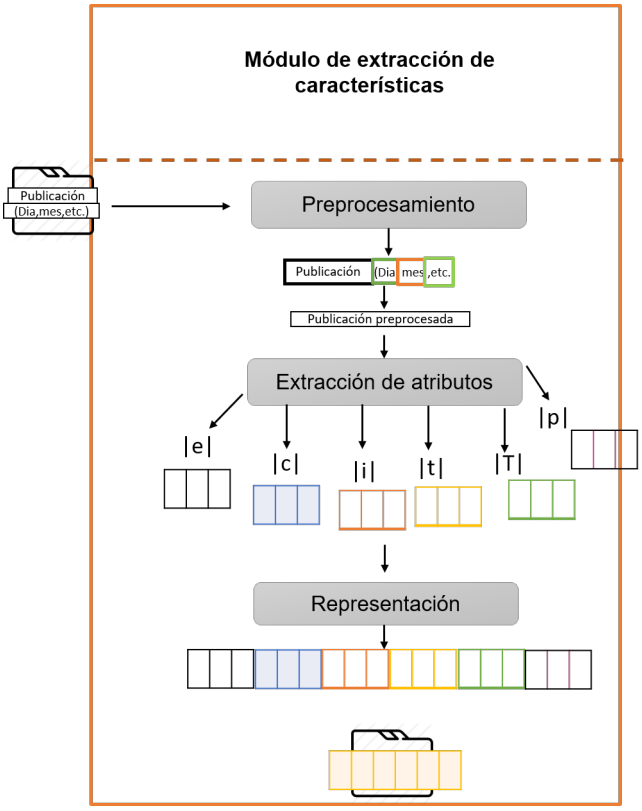

Figura 6.3: Módulo de extracción de Características

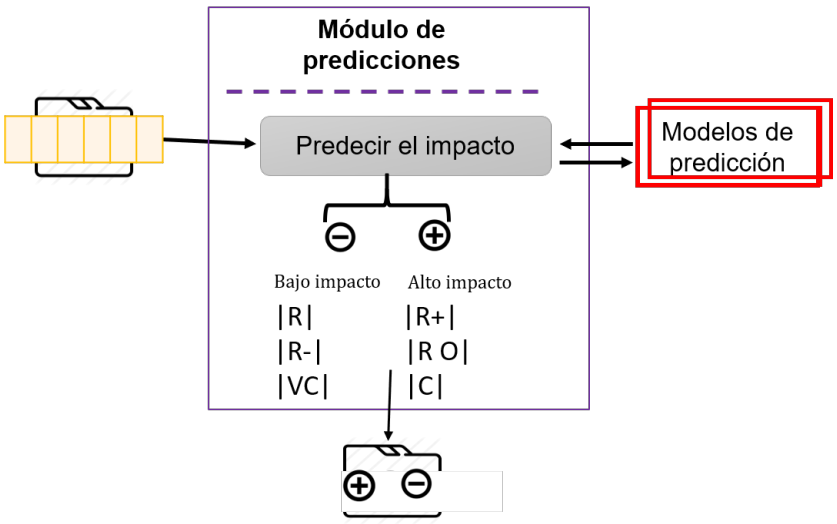

Figura 6.4: Módulo de predicciones

En la figura 6.5 se puede observar una gráfica con 6 barras horizontales, el tamaño de cada barra representa el impacto alto o bajo que tendrá cada una de las métricas ($|R|$, $|VC|$, $|C|$, $|R-|$, $|R+|$ y $|R\odot|$).

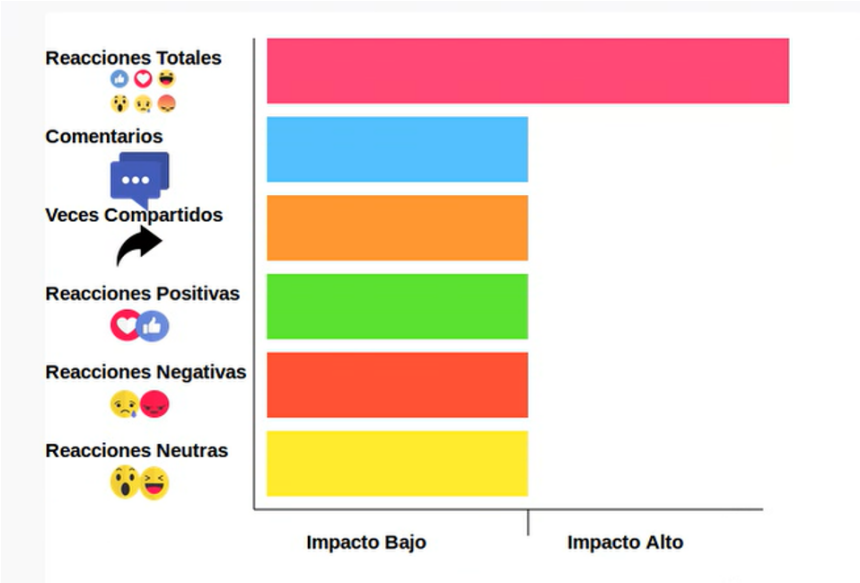

Figura 6.5: Representación visual de la predicción del impacto de una publicación de Facebook

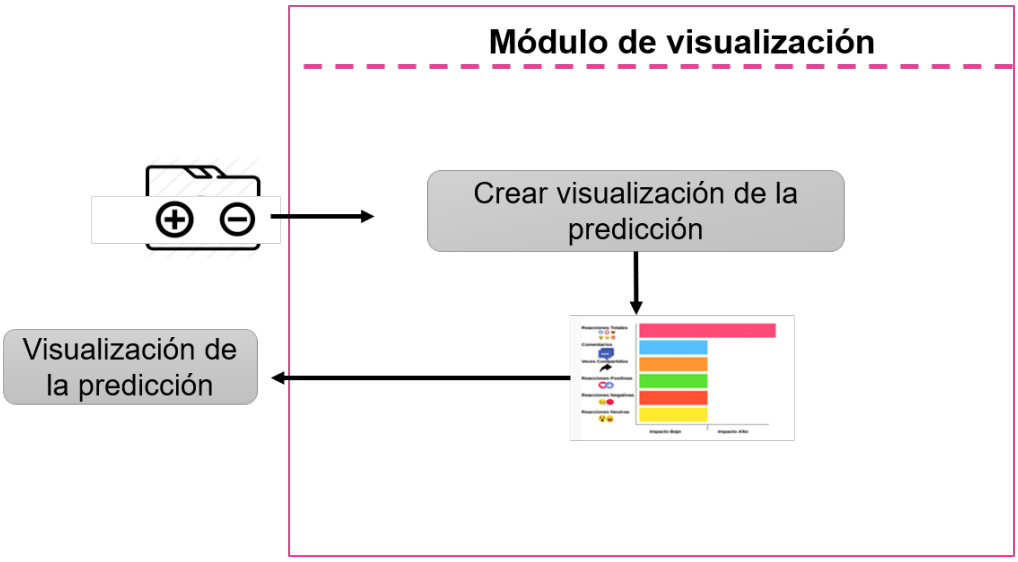

Figura 6.6: Módulo de visualización del la aplicación web

6.5 | Vista del sistema

En el siguiente apartado se muestran algunas capturas de pantallas de las páginas de la aplicación web. Así mismo, en esta sección se expone paso a paso como predecir el impacto de una nueva publicación de Facebook.

A continuación, en la figura 6.7 se muestra la página principal de la aplicación web, esta página tiene como objetivo incentivar a los visitantes a profundizar en el sitio. En

Figura 6.7: Pagina principal del sistema web. La siguiente imagen muestra la pantalla principal de la aplicación. Dicha pantalla contiene la información relevante sobre como predecir y medir el impacto de una publicación de Facebook. Así mismo, contiene un menú de navegación con tres vínculos: Predecir una publicación, información sobre el sistema y contactos. También la página principal contiene una serie de botones, el más relevante es *predecir el impacto*, este botón redirecciona al usuario a la pantalla de *predecir publicación* figura 6.10.

la figura 6.8 se muestra la página de contactos, la cual muestra la información de cada uno de los desarrolladores del sitio web.

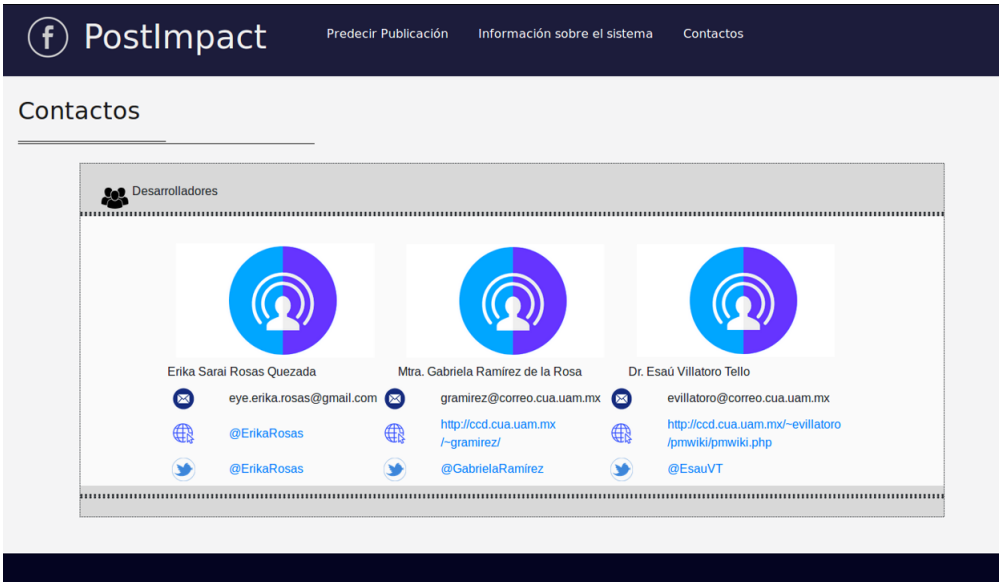

Figura 6.8: Página de contactos. En esta página se puede ver la información de contacto de cada uno de los miembros que colaboraron en este proyecto.

6.5.1 | Predecir el impacto de una publicación paso a paso

Para predecir el impacto de una nueva publicación debe de llevarse a cabo los siguiente:

1. Entrar a la página principal de la aplicación (ver página en la figura 6.7).
2. Dirigirse al menú de navegación de la página principal y dar clic sobre el enlace *predecir publicación*. Este enlace mostrará la página de la figura 6.9. El objetivo de esta página es, preguntar si la publicación que se desea predecir pertenece a un político o una empresa.
3. Una vez que se seleccionó alguna de las opciones, entonces se mostrará la página *Publicación a predecir* (ver página en figura 6.10). Esta página tiene como función capturar a través de un formulario, cada uno de los componentes que tendrá la nueva publicación. Los elementos que deben ser capturados en el formulario son:
   - Texto: En la parte superior izquierda de la página *Publicación a predecir*, se muestra una caja de texto. En esta se debe de escribir el texto, los emojis, los hashtags y los links que tendrá la publicación.
   - Número de vídeos, Número de links, Número de imágenes y Número de álbumes: En el recuadro correspondiente se debe de indicar el número absoluto de videos, links, imágenes y álbumes que tendrá la publicación.

Figura 6.9: Pagina de cuestiona miento entre empresa o político

- Número de seguidores y Número de me gusta: Se debe de indicar el número absoluto de seguidores y de me gusta que tiene la cuenta en donde se publicará el post.
- Hora, día, mes y año: Se debe de indicar en los recuadros, la hora, el día, mes y el año en el que se estima que se publicará.

4. Una vez que se llenó el formulario dar clic en el botón “predecir”, posteriormente, esperar a que la aplicación muestre la página *visualización* (ver página en de figura 6.11). En esta página se muestra gráficamente el impacto alto o bajo que tuvo cada una de las métricas (total de reacciones, comentarios, veces compartidos, reacciones negativas, positivas y neutras).

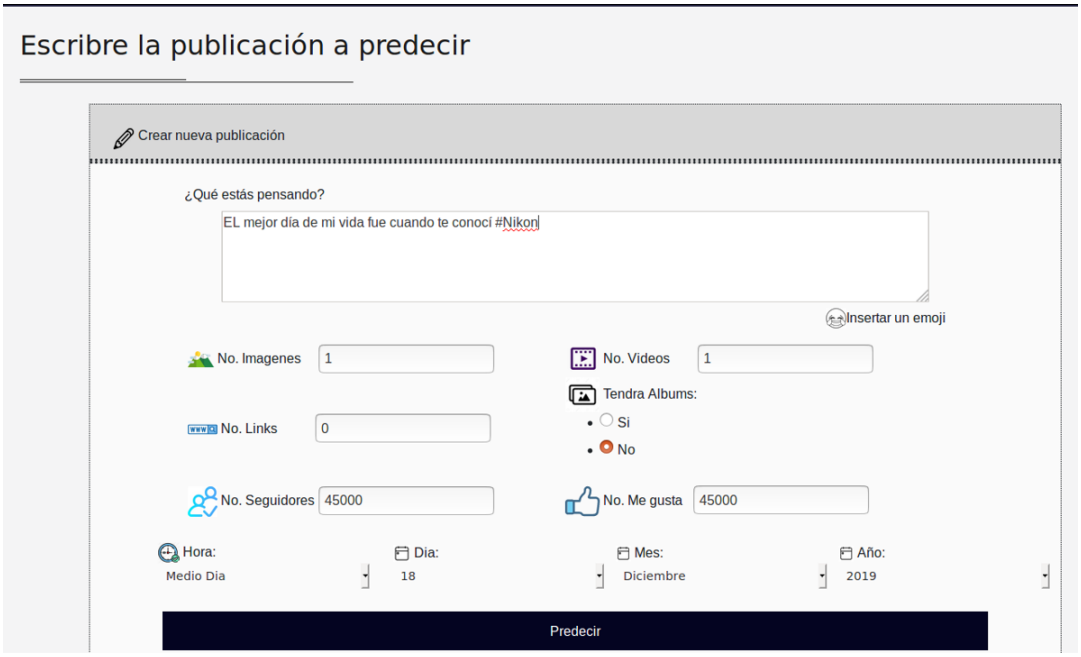

Figura 6.10: Página para capturar una publicación. La siguiente página muestra un formulario, a través del cual el usuario captura la publicación que desea predecir.

Figura 6.11: Página de visualización de resultados. En la siguiente página, se muestran gráficamente los resultados al predecir el impacto de una publicación en Facebook.

7

Conclusiones y trabajo futuro

Al principio de este trabajo se planteó como objetivo general, desarrollar una herramienta web que permita predecir el impacto que producirá una publicación de Facebook, a partir de emplear técnicas de aprendizaje computacional. Al final de esta investigación nuestro objetivo general se logró cumplir exitosamente. Ya que ahora, el resultado de esta investigación es una aplicación web que incorpora una serie de modelos predictivos capaces de anticipar el impacto de una publicación de Facebook.

Cabe mencionar, que el impacto de una publicación de Facebook según este trabajo, se mide a través del volumen que contenga una publicación en cada una de las siguientes métricas: Total de reacciones, reacciones negativas, reacciones positivas, reacciones neutras, comentarios y veces compartidos. Por lo tanto, la herramienta web predice el impacto alto o bajo de las seis métricas anteriores.

Para lograr el funcionamiento de la aplicación web, se requirió crear para cada métrica un modelo predictivo, basado en la clasificación supervisada de textos. Cada modelo es capaz de anticipar el impacto de una métrica en específico. Así que, para crear estos modelos fue necesario alcanzar cada uno de los objetivos específicos de este trabajo. Uno de los objetivos alcanzables consistió en: identificar aquellos atributos de una publicación que son relevantes para predecir el impacto. Por lo tanto, al principio de este trabajo, se planteó como hipótesis que el contenido textual (*el qué se escribe*) de una publicación es importante para predecir el impacto.

Para validar lo antes mencionado y dada la falta de un corpus estándar para evaluar este tipo de enfoques, asumimos como tarea extra recopilar y estandarizar 37, 585 publicaciones de Facebook, 14, 522 pertenecen a diez cuentas de empresas de renombre en México y 23, 063 publicaciones a catorce cuentas de políticos mexicanos con una

aspiración a un cargo público en el año 2018. Con el conjunto de datos se crearon dos subconjuntos, uno con publicaciones de empresas y otro con publicaciones de políticos.

Las diferencias más notables de ambos conjuntos son que: El tamaño del conjunto de datos de las empresas es mucho menor que el conjunto de los políticos. Por lo tanto, se infiere que sería relevante recopilar más publicaciones de empresas, para así mejorar los resultados de las predicciones. Otro aspecto relevante, es que las publicaciones de las empresas tienden a tener menos cantidad de texto, pero contienen más emojis y links que las publicaciones de los políticos. En ambos conjuntos las reacciones negativas son menores al número de reacciones positivas y neutras.

Teniendo ambos conjuntos de publicaciones (políticos y empresas), se procedió a alcanzar otro de los objetivos. El cual consistía en realizar una serie de experimentos para evaluar la pertinencia de los atributos identificados. Los resultados de los experimentos indican que:

- Entrenar a un clasificador de *Árboles de decisión*, con el *qué y cómo* se escribe una publicación, en combinación con todos los atributos propuestos (estilo, iteración, comportamiento, tiempo y popularidad) es relevante para predecir el impacto de las publicaciones de las empresas y de los políticos.
- Los resultados más bajos de los experimentos se obtienen al predecir las reacciones negativas, mientras que los más altos se obtienen al predecir el total de reacciones y reacciones positivas.
- Los resultados de los experimentos son mejores al predecir las métricas de los políticos, que al predecir las métricas de las empresas.

El último de los objetivos alcanzados fue crear con base en los mejores resultados los modelos predictivos e incorporarlos en una aplicación. Para concluir, inferimos que la aplicación web, será una herramienta que puede traer grandes ventajas a la hora de decidir publicar algo en las plataformas de redes sociales, pues ahora el gestor de contenidos de una determinada empresa o político puede determinar el impacto promedio, dado el impacto predicho de los comentarios, las veces compartidas, las reacciones totales, así como las reacciones positivas, las reacciones negativas y las reacciones neutras de una publicación.

Como trabajo futuro surgieron las siguientes ideas: Los modelos propuestos podrían enriquecerse con otras características estilísticas y de contenido. Con respecto al

contenido, sería interesante incorporar algunas características basadas en temas, como LDA o representaciones de segundo orden. Finalmente, existe cierta evidencia sobre la relevancia de detectar el sentimiento de la publicación como una característica, pensamos que sería sobresaliente evaluar qué tan beneficioso podría ser incorporar este tipo de características en nuestros modelos.

8. Referencias

[1] Smith Kit. *116 Estadísticas interesantes de las redes sociales*. 2019. URL: https://www.brandwatch.com/es/blog/116-estadisticas-de-las-redes-sociales/. (accessed: 22.09.2019).

[2] Ali Abdallah Alalwan, Nripendra P Rana, Yogesh K Dwivedi, and Raed Algharabat. "Social media in marketing: A review and analysis of the existing literature". In: *Telematics and Informatics* 34.7 (2017), pp. 1177–1190.

[3] Sérgio Moro, Paulo Rita, and Bernardo Vala. "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach". In: *Journal of Business Research* 69.9 (2016), pp. 3341–3351.

[4] Carsten D Schultz. "Proposing to your fans: Which brand post characteristics drive consumer engagement activities on social media brand pages?" In: *Electronic Commerce Research and Applications* 26 (2017), pp. 23–34.

[5] Alexandra Amado, Paulo Cortez, Paulo Rita, and Sérgio Moro. "Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis". In: *European Research on Management and Business Economics* 24.1 (2018), pp. 1–7.

[6] Enrique Bonsón, Sonia Royo, and Melinda Ratkai. "Citizens' engagement on local governments' Facebook sites. An empirical analysis: The impact of different media and content types in Western Europe". In: *Government Information Quarterly* 32.1 (2015), pp. 52–62.

[7] Qin Gao and Chenyue Feng. “Branding with social media: User gratifications, usage patterns, and brand message content strategies”. In: *Computers in Human Behavior* 63 (2016), pp. 868–890.

[8] Sérgio Moro and Paulo Rita. “Brand strategies in social media in hospitality and tourism”. In: *International Journal of Contemporary Hospitality Management* 30.1 (2018), pp. 343–364.

[9] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math; (March 1, 1997), Mar. 1997, pp. 1–3. ISBN: 0070428077.

[10] Leticia C. Cagnina. “ Representación de Documentos”. In: (2018), pp. 1–3.

[11] Nadia Patricia Araujo Arredondo. “Método Semisupervisado para la Clasificación Automática de Textos de Opinión”. In: (2009), pp. 11–13.

[12] Roque Enrique López Condori. “Método de Clasificación Automática de Textos basado en Palabras Claves utilizando Información Semántica: Aplicación a Historias Clínicas”. In: (2014), pp. 10–11.

[13] Ramos Mázquez Juan Carlos. “ Detección de acoso en mensajes de Twitter.” In: (2017), pp. 25–26.

[14] Leticia C. Cagnina. “ Representación de Documentos”. In: (2018), pp. 4–5.

[15] Alfaro Cardenas Pablo Juan Olivares. “ Clasificación automática de textos usando redes de palabras”. In: *Revista Signos* (2018), p. 3.

[16] Avinash Navlani. *Decision Tree Classification in Python*. URL: `https://www.datacamp.com/community/tutorials/decision-tree-classification-python`. (accessed: 24.09.2019).

[17] Aprende Machine Learning. *Clasificar con K-Nearest-Neighbor ejemplo en Python*. URL: `https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/`. (accessed: 24.09.2019).

[18] Aditya Mishra. *Metrics to Evaluate your Machine Learning Algorithm*. URL: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. (accessed: 25.09.2019).

[19] Juan Manuel Cabrera y Fabricio O. erez. " Clasificación de Documentos usando Naive Bayes Multinomial y Representaciones Distribucionales". In: (), p. 2.

[20] Clarabridge. *What is Customer Engagement?* 2019. URL: https://www.clarabridge.com/customer-experience-dictionary/customer-engagement/#. (accessed: 01.10.2019).

[21] Ferran Sabate, Jasmina Berbegal-Mirabent, Antonio Cañabate, and Philipp R Lebherz. "Factors influencing popularity of branded content in Facebook fan pages". In: *European Management Journal* 32.6 (2014), pp. 1001–1011.

[22] Ana Teresa Silva, Sérgio Moro, Paulo Rita, and Paulo Cortez. "Unveiling the features of successful eBay smartphone sellers". In: *Journal of Retailing and Consumer Services* 43 (2018), pp. 311–324.

[23] Tae Yano and Noah A Smith. "What's worthy of comment? Content and comment volume in political blogs". In: *Fourth International AAAI Conference on Weblogs and Social Media*. 2010.

A

Anexos

La secuencia de tablas de este anexo A.1, muestran los resultados que se obtuvieron para predecir las seis métricas (comentarios, veces compartidos, total de reacciones, reacciones positivas, neutras y negativas) de los políticos y de las empresas. Cada una de las tablas representa los resultados de una métrica. Así mismo, la primera columna de las tablas contiene el nombre del algoritmo *(BernoulliNB* ***(NB)****, DecisionTreeClassifier* ***(DT)****, KNeighborsClassifie* ***(KNN) y SVC)*** *)* con el que se realizó el experimento, sucesivamente las siguientes columnas representan el experimento realizado y debajo de cada una de estas columnas, se encuentran tres valores; los dos primeros contienen el valor **F-score** que se obtuvo al predecir la clase Alto Impacto **(+)** y Bajo impacto **(-)**. Mientras que el valor siguiente, representa la evaluación en F-macro **(FM)**, de ambas clases de cada uno de los resultados de los experimentos.

A.1 | Resultados al predecir las métricas de las empresas

A.1.1 | Resultados con el conjunto de experimentos **ExpA**

Tabla A.1: Resultados de los experimentos *ExpA* para predecir la métrica **comentarios** de las publicaciones de las *empresas*

Comentarios																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0.2	0.92	0.56
DT	0.03	0.93	0.48	0.25	0.87	0.56	0	0.93	0.465	0.01	0.92	0.465	0	0.93	0.465	0.42	0.91	0.665
KNN	0	0.93	0.465	0	0.92	0.46	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465
SVM	0	0.93	0.465	0	0.92	0.46	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465

Tabla A.2: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Totales** de las publicaciones de las *empresas*

Reacciones Totales																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0.33	0.88	0.605
DT	0.02	0.88	0.45	0.31	0.8	0.555	0	0.88	0.44	0.12	0.87	0.495	0.41	0.88	0.645	0.52	0.87	0.695
KNN	0	0.88	0.44	0	0.89	0.445	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44
SVM	0	0.88	0.44	0	0.89	0.445	0	0.88	0.44	0.09	0.88	0.485	0	0.88	0.44	0	0.88	0.44

Tabla A.3: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Negativas** de las publicaciones de las *empresas*

Reacciones Negativas																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5
DT	0.01	1	0.505	0.15	0.9	0.525	0	1	0.5	0	1	0.5	0	1	0.5	0.26	0.92	0.59
KNN	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5
SVM	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5

Tabla A.4: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Neutras** de las publicaciones de las *empresas*

Reacciones Neutras																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.83	0.415	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0.26	0.93	0.595
DT	0.02	0.93	0.475	0.22	0.87	0.545	0	0.93	0.465	0.01	0.93	0.47	0	0.93	0.465	0.42	0.91	0.665
KNN	0	0.93	0.465	0	0.94	0.47	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465
SVM	0	0.93	0.465	0	0.94	0.47	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465	0	0.93	0.465

Tabla A.5: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Positivas** de las publicaciones de las *empresas*

Reacciones Positivas																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0.36	0.88	0.62
DT	0.02	0.88	0.45	0.31	0.8	0.555	0	0.88	0.44	0.15	0.87	0.51	0.05	0.88	0.465	0.53	0.87	0.7
KNN	0	0.91	0.455	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44	0	0.88	0.44
SVM	0	0.88	0.44	0	0.89	0.445	0	0.88	0.44	0.1	0.88	0.49	0	0.88	0.44	0	0.88	0.44

Tabla A.6: Resultados de los experimentos *ExpA* para predecir la métrica **Veces Compartidos** de las publicaciones de las *empresas*

Veces Compartidos																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0.24	0.91	0.575
DT	0.02	1	0.51	0.28	0.86	0.57	0	0.92	0.46	0.03	0.91	0.47	0	0.92	0.46	0.45	0.9	0.675
KNN	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.93	0.465
SVM	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.93	0.465

A.1.2 | Resultados con el conjunto de experimentos **ExpB**

Tabla A.7: Resultados de los experimentos *ExpB* para predecir la métrica **comentarios** de las publicaciones de las *empresas*

Comentarios																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.41	0.92	0.66	0.42	0.91	0.67	0.43	0.9	0.67	0.42	0.91	0.67	0.42	0.91	0.67	0.42	0.91	0.67	0.42	0.91	0.67
DT	0.32	0.9	0.61	0.34	0.9	0.62	0.33	0.91	0.61	0.35	0.9	0.63	0.32	0.9	0.61	0.47	0.92	0.69	0.46	0.92	0.69
KNN	0.29	0.8	0.54	0.33	0.86	0.59	0.3	0.82	0.56	0.38	0.87	0.64	0.3	0.83	0.56	0.31	0.88	0.6	0.39	0.91	0.65
SVC	0.35	0.91	0.63	0.36	0.91	0.64	0.31	0.9	0.6	0.38	0.91	0.64	0.36	0.91	0.64	0.36	0.91	0.63	0.38	0.91	0.65

Tabla A.8: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Totales** de las publicaciones de las *empresas*

Reacciones Totales																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.49	0.86	0.68	0.5	0.87	0.69	0.49	0.85	0.67	0.49	0.87	0.68	0.48	0.87	68	0.49	0.87	0.68	0.52	0.87	0.69
DT	0.42	0.86	0.63	0.45	0.85	0.65	0.41	0.86	0.63	0.46	0.87	0.66	0.44	0.87	0.65	0.51	0.88	0.69	0.58	0.89	0.73
KNN	0.39	0.81	0.6	0.42	0.78	0.6	0.43	0.79	0.61	0.44	0.75	0.6	0.43	0.75	0.59	0.5	0.81	0.72	0.51	0.85	0.68
SVC	0.43	0.86	0.64	0.44	0.87	0.65	0.43	0.86	0.64	0.44	0.87	0.65	0.45	0.87	0.66	0.42	0.86	0.64	0.47	0.87	0.67

Tabla A.9: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Negativas** de las publicaciones de las *empresas*

Reacciones Negativas																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.14	0.98	0.58	0.15	0.97	0.58	0.15	0.97	0.58	0.15	0.97	0.58	0.15	0.98	0.59	0.14	0.97	0.57	0.15	0.98	0.59
DT	0.25	0.94	0.6	0.24	0.94	0.6	0.22	0.94	0.59	0.25	0.95	0.61	0.24	0.91	0.57	0.25	0.95	0.61	0.26	0.95	0.62
KNN	0.4	0.76	0.52	0.31	0.85	0.55	0.32	0.32	0.54	0.45	0.76	0.53	0.42	0.72	0.5	0.24	0.9	0.57	0.41	0.79	0.54
SVC	0.23	0.96	0.61	0.23	0.96	0.62	0.29	0.95	0.62	0.24	0.96	0.62	0.23	0.96	0.61	0.23	0.96	0.61	0.24	0.96	0.62

Tabla A.10: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Neutras** de las publicaciones de las *empresas*

Reacciones Neutras																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.34	0.9	0.62	0.3	0.92	0.61	0.32	0.9	0.62	0.33	0.91	0.62	0.33	0.91	0.62	0.33	0.91	0.62	0.33	0.91	0.62
DT	0.28	0.91	0.6	0.34	0.89	0.61	0.31	0.91	0.6	0.33	0.9	0.61	0.3	0.91	0.6	0.46	0.92	0.69	0.47	0.93	0.7
KNN	0.3	0.82	0.56	0.3	0.92	0.58	0.31	0.83	0.57	0.34	0.86	0.6	0.31	0.84	0.57	0.47	0.92	0.69	0.4	0.91	0.66
SVC	0.32	0.91	0.62	0.32	0.91	0.61	0.36	0.91	0.63	0.32	0.91	0.62	0.33	0.91	0.62	0.33	0.91	0.62	0.3	0.91	0.62

Tabla A.11: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Positivas** de las publicaciones de las *empresas*

Reacciones Positivas																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.5	0.86	0.68	0.51	0.87	0.69	0.49	0.87	0.68	0.49	0.87	0.68	0.49	0.87	0.68	0.5	0.87	0.68	0.52	0.88	0.7
DT	0.42	0.86	0.64	0.42	0.86	0.64	0.42	0.86	0.64	0.47	0.87	0.67	0.43	0.86	0.65	0.51	0.87	0.69	0.59	0.89	0.74
KNN	0.49	0.75	0.6	0.44	0.79	0.6	0.51	0.76	0.61	0.43	0.78	0.61	0.42	0.84	0.63	0.5	0.81	0.68	0.5	0.86	0.68
SVC	0.44	0.88	0.64	0.4	0.88	0.64	0.43	0.86	0.65	0.42	0.88	0.66	0.46	0.88	0.66	0.43	0.87	0.65	0.47	0.87	0.67

Tabla A.12: Resultados de los experimentos *ExpB* para predecir la métrica **Veces Compartidos** de las publicaciones de las *empresas*

Veces Compartidos																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.42	0.89	0.67	0.42	0.91	0.66	0.43	0.89	0.67	0.43	0.91	0.67	0.41	0.91	0.66	0.44	0.91	0.67	0.47	0.91	0.69
DT	0.36	0.9	0.63	0.4	0.89	0.64	0.39	0.9	0.64	0.39	0.9	0.65	0.38	0.9	0.64	0.48	0.91	0.7	0.52	0.92	0.72
KNN	0.35	0.89	0.62	0.35	0.87	0.61	0.35	0.83	0.59	0.39	0.79	0.58	0.35	0.88	0.62	0.4	0.83	0.62	0.47	0.91	0.69
SVC	0.4	0.9	0.65	0.39	0.9	0.65	0.4	0.9	0.65	0.41	0.91	0.66	0.41	0.91	0.66	0.41	0.91	0.66	0.41	0.91	0.66

A.2 | Resultados al predecir las métricas de los políticos

A.2.1 | Resultados con el conjunto de experimentos **ExpA**

Tabla A.13: Resultados de los experimentos *ExpA* para predecir la métrica **comentarios** de las publicaciones de los *políticos*

Comentarios																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0.32	0.9	0.61
DT	0	0.92	0.46	0.22	0.84	0.53	0	0.92	0.46	0.16	0.89	0.525	0.39	0.93	0.66	0.55	0.91	0.73
KNN	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0.39	0.93	0.66	0.39	0.93	0.66
SVM	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0.39	0.93	0.66	0.39	0.93	0.66

Tabla A.14: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Totales** de las publicaciones de los *políticos*

Reacciones Totales																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	1	0.5	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0.5	0.77	0.635	0.51	0.84	0.675
DT	0	0.89	0.445	0.26	0.8	0.53	0	0.89	0.445	0.2	0.85	0.525	0.42	0.92	0.67	0.65	0.91	0.78
KNN	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0.42	0.92	0.67	0.42	0.92	0.67
SVM	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0.42	0.92	0.67	0.42	0.92	0.67

Tabla A.15: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Negativas** de las publicaciones de los *políticos*

Reacciones Negativas																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0.09	0.97	0.53
DT	0	1	0.5	0.19	0.85	0.52	0	1	0.5	0.14	0.94	0.54	0	1	0.5	0.44	0.91	0.675
KNN	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5
SVM	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5

Tabla A.16: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Neutras** de las publicaciones de los *políticos*

Reacciones Neutras																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0.25	0.93	0.59
DT	0	0.94	0.47	0.16	0.88	0.52	0	0.94	0.47	0.15	0.92	0.535	0	0.94	0.47	0.47	0.93	0.7
KNN	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47
SVM	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47	0	0.94	0.47

Tabla A.17: Resultados de los experimentos *ExpA* para predecir la métrica **Reacciones Positivas** de las publicaciones de los *políticos*

Reacciones Positivas																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0.51	0.78	0.645	0.53	0.84	0.685
DT	0	0.89	0.445	0.26	0.8	0.53	0	0.89	0.445	0.2	0.85	0.525	0.46	0.91	0.685	0.66	0.92	0.79
KNN	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0	0.89	0.445	0.42	0.92	0.67	0.42	0.92	0.67
SVM	0	0.89	0.445	0	0.89	0.445	0	1	0.5	0	0.89	0.445	0.42	0.92	0.67	0.42	0.92	0.67

A.2.2 | Resultados con el conjunto de experimentos **ExpB**

Tabla A.18: Resultados de los experimentos *ExpA* para predecir la métrica **Veces compartidos** de las publicaciones de los *políticos*

Veces compartidos																		
Algoritmo	*Exp1a*			*Exp2a*			*Exp3a*			*Exp4a*			*Exp5a*			*Exp6a*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0	0.89	0.445	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0.42	0.89	0.655
DT	0	0.92	0.46	0.2	0.85	0.525	0	0.92	0.46	0.13	0.89	0.51	0.5	0.95	0.725	0.59	0.93	0.76
KNN	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0.5	0.95	0.725	0.5	0.95	0.725
SVM	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0	0.92	0.46	0.5	0.95	0.725	0.5	0.95	0.725

Tabla A.19: Resultados de los experimentos *ExpB* para predecir la métrica **Comentarios** de las publicaciones de los *políticos*

Comentarios																				
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.36	0.38	0.63	0.38	0.9	0.64	0.36	0.9	0.63	0.37	0.9	0.64	0.37	0.9	0.64	0.37	0.9	0.64	0.36	0.9	0.63
DT	0.33	0.9	0.59	0.31	0.88	0.6	0.3	0.88	0.59	0.34	0.89	0.61	0.33	0.89	0.61	0.48	0.91	0.69	0.57	0.93	0.75
KNN	0.3	0.81	0.55	0.3	0.8	0.55	0.31	0.8	0.55	0.29	0.84	0.56	0.31	0.85	0.58	0.48	0.9	0.69	0.48	0.91	0.69
SVC	0.36	0.9	0.63	0.37	0.9	0.63	0.37	0.9	0.63	0.36	0.9	0.63	0.38	0.9	0.64	0.46	0.93	0.7	0.5	0.93	0.72

Tabla A.20: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Totales** de las publicaciones de los *políticos*

Reacciones Totales																				
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.44	0.88	0.66	0.44	0.88	0.66	0.43	0.88	0.66	0.42	0.88	0.65	0.44	0.88	0.66	0.44	0.88	0.66	0.44	0.88	0.66
DT	0.36	0.85	0.61	0.37	0.85	0.61	0.36	0.86	0.61	0.42	0.87	0.64	0.38	0.86	0.62	0.57	0.9	0.73	0.68	0.92	0.8
KNN	0.33	0.8	0.56	0.34	0.84	0.59	0.35	0.8	0.58	0.47	0.74	0.55	0.37	0.76	0.57	0.56	0.89	0.73	0.54	0.89	0.71
SVC	0.43	0.87	0.66	0.44	0.88	0.66	0.43	0.87	0.65	0.44	0.88	0.66	0.43	0.87	0.65	0.55	0.92	0.73	0.58	0.92	0.75

Tabla A.21: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Negativas** de las publicaciones de los *políticos*

Reacciones Negativas																				
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.27	0.9	0.58	0.26	0.91	0.58	0.23	0.92	0.58	0.25	0.92	0.59	0.24	0.92	0.59	0.24	0.92	0.59	0.27	0.91	0.6
DT	0.22	0.9	0.56	0.22	0.9	0.56	0.2	0.89	0.55	0.26	0.9	0.58	0.24	0.9	0.57	0.34	0.9	0.62	0.41	0.93	0.68
KNN	0.3	0.8	0.53	0.38	0.74	0.52	0.35	0.78	0.54	0.27	0.84	0.55	0.37	0.76	0.53	0.49	0.77	0.58	0.48	0.78	0.58
SVC	0.22	0.92	0.57	0.21	0.93	0.57	0.21	0.92	0.58	0.23	0.93	0.59	0.22	0.93	0.58	0.25	0.93	0.6	0.29	0.93	0.62

Tabla A.22: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Neutras** de las publicaciones de los *políticos*

Reacciones Neutras																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.29	0.91	0.6	0.29	0.92	0.6	0.29	0.92	0.6	0.29	0.92	0.6	0.28	0.92	0.6	0.28	0.92	0.6	0.29	0.92	0.6
DT	0.22	0.9	0.56	0.22	0.9	0.56	0.22	0.9	0.56	0.26	0.9	0.58	0.28	0.91	0.59	0.37	0.91	0.64	0.51	0.94	0.72
KNN	0.22	0.86	0.54	0.23	0.87	0.55	0.24	0.84	0.54	0.25	0.87	0.56	0.28	0.84	0.56	0.32	0.89	0.6	0.37	0.92	0.65
SVC	0.27	0.92	0.59	0.26	0.92	0.59	0.28	0.92	0.6	0.27	0.92	0.6	0.29	0.92	0.6	0.33	0.92	0.62	0.38	0.92	0.65

Tabla A.23: Resultados de los experimentos *ExpB* para predecir la métrica **Reacciones Positivas** de las publicaciones de los *políticos*

Reacciones Positivas																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.44	0.88	0.66	0.44	0.88	0.66	0.44	0.88	0.66	0.44	0.88	0.66	0.43	0.88	0.66	0.44	0.88	0.66	0.45	0.88	0.66
DT	0.36	0.85	0.61	0.37	0.86	0.62	0.37	0.86	0.61	0.43	0.87	0.65	0.37	0.85	0.61	0.59	0.9	0.74	0.68	0.92	0.8
KNN	0.37	0.71	0.53	0.34	0.81	0.58	0.35	0.79	0.57	0.35	0.83	0.59	0.36	0.76	0.56	0.57	0.89	0.73	0.54	0.89	0.71
SVC	0.43	0.87	0.65	0.44	0.88	0.66	0.44	0.88	0.66	0.44	0.88	0.66	0.43	0.88	0.65	0.56	0.92	0.74	0.58	0.92	0.75

Tabla A.24: Resultados de los experimentos *ExpB* para predecir la métrica **Veces Compartidos** de las publicaciones de los *políticos*

Veces Compartidos																					
Algoritmo	*Exp1b*			*Exp2b*			*Exp3b*			*Exp4b*			*Exp5b*			*Exp6b*			*Exp7b*		
	(+)	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*	*(+)*	*(-)*	*(F-M)*
NB	0.45	0.91	0.67	0.44	0.92	0.68	0.43	0.91	0.67	0.41	0.91	0.66	0.43	0.91	0.67	0.44	0.92	0.68	0.45	0.92	0.68
DT	0.35	0.9	0.61	0.34	0.9	0.62	0.37	0.86	0.61	0.4	0.9	0.65	0.37	0.9	0.63	0.55	0.93	0.87	0.63	0.94	0.78
KNN	0.3	0.85	0.58	0.32	0.84	0.58	0.3	0.87	0.58	0.34	0.88	0.61	0.3	0.84	0.57	0.5	0.93	0.72	0.54	0.93	0.73
SVC	0.41	0.91	0.66	0.42	0.91	0.66	0.41	0.91	0.66	0.43	0.91	0.67	0.42	0.91	0.67	0.57	0.95	0.76	0.59	0.95	0.77