PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

MÉTODO DE PREDICCIÓN ELECTORAL A TRAVÉS DE MODELO BASADO EN ANÁLISIS DE SENTIMIENTOS EN TWITTER

CAMILA ALEJANDRA FIGUEROA RODRÍGUEZ

Profesor Guía: **Hector Allende Cid**
Profesor Co-referente: **Wenceslao Palma Muñoz**

Carrera: **Ingeniería Civil Informática**

Agosto 2018

Resumen

El aumento exponencial de usuarios activos en los últimos años en redes sociales, ha motivado el surgimiento de nuevas técnicas para la realización de estudios socio-demográficos que utilizan dichas aplicaciones para la obtención de un gran volumen de datos, y a partir de esta información, poder generar estadísticas o cifras que permitan conocer la opinión de un grupo determinado de personas. En el siguiente informe se explicarán los modelos y métodos que se proponen para la obtención de cifras predictivas en el proceso de elecciones presidenciales 2017 realizado en Chile, por medio de la extracción de datos en una red social; Twitter. Se busca obtener la menor diferencia posible con respecto a los resultados obtenidos posteriormente en urnas de votación, para poder considerar a ésta técnica como un método de predicción válido para eventos relacionados a votaciones electorales..

Palabras Clave: Twitter, Predicción Electoral, Análisis de Sentimientos, Bolsa de Palabras, Aprendizaje Supervisado, Aprendizaje de Máquina.

Abstract

In recent years, the exponential increase of active users in social networks has motivated the emergence of new techniques for carrying out socio-demographic studies. These new studies use such techniques to obtain large volumes of data, and, from this information, to generate statistics or figures that allow one to access the opinions of a specific group of people. The following report will explain, through the extraction of data in the social network Twitter, the models and methods proposed for obtaining predictive figures in the process of the 2017 presidential elections in Chile. In order to be able to consider this technique as a valid method of prediction of events related to electoral votes, the report seeks to obtain the smallest possible difference between the social media data and the results obtained later from polling stations.

Palabras Clave: Twitter, Presidential prediction, Sentiment Analysis, Bag of Words, Supervised learning, Machine learning.

Agradecimientos

Quiero expresar mi agradecimiento a todas las personas que directa e indirectamente han participado en mi desarrollo académico y de la presente tesis, la cual, ha requerido de mucho esfuerzo y trabajo .

En primer lugar, a mi madre, por su innegable y constante sacrificio y apoyo fundamental, el cual, me ha permitido llegar a este punto de mi carrera. Durante todos estos años, ella fue y será quien ha formado gran parte de mis buenos valores y quien ha estado siempre para mi, en las buenas y en las no tan buenas.

Agradecer también a todos aquellos que contribuyeron en darme fuerzas para poder continuar en los momentos de dificultad en todo ámbito; amigos y conocidos que de manera totalmente desinteresada han contribuido a mi aprendizaje y me han dado una palabra de aliento cuando la he necesitado.

Sin duda, nada hubiese sido igual, sin el apoyo familiar y el de mi pareja, quien ha estado presente durante este ultimo periodo, y de quien me siento muy orgullosa, ya que ha sabido brindarme de diferentes formas todo su apoyo y contensión incondicional, entregándome mucho amor día a día, a pesar de las adversidades. He logrado aprender en esta última etapa de carrera, que la vida puede presentar muchos cambios de rumbo, y no por eso hay que decaer.

Por último, y no por eso menos importante, agradecer a mis profesores Hector Allende y Wenceslao Palma por guiarme en este camino de investigación, por su tiempo y capacidad de transmitir los conocimientos con los que cierro esta etapa.

"Un poco más de persistencia, un poco más de esfuerzo, y lo que parecia un irremediable fracaso, puede resultar en éxito glorioso" (Elbert Hubbard)

Índice

Lista de Figuras

1. Introducción

Desde el año 2017, el número de usuarios en el mundo, que utiliza aplicaciones desde el móvil ha aumentado exponencialmente, superando en un 34 porciento al del año 2016 (2549 millones de personas), este gran aumento ha motivado la realización de ciertas investigaciones que buscan obtener resultados sociodemográficos, ya que se descubrió que, a través de las redes sociales, se puede conseguir la más diversa concentración de opiniones y preferencias. Todo lo anterior abre la puerta a un nuevo mundo de análisis social; poder sacar conclusiones de la opinión a nivel global, país, o bien de manera individual, con respecto a un tema, producto, servicio, candidato a presidente, etc.

Según estudios realizados por We Are Social (Marketing4ECommerce, 2017) a comienzos del 2017, existen más de 3750 millones de usuarios en línea alrededor del mundo. Un 45 porciento de Latinoamérica es activa en las redes sociales. Lo anterior ha significado un aumento de un 30 porciento del uso del tráfico en la internet, y un incremento en el número de participantes en redes sociales, que alcanza un 20 porciento, todo esto, en comparación al año 2016. Más de un tercio de la población mundial utiliza alguna de estas plataformas masivas de comunicación social online.

BBVA Research (Emol, 2016) realizó durante el año 2016 un estudio comparativo sobre el uso internet en Chile, en relación a otros países de Latinoamérica y el mundo titulado " Contexto digital: Alianza del Pacífico", el cual, arrojó que el 95 porciento de los chilenos de entre 18-34 años se admite usuario de redes sociales. Todo lo anterior supera el 85 porciento que promedian países desarrollados fuera del continente, Estados Unidos o Reino Unido. En base a éstas cifras es que se puede considerar a Chile como un buen candidato para la realización de estudios de análisis de sentimientos.

En la actualidad Chile posee ciertos métodos de predicción basados en encuestas, las cuales son utilizadas para predecir qué candidato presidencial será electo, pero, éstos han carecido de credibilidad para la población en los últimos años, debido al gran sesgo que poseen entre los números obtenidos por las encuestas y las cifras reales obtenidas posterior a las votaciones en urna. El estudio realizado se centra en el contexto actual nacional, donde dos grandes partidos políticos se enfrentan; Frente amplio y Chile Vamos. Se busca disminuir por medio de esta nueva estrategia, el sesgo obtenido entre el método de predicción y resultados reales obtenidos posterior a las votaciones. Identificando de esta manera, a los grupos sociales que están a favor o en contra de cada candidato de alguno de los dos partidos políticos, y sus opiniones.

El análisis de sentimientos permite, la obtención de parámetros globales que den a conocer de manera genérica cual es la opinión de un grupo en particular con respecto a algún tema puntual. Se puede por medio de ésta técnica, conocer si para los participantes de una aplicación en específico les agrada o no una idea, un sistema o bien, un candidato político. Todo esto mediante el estudio y análisis de las palabras que contiene cada opinión publicada en alguna red social, por medio del uso de ciertos algoritmos de clasificación y la posterior aplicación de técnicas de aprendizaje supervisado en máquinas de soporte vectorial. El siguiente documento, explica como se implementa esta técnica para la obtención de un clasificador cuyo fin es obtener cifras predictivas para el proceso de primarias 2018. Se detallarán los algoritmos y técnicas a utilizar y se indicarán los resultados obtenidos por medio de diferentes representaciónes de datos, ya que estas pueden jugar un papel fundamental en el resultado de las técnicas a aplicar, así se podrá conocer cual es la mejor dado el contexto de clasificación de Tweets.

2. Descripción del Problema

Hoy en día existen en Chile distintos métodos para poder predecir el comportamiento de la ciudadanía en los procesos de votaciones en urna; (LaNación, 2017) CEP, Admimark y Cadem , son encuestas realizadas a ciertos sectores más representativos de la población para la obtención de números que reflejen que candidato presidencial será el más votado. Estos pronósticos han significado un gran avance en materia de predicción electoral, pero en años anteriores ha quedado en visto, la poca exactitud de sus porcentajes, un sesgo importante en comparación a los reales números obtenidos posterior a las votaciones realizadas en urna. Lo que ha significado que se les acuse de intentar privilegiar a ciertos candidatos por sobre otros, o bien, que este tipo de encuestas pierdan la credibilidad ante el público.

Es por esto que surge la necesidad de poder encontrar un método más preciso, que permita estimar los porcentajes de aprobación del público a candidatos políticos. Hasta el momento, la predicción de las cifras exactas es algo que sale de lo imaginable, pero se pretende que el sesgo sea el menor posible, en comparación al resultado en urnas de votación.

Hipótesis

En el área de Aprendizaje de máquina (o machine learning) los resultados de los algoritmos utilizados para la clasificación de etiquetas y el análisis de los sentimientos asociados a dicha catalogación, están directamente relacionados con la intención de voto de la persona que publica cada Tweet, lo cual permitirá obtener un sesgo, o error cuadrático medio (MAE), bajo el 5 porciento finalizado el proceso de predicción electoral en contexto nacional de elecciones presidenciales de Chile.

3. Objetivos

3.1. Objetivo General

Generar un modelo que permita extraer, por medio de métodos de clasificación, el punto de vista y opinión de usuarios chilenos en Twitter, hacia un candidato presidencial en el proceso de primarias 2017. Una vez realizado este modelo, se busca obtener datos predictivos que permitan obtener el menor sesgo posible con los resultados en la votación llevada a cabo en todo el país.

3.2. Objetivos Específicos

Como objetivos específicos del estudio, se tienen los siguientes:

- Generar un corpus con datos etiquetados de análisis de sentimientos y representarlos de manera numérica, realizando un preprocesamiento a los datos.
- Determinar los modelos y algoritmos adecuados para la investigación
- Aplicar técnicas para la identificación de contexto en cada Tweet.
- Obtener cifras predictivas y métricas que muestren un buen desempeño.

4. Estado del Arte

Debido a la masiva convocatoria a el uso de redes sociales, es que han surgido nuevas herramientas enfocadas al análisis de sentimientos, puesto que, es una técnica que está tomando cada vez más fuerza, dado el gran volumen de datos que hoy en día circula en la internet, a través del contenido que los mismos usuarios publican en las diferentes aplicaciones de comunicación social.

Con respecto a los estudios realizados para el caso de análisis de sentimientos en elecciones presidenciales, ***Análisis de sentimientos y predicción de eventos en Twitter*** (García, 2014), es una tesis enfocada a la clasificación de publicaciones realizadas en Twitter, a través de análisis léxico y procesamiento de palabras, por medio de una máquina de soporte vectorial, haciendo uso de un diccionario que almacena términos asociados a sentimientos positivos o negativos.

También la tesis ***Análisis de sentimiento y clasificación de texto mediante Adaboost concurrente*** (Castro, 2016) se enfoca principalmente a la clasificación en redes sociales por medio de extracción de Tweets a través de una herramienta llamada Analitic y la aplicación de un algoritmo de concurrencia (Adaboost) utilizando un diccionario de palabras.

Aplicaciones como ***Social Bakers*** (SocialBakers, 2008) surgieron para intentar obtener una similitud, con los porcentajes de adherente a cada candidato político en un contexto presidencial, todo ello por medio del análisis y cuantificación de la cantidad de usuarios seguidores en Twitter y Facebook que posee cada candidato presidencial, buscando predecir cual candidato sería el más votado, por medio de la generación de estadísticas y gráficas representativas de las cifras obtenidas.

En materia de análisis de sentimiento existen herramientas léxicas como ***Opinion lexicon*** (Hu M, 2004), la cual se centra en la clasificación de mensajes que contengan emoticones (manifestación de una emoción por medio del uso de ciertos caracteres, por ejemplo ":-)"que representa alegría o felicidad) positivos o negativos. Si el mensaje en cuestión no los posee, no se puede clasificar. También, Twitter Sentiment es una herramienta que posteriormente fue llamada como ***Sentimente140*** (Go A, 2009) , y propone un conjunto de tres clasificadores construidos que poseen un almacén de gran cantidad de Tweets con emociones asociadas, con ello se realizan comparaciones para saber si un Tweet es positivo o negativo.

5. Marco Teórico

En esta sección se presentarán los conceptos y técnicas que se utilizan para la realización de proyectos de predicción a través de redes sociales.

5.1. Análisis de sentimientos

Hace referencia al procesamiento del lenguaje natural o escrito (Bermingham and Smeaton, 2011; Andranik Tumasjan, 2010; Deltell, 2015; Enrique Alonso, 2013) (en este contexto denominado, lingüística computacional) utilizado para la identificación y extracción de información proveniente de ciertas fuentes (Ej. Aplicaciones, sitios web). Esta información permite (mediante el uso y avance de ciertas tecnologías), obtener estadísticas respecto a la opinión de los usuarios participantes de una red social. En la actualidad, existen herramientas que permiten clasificar directamente un listado almacenado de publicaciones en redes sociales por medio de "Tags" o "Etiquetas" que determinan si un mensaje publicado posee una opinión "positiva" o "negativa" con respecto a alguna temática en particular.

De la variedad de emociones que posee el ser humano, podemos destacar seis tipos: Alegría, Sorpresa, Disgusto, Miedo y Tristeza.

Dentro del proceso de análisis de sentimientos se realiza el de procesamiento de las palabras contenidas en una oración. Esta fase se puede realizar de dos maneras distintas; Análisis semántico y análisis sintáctico.

5.1.1. Análisis sintáctico

El analizador sintáctico se encarga de chequear el texto de ingreso a un sistema, basado a una gramática o diccionario de palabras dado previamente. Es decir, a partir de un documento que posee todas las palabras que normalmente se asocian a ciertas emociones, asignar pesos a cada una, y determinar si es que un Tweet es positivo o negativo dependiendo de que términos este compuesto.

5.1.2. Análisis semántico

Corresponde al significado asociado a las estructuras formales (sintáxis) del lenguaje. Se almacena un diccionario o bien una bolsa de palabras, que según una clasificación previa dada el contexto real de la oración (útil para casos en dónde se utilicen palabras consideradas como negativas en un Tweet con intención real positiva) se pueda catalogar, si es que el Tweet completo posee un sentimiento positivo o negativo hacia un candidato.

5.1.3. Aprendizaje Computacional

Busca analizar los datos por medio de aprendizaje supervisado y de automática, para ello, inicialemente se debe realizar una fase de entrenamiento, la cual permitirá sentar las bases para que la máquina pueda catalogar un futuro grupo de Tweets nuevos (sin clasificar), de esta manera se pretende obtener el menor sesgo posible en comparación a los resultados obtenidos de manera inicial, realizando pruebas hacia los datos de entrada y salida y posteriormente validando. Dentro de las técnicas del aprendizaje computacional se tiene : El uso de máquinas de soporte vectorial (SVM), Naive Bayes y clasificadores (García, 2014).

5.1.4. Aprendizaje supervisado y no supervisado

El aprendizaje supervisado (Becerra, 2016) corresponde a una técnica que permite entrenar a la máquina de soporte vectorial a partir de datos clasificados de manera inicial, es decir que, a partir de datos previos clasificados, se pretende a partir de esto que el algoritmo y la máquina obtengan los resultados deseados al ingresar datos nuevos que no han sido clasificados. A partir de las entradas válidas se pretende que la máquina "aprenda" a reconocer patrones y logre identificar a las nuevas entradas y a partir de ello se obtengan resultados válidos, o, en otras palabras, que el sistema logre actuar según lo esperado a partir de los ejemplos iniciales.

Los sistemas de clasificación supervisados son aquellos en los que, a partir de un conjunto de ejemplos ya categorizados previamente (conjunto de entrenamiento), intentan asignar una clasificación a un segundo conjunto de ejemplos de prueba. A diferencia de los sistemas de clasificación no supervisados, los cuales no disponen de una base de ejemplos previamente clasificados, sino que únicamente a partir de las propiedades de los elementos de entrada, buscan dar una agrupación (clasificación, clustering) de los ejemplos según su similaridad, es por ello que para éste proyecto se utilizará la técnica de aprendizaje supervisado.

5.2. Algoritmos mas utilizados

A continuación, se presentan los algoritmos utilizados en procesos de predicción, todos estos pueden ser aplicados mediante el uso de distintos lenguajes de programación (Ej: utilizando la librería de aprendizaje de máquinas provista por "Scikit-learn") (Sckit-learn, 2017).

5.2.1. Árboles de decision

Los árboles de decisión son un método no paramétrico de aprendizaje supervisado utilizado para clasificación y regresión. La meta principal es la creación de un modelo que predice el valor de una variable específica, mediante el aprendizaje de reglas de decisión simples inferidas de las características de los datos de entrenamiento.

5.2.2. Naive Bayes

Este método (Murphy, 2006) corresponde a uno de los más utilizados para el procesamiento de datos en análisis de sentimientos, posee una baja complejidad de implementación y un buen desempeño en relación a sus resultados. Es un método de probabilidad el cual se basa en las siguientes fórmulas de cálculo de pertenencia:

$$P(a|b) \propto P(a) \prod_{1 \leq h \leq n_b} P(t_h|a) \tag{1}$$

Desde aquí podemos calcular la probabilidad de que el término t_h suceda en un objeto de tipo a como $P(t_h|a)$.

Para el caso particular del análisis de sentimientos es que se puede utilizar la siguiente expresión de probabilidad:

$$P(a|b) = \frac{P(a)P(b|a)}{P(b)} \tag{2}$$

Donde $P(a|b)$ corresponde a la probabilidad de que dada un elemento de la bolsa de palabras b, corresponda a una de tipo a que ya haya sido clasificada previamente con algun sentimiento asociado. Obteniendo esta probabilidad se puede obtener la polaridad de la frase asociada, para determinar si un Tweet es positivo o negativo hacia un candidato.

$P(b|a)$ corresponde a la probabilidad de que una palabra se encuentre y se extraiga directamente de la base de entrenamiento. Por otro lado $P(b)$ es un factor de normalización.

5.2.3. Random Forest

Un "Random Forest" es un meta estimador, que entrena un conjunto de árboles de decisión en varias submuestras del set de datos y utiliza los promedios de dichos árboles para mejorar la exactitud predictiva y controlar los posibles sesgos que se pueden generar en vez los árboles son entrenados.

5.2.4. AdaBoost

Un clasificador AdaBoost es un meta estimador, el cual empieza entrenando un clasificador con todo el set de datos original. Luego empieza a repetir este proceso generando copias de los clasificadores, pero la diferencia es que este meta estimador empieza a asignar mayores pesos a instancias las cuales fueron clasificadas incorrectamente, esto para poder ajustar a los subsiguientes clasificadores que se generan a instancias de datos más difíciles.

Para lo anterior es que se hace uso de la bolsa de palabras (bag of words) (G. Paltoglou, 2013) para poder obtener información útil para la determinación de la polaridad en cada Tweet.

5.2.5. Máquina de soporte vectorial lineal (SVM)

Las máquinas de soporte vectorial (SVM), son modelos de aprendizaje supervisados asociados a los algoritmos de aprendizaje que analizan los datos usados para la clasificación y análisis de regresión. Dado un conjunto de ejemplos de entrenamiento, cada uno etiquetado en una o más categorías, un algoritmo de entrenamiento de SVM construye un modelo que asigna nuevos ejemplos en distintas categorías, haciéndolo un clasificador binario linear.

Un modelo SVM es una representación de los ejemplos como puntos en un espacio, mapeados de tal manera que los ejemplos de distintas categorías están divididos por un espacio lo más ancho posible. Para el proceso de clasificación, los nuevos ejemplos (datos a clasificar), son mapeados en este espacio y luego se predice la categoría a la cual estos pertenecen dependiendo de qué lado del espacio caen. Las SVM se denominan lineales, cuando la función encargada de hacer las separaciones las realizas con rectas entre los planos.

6. Clasificación de Tweets

Se presentarán las estrategias utilizadas para la clasificación de Tweets, para la posterior obtención de cifras estimativas a través de la predicción de datos.

6.1. Uso de Herramienta Analitic PRO

La herramienta Analitic PRO (AnaliticPro, 2018) permite dividir los Tweets por candidato al cual está dirigido dicho comentario u opinión emitida por Twitter, es decir, que podemos observar en una lista todas las publicaciones, de las diferentes cuentas nacionales, que son realizadas mencionando al candidato correspondiente. Es por ello que se puede dimensionar el volumen de opiniones que realiza la población en cierto periodo determinado. Se logran identificar las cuentas Chilenas a través de la zona horaria inscrita en el Tweet correspondiente; A través de ello se han podido identificar al rededor de cuatrocientos setenta mil cuentas en el país, de las cuales se obtienen millones de publicaciones, para conformar un gran almacén de datos, que será utilizado para la clasificación y uso de datos, para los procesos de entrenamiento y prueba en la máquina.

Se ha apreciado un aumento significativo en el volumen de Tweets durante el periodo correspondiente a la presentación de un candidato en particular, en los medios de comunicación masiva. Es en esos periodos donde los usuarios comienzan a publicar sus opiniones a favor o en contra de un candidato, a modo de apoyo a las ideas propuestas, o bien, a modo de manifestación de su desagrado ante el político en cuestión, por medio de expresiones que denotan disgusto o negatividad. Una vez captado todo el volumen de Tweets, y separados por el presidenciable correspondiente, estos se clasifican por medio de la herramienta y máquina utilizada, en positivo o negativo, asociando esta información al candidato mencionado en la publicación.

6.2. Representación computacional de textos

Por lo general, el análisis de sentimientos intenta reconocer estos por medio de la detección de ciertas palabras clave, que se guardan en un diccionario de datos y permiten, determinar la emoción asociada a una frase. Este almacén de palabras es conocido como "Bag of words" (G. Paltoglou, 2013) (bolsa de palabras), la cual es una representación simplificada que utiliza el procesamiento del lenguaje natural y la cantidad de repeticiones en la oración correspondiente.

	eveyrthing	interesting	learning	lerning	like	Machien	machine	not	predicts	problems	solving	sure	What
1	0	1	0	0	1	0	0	0	0	1	1	0	0
2	0	0	1	0	0	0	1	0	0	0	0	0	1
3	0	0	0	0	0	0	0	1	0	0	0	1	0
4	1	0	0	1	0	1	0	0	1	0	0	0	0

Figura 1: Representación vectorial de un texto dividido en palabras.

De manera inicial se consideraba el hecho de que la información que se almacena permitía entrenar a uno (o más) clasificadores. Utilizando una oración, se almacenaron sus palabras por separado, como también, la cantidad de veces que se repetía cada palabra asociada. (Ej. "Prefiero dar mi voto a mi candidato, ese candidato es el mejor", almacenará en la bolsa de datos un vector

con las palabras componente de la publicación:

[“Prefiero”, “dar”,”mi”,”voto”,”a”,”mi”,”candidato”,”ese”,”es”,”el”,”mejor”].

Se consideró un vector de ocurrencias que posee el número de veces que se repite la palabra en la posición correspondiente:

[1,1,1,1,1,1,2,1,1,1].

Asignando un valor “2” a la palabra “candidato” que se repite dos veces en la oración).

6.2.1. Representación inicial de Tweets por medio de TF-IDF

Para poder representar el contenido de cada Tweet como un número que más adelante pueda ser procesado por un algoritmo, es que se ha utilizado Tf-IDF (Term Frequency - Inverse document frequency) el cual , corresponde a un valor resultante de la ponderación entre la frecuencia de aparición de un término en un Tweet a clasificar y su frecuencia inversa de aparición en la bolsa de palabras. De esta manera, se puede conocer el "peso" de cada Tweet, y posterior a ello, obtener su polaridad (intención de voto positiva o negativa).

$$tf(t,d) = \frac{f(t,d)}{max\,\{f(w,d) : w\,\epsilon\,d\}} \tag{3}$$

$tf(t,d)$ corresponde a la frecuencia normalizada del término en el Tweet. Dicho de otra forma, la “cantidad de veces" que aparece dicha palabra en la publicación.

$$idf(t,D) = log\frac{|D|}{|\{d\,\epsilon\,D : t\,\epsilon\,d\}|} \tag{4}$$

$idf(t,D)$ corresponde a la frecuencia inversa de documento, siendo una medida, para determinar si el término es común o no, en la bolsa de palabras. Se obtiene dividiendo el número total de términos contenidos en la bolsa de palabras, por el número de Tweets que la poseen, posterior a ello, se toma el logaritmo de ese cociente.Finalmente se obtiene el valor TF-IDF(t,d,D) ponderando (3) con (4).

$$TF - IDF(t,d,D) = tf(t,d) \times idf(t,D) \tag{5}$$

Un peso alto en TF-IDF se alcanza con una elevada frecuencia de término (en el Tweet correspondiente) y una pequeña frecuencia de ocurrencia del término en la bolsa de palabras. Como el cociente dentro de la función logaritmo del idf es siempre mayor o igual que 1, el valor del idf (y del TF-IDF) es mayor o igual que 0. Cuando un término aparece muchas veces en la bolsa de palabras, el cociente dentro del logaritmo se acerca a 1, ofreciendo un valor de idf y de TF-IDF cercano a 0.

6.2.2. Representación de Tweets por medio de Word2Vec

Correspondiente a la segunda forma de representación que se evaluará, descubierta hace poco tiempo y basada en principios postulados por Tomás Mikolov, desde la librería Gensim a través de Python. Corresponde a un algoritmo de procesamiento de datos, en el cual se aplica la técnica de aprendizaje no supervisado, representando a un Tweet como un vector de palabras, en donde

cada término del texto es representado por un vector de un bit activo (hot vector) (et al, 2013) , su característica principal es que permite poder considerar a una palabra de entrada y a sus vecinas, por lo que se puede incorporar algo del “contexto“ de la frase que trabaja en base a frecuencias de aparició, lo cual se diferencia de Tf-Idf en el sentido de que este último, se basa meramente en frecuencias de aparición de palabras y que en casos en donde pudiesen presentarse muchas palabras "negativas", se obtendría un resultado "negativo"pudiendo ser otro el contexto de la frase, lo cual podría llevar a clasificaciones erroneas. Pero, los resultados de Word2Vec en este caso, no aseguran un alto rendimiento, es por eso, que se pondrá a prueba a continuación.

Para ejemplificar su funcionamiento, consideremos la palabra “voto” de la frase “Yo no pienso dar mi voto a Piñera”, la cual se representaría de la siguiente forma:

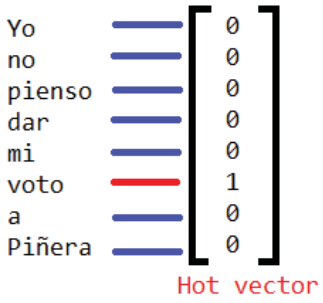


Figura 2: Representación de vector de un bit o Hot vector.

Tal como se observa, un Hot vector, almacena solo un bit “1” correspondiente a la palabra de entrada.

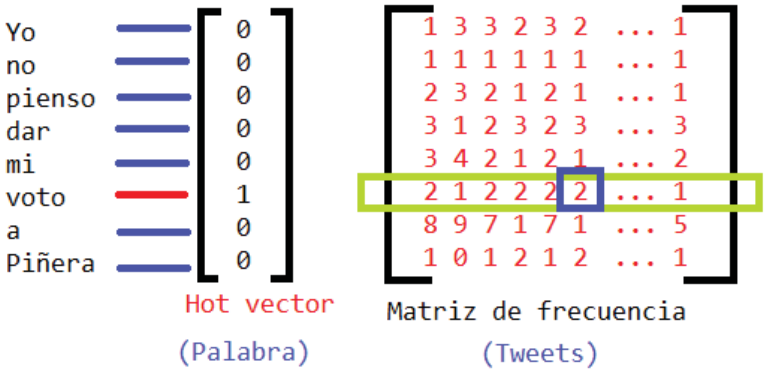


Figura 3: Representación de búsqueda de elemento mediante matriz de frecuencia y vector de bit activo

Esto permite realizar una búsqueda de la palabra en un Tweet y obtener su frecuencia de aparición correspondiente, por medio de una red neuronal en donde la capa oculta es utilizada para realizar una búsqueda de la palabra en un Tweet.

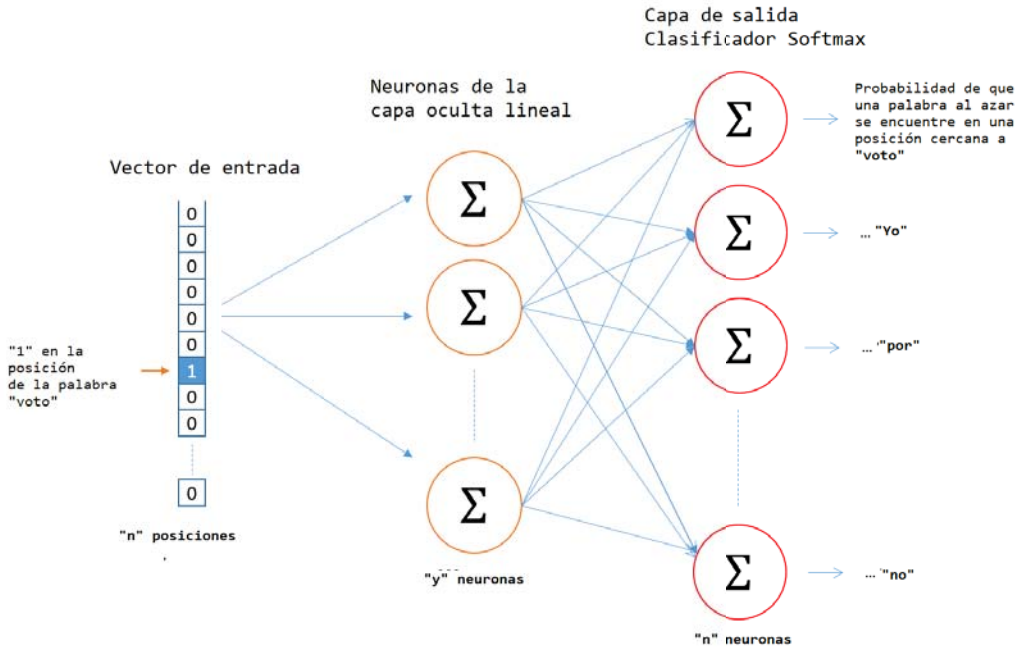


Figura 4: Funcionamiento de red neuronal con vector de entrada.

Este algoritmo es capaz de trabajar mediante una palabra de entrada y tener en cuenta a las que la acompañan, es decir, a aquellos términos que están cierta posición a la izquierda o derecha de ésta; En particular, se trabaja a las palabras vecinas con un cierto valor de cercanía, hablamos de un valor que es definido dependiendo de las necesidades del problema y que por defecto es típicamente 5, es decir, 5 palabras a la izquierda o derecha (10 en total). El algoritmo obtiene la probabilidad de que una palabra sea vecina de otra en base a la cercanía definida para el problema en específico.

Es así como por ejemplo (y asumiendo que se refiere al nombre "Unión Soviética de Rusia"), las palabras "Soviética" y "Rusia", se encontrarán con una probabilidad más alta de cercanía a diferencia de otras palabras que pudiesen ser ingresadas como "casa" o "roja", que tendrán una probabilidad más baja. La representación de estos datos es binaria, por lo que la palabra que ingresa se identifica con un "1" en el vector de palabras, y con ceros las restantes; Lo último permite que el tamaño del universo de trabajo sea menor, debido a que si tuviésemos 10.000 palabras almacenadas dentro de nuestro diccionario, y se almacenaran Tweets con 300 características en cada uno, tendríamos que trabajar con una matriz de 10.000 x 300, a diferencia de la primera opción, que toma el vector de ceros y uno correspondiente y lo multiplica por el vector de 10.000 palabras, obteniendo finalmente un vector, lo cual reduce considerablemente el tamaño de elementos a procesar.

El valor de cercanía de las palabras vecinas puede ser modificado dependiendo de la naturaleza del problema. Esta red además de almacenar los pares, es capaz de discriminar cuando dos Tweets, o dos vectores de palabras son similares, cuando contienen relaciones entre mismas palabras, es por eso que para el caso del Tweet "Yo no quiero dar mi voto a Piñera", y el Tweet "Por Piñera

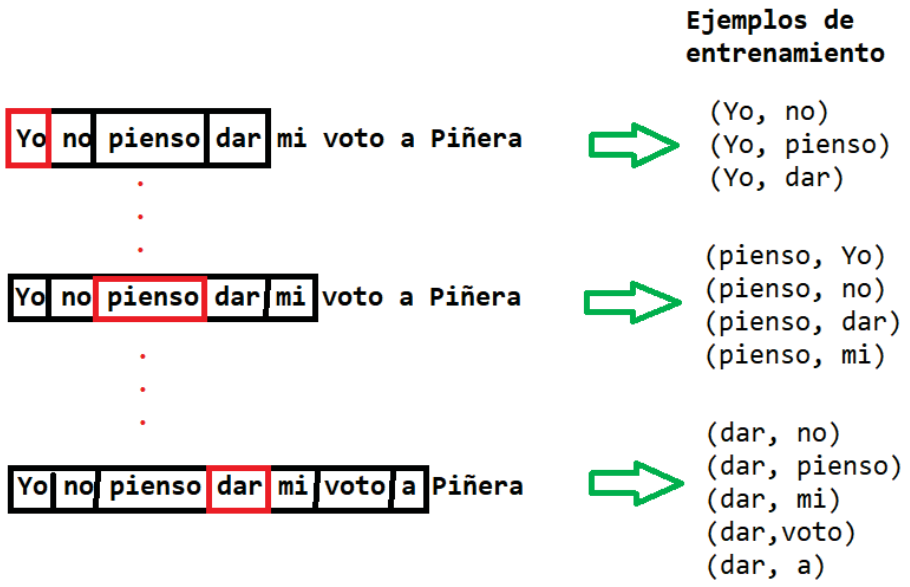


Figura 5: Ejemplo de lógica de palabras cercanas

yo no voto porque no es de mi gusto", se almacenarían las palabras que darían origen a la frase "Yo no quiero dar mi voto a Piñera por porque no es de mi gusto". Lo cual, a diferencia de una representación del tipo TF-IDF, permite realizar una identificación del contexto de una palabra o frase, permitiendo clasificar por el sentido del Tweet más que por frecuencia de aparición.

6.2.2.1. Representación de Tweets por medio de Doc2Vec

Este tipo de representación proviene de Word2Vec, a través de la misma librería Gensim. A diferencia de su origen, el cual, enviaba un Hot vector (vector de solo un bit activo) a la capa intermedia, o capa oculta de la red neuronal, para que esta realizara una búsqueda de la palabra en todo el diccionario almacenado, Doc2Vec almacena todo el Tweet en un mismo vector de tamaño N, dando otra alternativa de representación a las palabras contenidas dentro de un Tweet. El rendimiento de este es muy similar al de Word2Vec, por lo que en base a los resultados de este ultimo es que se pueden estimar mas menos los de Doc2Vec.

La lógica de esta técnica es similar a Word2Vec, puesto que, también el algoritmo es capaz de determinar cuales son las palabras cercanas o vecinas a un término de entrada.

6.3. K-Fold Cross Validation y segmentación de datos

Tras haber representado las palabras de una manera numérica, se obtiene la matriz de términos (representados por "pesos") de cada tweet. Dicha matriz de valores, está acompañada a un vector de etiquetas (el cual se obtuvo a gracias a la clasificación de Tweets realizada manualmente por ayudantes, por medio de la herramienta Analitic PRO).

$$T, E = \begin{pmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,140} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,140} \\ \vdots & \vdots & \vdots & \vdots \\ t_{n,1} & t_{n,2} & \cdots & t_{n,140} \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_4 \end{pmatrix} \tag{6}$$

En donde cada fila de la matriz T, corresponde a cada Tweet, y sus columnas, las palabras que lo componen. Y a su vez, la matriz E, es la etiqueta asociada al Tweet de la fila en la que se posiciona.

Debido a que se utilizará la técnica de aprendizaje supervisado, es que se dividirá una porción de los datos obtenidos, dada a clasificación por etiquetas de positivo y negativo, en 80 porciento de entrenamiento y 20 porciento de pruebas.

Para intentar disminuir cierto nivel de sesgo en la predicción de los datos, debido a cierta tendencia de los datos de entrenamiento a ser positivos o negativos, es que se realizó un Shuffle (reordenamiento de términos al azar) sin perder la relación de cada fila de la matriz E con sus etiquetas correspondientes.

K-fold corresponde a una técnica de validación cruzada, utilizada para realizar entrenamiento y pruebas en algoritmos de clasificación y predicción de eventos. Ayuda a la disminución del sesgo proveniente de la acumulación de tendencias positivas o negativas. Consiste en realizar ciclos iterativos, en donde cierto numero de datos es probado con secciones de los datos de entrenamiento, hasta alcanzar la totalidad de estos (Figura 2), trabajando en base a la media aritmética. Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados.

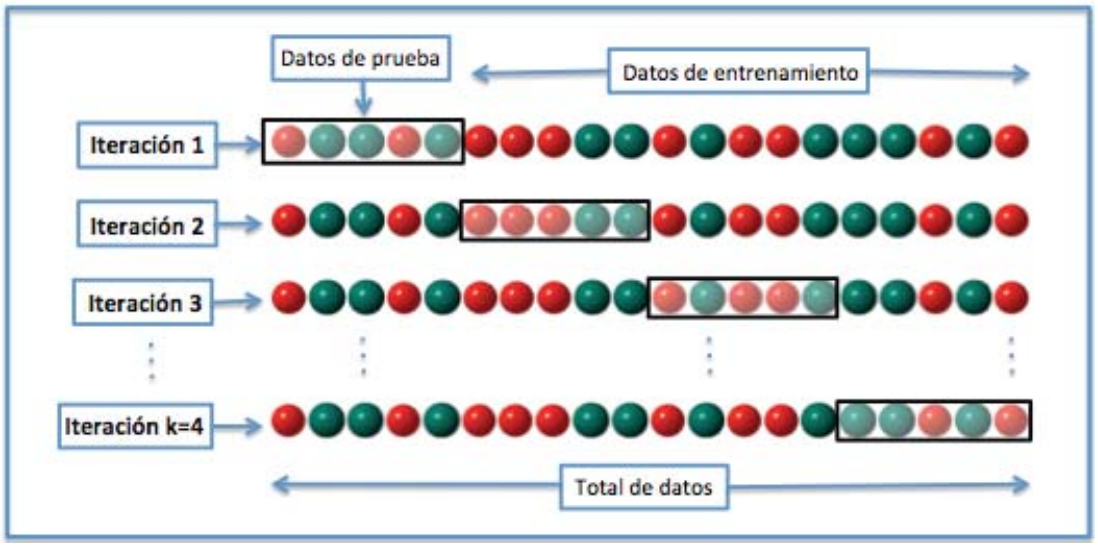


Figura 6: Funcionamiento de técnica de validación cruzada K-Fold.

6.4. Clasificadores de votación

Los clasificadores de votación son estimadores que se agrupan y se entrenan todos con un mismo conjunto de datos, y a su vez se les entrega el mismo conjunto para realizar las estimaciones/clasificaciones. Para esto existen dos tipos de clasificadores de votación:

- Clasificadores de votación por regla de mayoría, los cuales la etiqueta final para la muestra a categorizar se elige por la mayoría de etiquetas entregadas para dicha muestra.
- Clasificadores de votación por votación suave, los cuales predicen la categoría de la muestra mediante la elección de la probabilidad acumulada de las categorías por todos los clasificadores.

En el desarrollo del proyecto se utilizaron dos clasificadores de votación, los cuales consideraron todos los estimadores descritos anteriormente (Árboles de decisión, Random Forest, AdaBoost y SVM linear).

6.5. Procesamiento de mensajes

Por medio de aplicaciones de clasificación que se basan en las palabras utilizadas en la publicación, se pudieron obtener un porcentaje de marcas, éstas palabras se encuentran en una base de datos de términos, la cual almacena todas aquellas que muestran “agrado” y “desagrado”, y a su vez se asocia el contexto de la frase correspondiente (Ej. “Me parece pésimo que no existan más candidatos como Sebastían Piñera”, podría ser interpretada como “negativo”, debido a que dentro del contenido del Tweet existía la palabra “pésimo”, cuando en realidad, el comentario hacía una referencia positiva al candidato). Debido a lo anterior es que se han determinado los siguientes escenarios de ocurrencia:

- Que el Tweet en cuestión haga referencia a un candidato en específico, pero el contenido esté ilegible, o bien con links con referencias a otros sitios (spam).
- Comentarios fuera del contexto político, que opinen de temas no relacionados al candidato
- Opiniones con abreviaciones y modismos clásicos chilenos. La informalidad permite que el lenguaje utilizado no es el más apropiado, palabras sin acentuaciones, abreviaturas o garabatos (Ej. “xq” en vez de “por qué” o “bkn” en vez de “genial”, etc).
- Tweets que posean tanto palabras positivas como negativas o con lectura ambigua (Ej. “El discurso del candidato fue para nada algo bueno”).

Haciendo alusión a la gran cantidad de modismos que existen hoy en día en Chile, es que se requirió de un equipo de ayudantes que pudiese marcar, el porcentaje de Tweets restante, de forma manual en la herramienta, trabajo que tomó unos cuantos meses, debido al gran volumen de Tweets que están involucrados, considerando tambipen aquellas publicaciones que poseen ironías, o bien, manifiestan su agrado por medio de expresiones que son asociadas por el clasificador como de desagrado, y en realidad el trasfondo del comentario está asociado a algo positivo; Lo anterior puede producir datos fallidos, ya que el algoritmo utilizado primeramente no era capaz de detectar aquellas ironías o ambigüedades, es por ello que se prefirió realizar una clasificación "manual"de éstas publicaciones. De esta manera el equipo fue clasificado un gran volumen de Tweets, para asegurar una mayor fiabilidad de los datos a trabajar.

Serán catalogados todos aquellos Tweets que posean un sentido lógico, y puedan ser calificados como positivos o negativos tanto por el algoritmo de clasificación, como por el equipo de trabajo asociado. El resto de las publicaciones, se considerarán como "ruido", o bien, datos que serán clasificados como "neutro" o sin categoría. Se estima que aproximadamente entre un 70 u 80 porciento de los datos obtenidos corresponden a ello.

Además de la clasificación de "positivo" y "negativo", se posee también otro tipo de clasificación que se puede añadir a la anterior; Si es que el Tweet en cuestión hace alusión a:

- Intención de voto explícito a favor del candidato (Ej. "Yo en estas primarias 2017, voto Kast", "Piñera tienes mi voto").
- Intención de voto explícita en contra del candidato (Ej. "No pienso votar por Kast", "Piñera perdiste mi voto").

Esta clasificación añadida, es aplicada por medio de un botón adicional a los de positivo y negativo en la herramienta, siendo de utilidad para conocer de manera más certera el porcentaje de población que está decidido a votar por cierto candidato. Existen, según lo estudiado, dos tipos de Tweets:

- Provenientes de una opinión propia, emitida en cualquier momento.
- Realizados posterior a algún evento de comunicación masiva en el cual se manifestaron ciertos candidatos.

6.6. Uso de GridSearchSV

El rendimiento de cada algoritmo a utilizar depende de los valores de configuración que se utilicen para el procesamiento de los datos. Estos elementos juegan un rol fundamental en los valores de las métricas que se obtendrán finalmente. Una mejor configuración en el algoritmo puede conllevar un aumento en el desempeño de este mismo para la clasificación de Tweets.

Dentro de la librería SkitLearn se tienen distintas funcionalidades, una de ellas es GridSearchSV, la cual permite a través de la obtención de una grilla de parámetros requeridos, el cálculo de los mejores valores de configuración con los que un algoritmo de clasificación en particular podría rendir mejor.

De manera genérica se solicita a GridSearchSV la obtención de la mejor configuración para un algoritmo en particular según un rango dado:
$C : [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$,
kernel : ['rbf' , 'linear' , 'sigmoid'] ,
gamma :[0.1 , 0.2 , 0.3 , 0.4 , 0.5]
Lo anterior es un ejemplo de como se le entrega a la función la grilla de parámetros que pueden poseer las distintas variables de configuración de un algoritmo en particular.

El parámetro gamma define hasta dónde llega la influencia de un solo grupo de entrenamiento. Un gamma bajo implica una influencia lejana de clasificación con respecto a la frontera de decisión, por el contrario, un gamma alto implica una influencia cercana de clasificación en relación a esta frontera. Dependiendo del modelo y el algoritmo a utilizar, el valor puede variar, la función

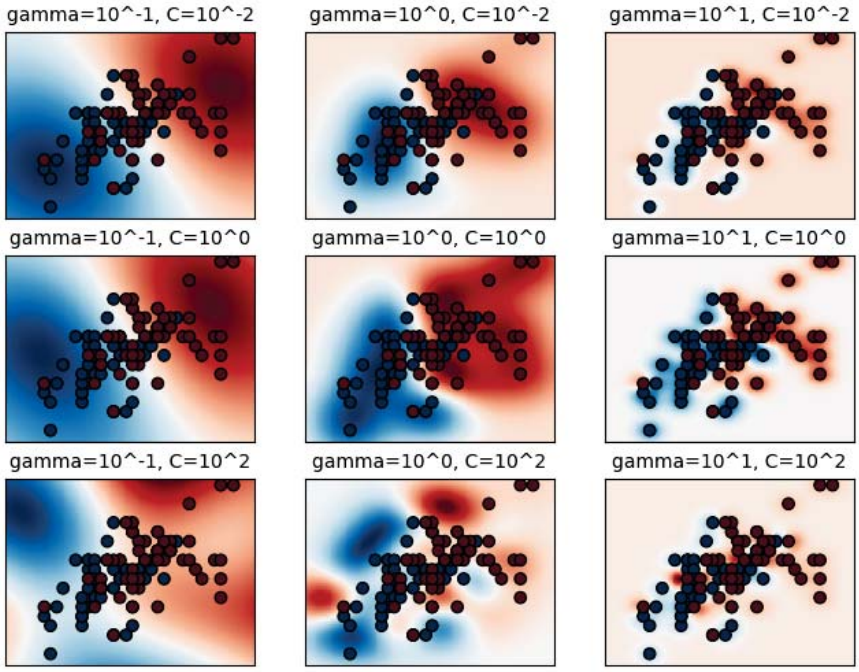


Figura 7: Ejemplo de representación de valor Gamma para SVM radial (rbf)

entregará el número que permita obtener mejores resultados en las métricas.

El parámetro C intercambia errores de clasificación de ejemplos de entrenamiento contra la simplicidad de la superficie de decisión. Una C baja hace que la superficie de decisión sea lisa, mientras que una C alta tiene como objetivo clasificar correctamente todos los ejemplos de entrenamiento dando libertad al modelo para seleccionar más muestras como vectores de soporte.

El kernel corresponde a la función matemática utilizada para la transformación de los datos dentro de la máquina que se vaya a utilizar. Dentro de los que existen se tienen; lineal, de función de base radial (rbf), sigmoideo.

El kernel lineal, proporciona una función lineal matemática que no permite al algoritmo aprender comportamientos distintos, sigue una linea de comportamiento.

El kernel RBF, utiliza una función de base radial que permite generar un modelo predictivo para una o más variables dependientes basado en una o más variables predictoras. Obteniendo un plano que posee pretuverancias en forma de campana, donde el valor que mas se acerca al esperado se encuentra posicionado en el centro a la altura máxima. Los valores que se asemejen al mejor obtenido se situarán al rededor del radio de cada uno de los mejores locales.

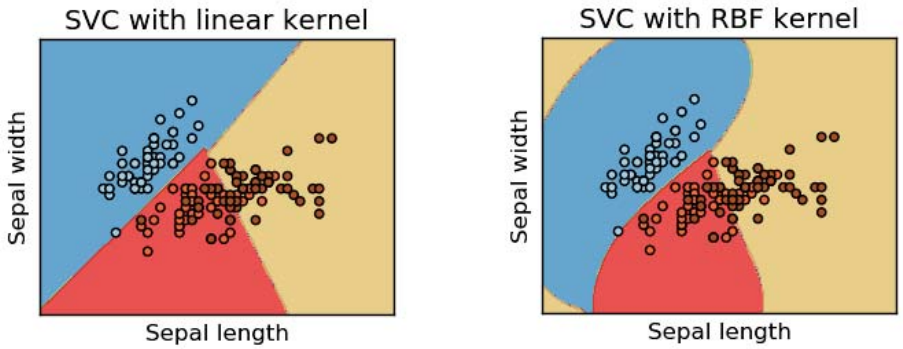


Figura 8: Representación de comportamiento de función lineal versus funcion de base radial en máquina SVC

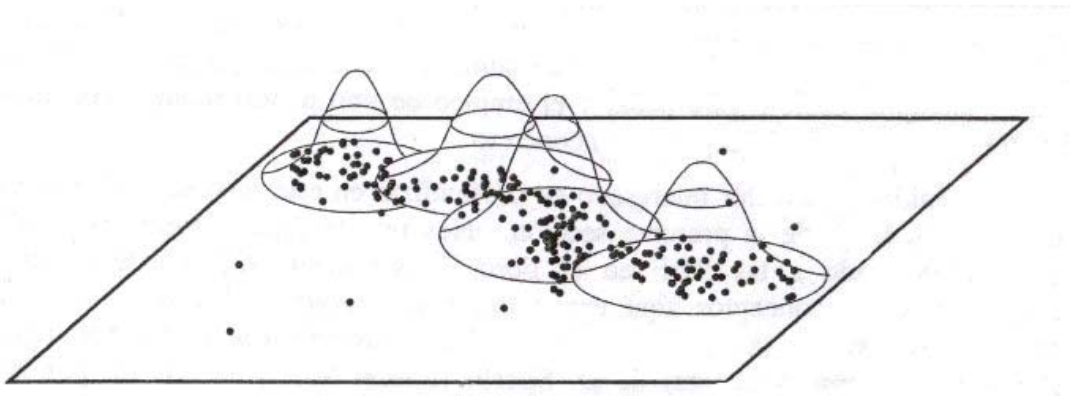


Figura 9: Representación de comportamiento de función de base radial

7. Modelos a utilizar

Se expondrá las estrategias a utilizar para el proceso de captación de datos a través de análisis de sentimientos, los cuales serán utilizados para predecir las cifras y datos de interés de los candidatos postulantes a la presidencia en Chile

7.1. De predicción mediante menciones positivas

Para poder entrenar este modelo y realizar las predicciones, primero se define una ventana de diez días antes de las votaciones, en donde para generar el conjunto a predecir, se obtienen todos los tweets relacionados a todos los candidatos que van a participar de las primarias en dicha ventana. Luego para poder generar el conjunto de entrenamiento de los modelos, se definió la fecha del primero de junio (01/06/2017), como fecha de inicio para la construcción del conjunto de tweets por candidatos, por consiguiente, la fecha final para la construcción correspondería al día anterior del inicio de los diez días a predecir hasta las 23:59:59.

Este modelo para obtener la intención de voto, realiza la suposición de que un comentario positivo refleja la intención de voto independiente del usuario, sin contar en cuenta ni menciones negativas, ni las menciones neutrales. Esto significa que, si la mención positiva corresponde al candidato 1, este mensaje se le asigna una nueva etiqueta la cual indica la intención de voto para ese candidato (ej: pos-Candidato1). Cabe destacar que este modelo es susceptible al sesgo producido por posibles cuentas bots y RT (re-tweets), debido a que los bots pueden generar y replicar una mayor cantidad de mensajes que pueden ser catalogados como positivos por los expertos encargados de categorizar los tweets para cada candidato. Finalmente, la predicción se realiza tomando el conjunto de predicción y clasificando cada tweet en este conjunto según las etiquetas de intención de voto involucradas. Finalmente, estos mensajes ya etiquetados se tabulan y se calculan las proporciones de votos por candidato según el modelo.

7.2. De relación de mensajes positivos/negativos

Este enfoque consiste en la construcción de cinco clasificadores, los cuales están especializados para cada uno de los independiente de la coalición a la cual el candidato pertenezca. A diferencia del modelo anterior, estos modelos en vez de ver el volumen bruto de menciones positivas por candidatos, se enfocan principalmente en detectar la posible intención de votos para cada usuario que haya emitido algún comentario de uno o más candidatos.

Para poder entrenar estos modelos, se realiza el mismo proceso para generar la base de tweets de entrenamiento que en el modelo descrito anteriormente, luego lo primordial es encontrar a todos los usuarios que hayan emitido mensajes para uno o más candidatos y luego ver si la resta entre mensajes positivos y negativos es mayor a cero para cada candidato (implicando que los mensajes positivos son más que los negativos). Si lo mencionado anteriormente se cumple, todos los mensajes de ese usuario son catalogados como la intención de voto para el candidato, de lo contrario, se marcan como una no-intención de voto. Esto mismo se realiza para todos los candidatos, generando un conjunto de datos y modelos especializados por candidato.

8. Predicción de datos futuros

Por medio de los datos de entrada que constituyeron el aprendizaje adquirido por la máquina, se pueden realizar estimaciones respecto a cual será el valor estimativo de un evento futuro. Por medio de la predicción se podrá dar a conocer el porcentaje de aprobación o desaprobación de un candidato en base a la opinión de los usuarios de Twitter.

A través de la obtención individual de cada uno de los candidatos se tendrá que realizar una comparativa general, para generar una gráfica o estadística que permita dar a conocer cual será el candidato con mayor preferencia, tal como hoy en día realizan encuestas como ADIMARK, CADEM o CEP, mostrando gráficas como la siguiente:

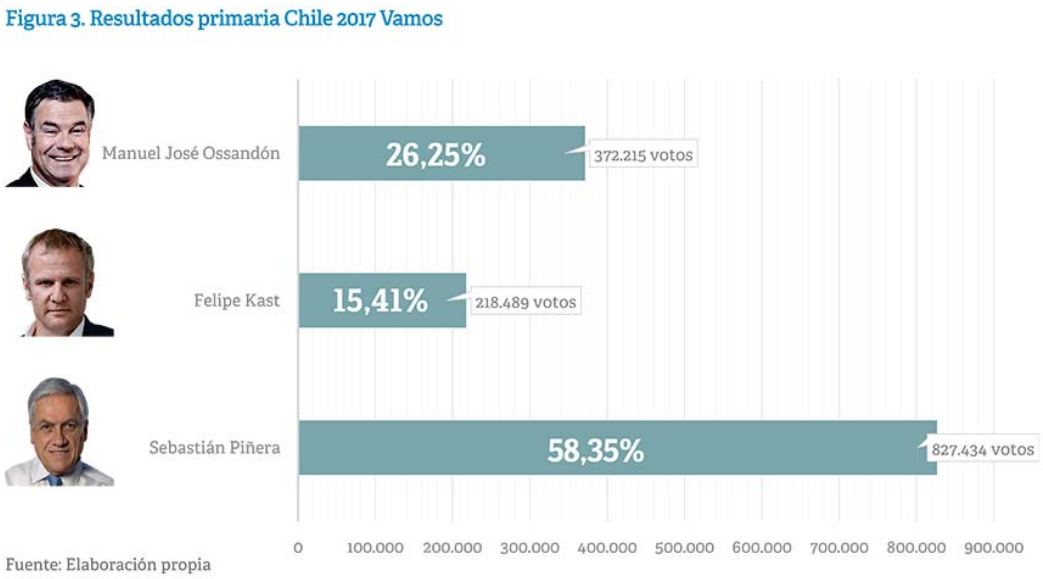


Figura 10: Ejemplo de predicción electoral encuesta CEP 2017 para primarias presidenciales

La predicción y sus resultados dependen de que tan eficiente sea el clasificador, ya que según lo que aprenda la máquina, serán los resultados que arrojará con futuras entradas. Es por ello que es fundamental la elección de un algoritmo de alto rendimiento para la predicción final.

9. Métricas a utilizar

En esta sección se presentarán las métricas que serán utilizadas para la evaluación del rendimiento de la solución propuesta, para realizar con ello, una comparación en base a las cifras obtenidas en el proceso de primarias 2017.

9.1. F1-Score (Puntaje F1)

En el análisis estadístico de la clasificación binaria , el puntaje F1 (también puntaje F o medida F) es una medida de la precisión obtenida en una prueba. Considera tanto la precisión como la exhaustividad de la prueba, para calcular la puntuación: "Precision" es el número de resultados positivos correctos dividido por el número de todos los resultados positivos:

$$Precision = \frac{\{palabras\,relevantes\} \cap \{palabras\,totales\}}{\{palabras\,totales\}} \tag{7}$$

"Exhaustividad" es el número de resultados positivos correctos dividido por el número de resultados positivos positivos resultados que deberían haber sido devueltos:

$$Exhaustividad = \frac{\{palabras\,relevantes\} \cap \{palabras\,totales\}}{\{palabras\,relevantes\}} \tag{8}$$

El puntaje F1 es el promedio armónico de la precisión y la exhaustividad , donde un puntaje F1 alcanza su mejor valor en 1 (precisión perfecta y exhaustividad) y peor en 0:

$$F1 - Score = 2\frac{Precision \cdot Exhaustividad}{Precision + Exhaustividad} \tag{9}$$

9.2. Accuracy (Exactitud)

Función que calcula la precisión , ya sea la fracción (valor predeterminado) o el recuento (normalizacion = Falso) de las predicciones correctas. En la clasificación multietiqueta, la función devuelve la precisión del subconjunto. Si todo el conjunto de etiquetas predichas para una muestra coincide estrictamente con el verdadero conjunto de etiquetas, entonces la precisión del subconjunto es 1.0; de lo contrario, es 0.0. Si $\widehat{y}$ es el valor predicho de la muestra x el valor verdadero correspondiente, la fracción de predicciones correctas n se define como:

$$Accuracy(x, \widehat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} 1(x = \widehat{y}) \tag{10}$$

9.3. ROC-AUC Score (Puntaje de área bajo la curva)

Este valor permite la representación de la razón o ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos) según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo). ROC también puede significar Relative Operating Characteristic (Característica Operativa Relativa) porque es una comparación de dos características operativas (VPR y FPR) según cambiamos el umbral para la decisión.

El análisis de la curva ROC, o simplemente análisis ROC, proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente de (y antes de especificar) el coste de la distribución de las dos clases sobre las que se decide. La curva ROC es también independiente de la distribución de las clases en la población. El análisis ROC se relaciona de forma directa y natural con el análisis de coste/beneficio en toma de decisiones diagnósticas.

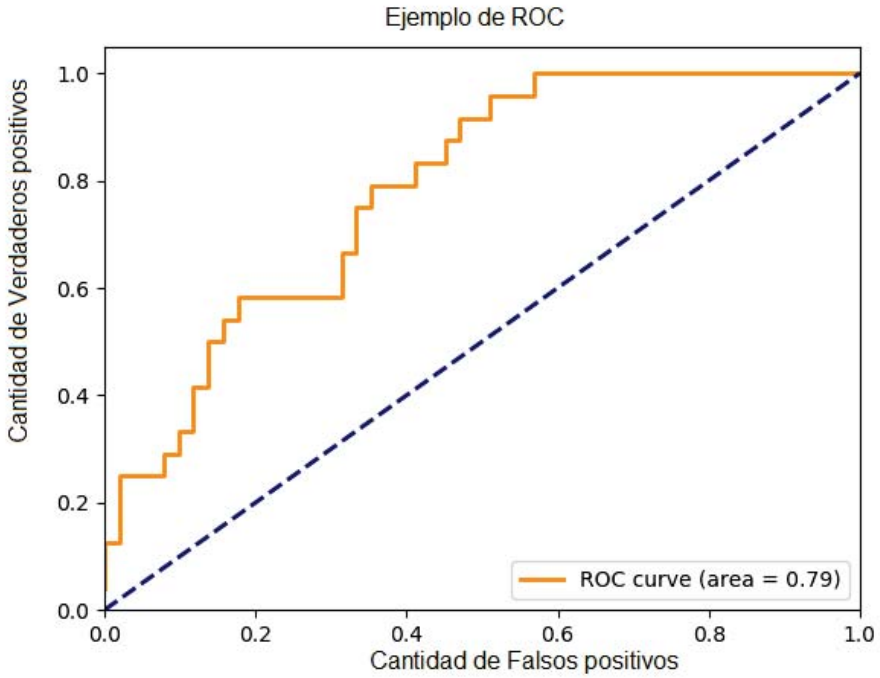


Figura 11: Ejemplo gráfico de representación de ROC-AUC Score

10. Clasificación por medio de TF-IDF

Por medio del uso de 40 iteraciones de experimentación, además de la utilización de la técnica k-fold cross validation para el proceso de aprendizaje y prueba de los datos con k=4 (4 folds), se pudieron obtener las métricas de rendimiento de clasificación con el set de datos de estudio correspondiente al candidato Sebastián Piñera, en el periodo comprendido entre el 5 de Mayo del 2017 y 1 de Julio del 2017 (periodo de primarias electorales), para cada uno de los algoritmos propuestos; Máquina de soporte vectorial, árbol de decisión y Naive Bayes (o Bayes ingenuo).

10.1. Mejor configuración para máquina de soporte vectorial (SVM) dado un rango de prueba

Por medio de la utilización de GridSearchCV, a través de SkLearn de Python 3.0, dada la siguiente grilla de parámetros:

C=[1,2,3,4,5,6,7,8,9,10]
kernel=[lineal,rbf,sigmoideo]
gamma=[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]

En una fase inicial (representando los datos por medio de TF-IDF) se pudo obtener que la mejor configuración a partir de los valores dispuestos anteriormente, viene dada por: C=9, gamma=0.1 y kernel RBF. Por lo que serán estos los valores que se utilizaron para obtener los desempeño que entregan las métricas. Cabe mencionar que estos parámetros pueden mejorar más el rendimiento del algoritmo conforme se vaya abarcando un rango mas amplio de valores.

10.2. Resultados de métricas para SVM

Luego de elegir la mejor configuración para la máquina, dentro de un rango de prueba dispuesto, se ejecutó el algoritmo, siendo el que otorgó las mejores métricas finales de clasificación.

Con métricas de desempeño que superan a los otros algoritmos, se obtuvo que de los 557 elementos que conformaban la clase "Positivo", se tuvieron 525 aciertos y 32 fallos. Por otro lado para la segunda clase "Negativo"que poseía 347 elementos, se pudieron clasificar adecuadamente 290 de ellos, fallando en 57. Promediando todos los resultados de los experimentos se posee un puntaje F1 del 90 porciento. A su vez, un área bajo la curva (roc auc score) del 89 porciento, y una exactitud del 90 porciento.

10.3. Resultados de métricas para Árboles de decisión

El algoritmo de árbol de decisión destacó en relación a la demora de ejecución, siendo el más rápido en procesar y generar resultados en comparación a los otros dos. Su rendimiento en comparación fue más bajo, pero aún así posee métricas que superan el 79 porciento.

Tras la ejecución del algoritmo de árbol de decisión, se obtuvo que de los 553 elementos que conformaban la clase "Positivo", se tuvieron 460 aciertos y 93 fallos. Por otro lado para la segunda clase "Negativo"que poseía 351 elementos, se pudieron clasificar adecuadamente 270 de ellos, fallando en 81. Promediando todos los resultados de los experimentos se posee un puntaje F1 del 80 porciento. A su vez, un área bajo la curva (roc auc score) del 80 porciento, y una exactitud del 80 porciento.

RESUMEN GENERICO , F1-SCORE Y MATRIZ DE CONFUSIÓN
AL PRIMER EXPERIMENTO REALIZADO CON K=5(FOLDS)

	precision	recall	f1-score	support
0	0.90	0.94	0.92	557
1	0.90	0.84	0.87	347
avg / total	0.90	0.90	0.90	904

Matriz de confusión: [[525 32]
[57 290]]

PROMEDIO DE METRICAS POST 40 EXPERIMENTOS

promedio F1 : 0.900789606451
promedio ROC_AUC : 0.889142120975
promedio exactitud : 0.901548672566

Figura 12: Resultado métricas obtenidas por medio de Tf-Idf con SVM

RESUMEN GENERICO , F1-SCORE Y MATRIZ DE CONFUSIÓN
AL PRIMER EXPERIMENTO REALIZADO CON K=5(FOLDS)

	precision	recall	f1-score	support
0	0.82	0.86	0.84	536
1	0.78	0.72	0.75	368
avg / total	0.80	0.81	0.80	904

Matriz de confusión: [[462 74]
[102 266]]

PROMEDIO DE METRICAS POST 40 EXPERIMENTOS

promedio F1 : 0.80393780311
promedio ROC_AUC : 0.792383192732
promedio exactitud : 0.805309734513

Figura 13: Resultado métricas obtenidas por medio de Tf-Idf con árbol de decisión

10.4. Resultados de métricas para Naive Bayes

A pesar de que Naive Bayes es considerado un clasificador que no destaca por sus métricas finales, sino que más bien por la característica que da a los datos de poder trabajarlos de manera descriptiva, para poder realizar análisis con respecto a las palabras más destacadas o importantes. El algoritmo arrojó resultados postivos con el conjunto de datos de prueba, obteniendo cifras por sobre el 85 porciento.

RESUMEN GENERICO , F1-SCORE Y MATRIZ DE CONFUSIÓN
AL PRIMER EXPERIMENTO REALIZADO CON K=5(FOLDS)

	precision	recall	f1-score	support
0	0.85	0.97	0.91	539
1	0.95	0.75	0.83	365
avg / total	0.89	0.88	0.88	904

Matriz de confusión: [[524 15]
[93 272]]

PROMEDIO DE METRICAS POST 40 EXPERIMENTOS

promedio F1 : 0.877415349465
promedio ROC_AUC : 0.858688082954
promedio exactitud : 0.880530973451

Figura 14: Resultado métricas obtenidas por medio de Tf-Idf con Naive Bayes

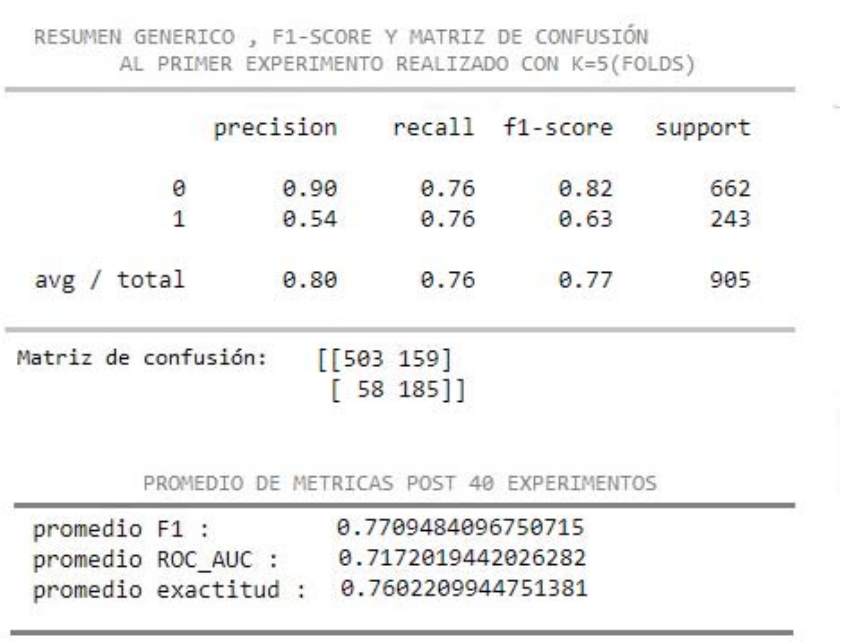
De lo anterior se puede apreciar que bajo los experimentos realizados, se obtuvo que de los 539 elementos que conformaban la clase "Positivo", se tuvieron 524 aciertos y 15 fallos. Por otro lado para la segunda clase "Negativo"que poseía 365 elementos, se pudieron clasificar adecuadamente 272 de ellos, fallando en 93. Promediando todos los resultados de los experimentos se posee un puntaje F1 del 88 porciento. A su vez, un área bajo la curva (roc auc score) del 86 porciento, y una exactitud del 83 porciento.

11. Clasificación por medio de Word2Vec

Por medio del uso de 40 iteraciones de experimentación, además de la utilización de la técnica k-fold cross validation para el proceso de aprendizaje y prueba de los datos con k=5 (5 folds), y una rango de vecinos “w“ de tamaño 5 , se pudieron obtener las métricas de rendimiento de clasificación con el mismo set de datos de estudio anterior, correspondiente al candidato Sebastián Piñera, en el periodo comprendido entre el 5 de Mayo del 2017 y 1 de Julio del 2017 (periodo de primarias electorales), para cada uno de los algoritmos propuestos; Máquina de soporte vectorial, árbol de decisión y Random Forest.

11.1. Resultados de métricas para SVM

Al igual que en el caso anterior de clasificación por medio de Tf-Idf, quien obtuvo las cifras más altas en relación al trabajo realizado fue máquina de soporte vectorial SVM.



RESUMEN GENERICO , F1-SCORE Y MATRIZ DE CONFUSIÓN
AL PRIMER EXPERIMENTO REALIZADO CON K=5(FOLDS)

	precision	recall	f1-score	support
0	0.90	0.76	0.82	662
1	0.54	0.76	0.63	243
avg / total	0.80	0.76	0.77	905

Matriz de confusión: [[503 159]
[58 185]]

PROMEDIO DE METRICAS POST 40 EXPERIMENTOS

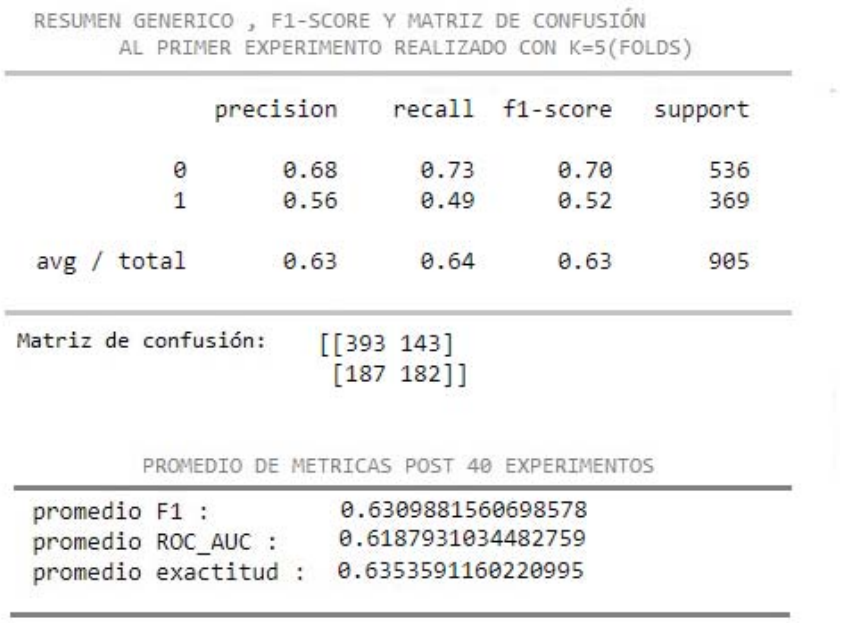
promedio F1 : 0.7709484096750715
promedio ROC_AUC : 0.7172019442026282
promedio exactitud : 0.7602209944751381

Figura 15: Resultado métricas obtenidas por medio de Word2Vec con SVM

Con métricas de desempeño que superan a los otros algoritmos, se obtuvo que de los 243 elementos que conformaban la clase "Positivo", se tuvieron 185 aciertos y 58 fallos. Por otro lado para la segunda clase "Negativo"que poseía 662 elementos, se pudieron clasificar adecuadamente 503 de ellos, fallando en 159. Promediando todos los resultados de los experimentos se posee un puntaje F1 del 77 porciento. A su vez, un área bajo la curva (roc auc score) del 71 porciento, y una exactitud del 76 porciento.

11.2. Resultados de métricas para Árboles de decisión

El algoritmo de árbol de decisión destacó en relación a la demora de ejecución, siendo el más rápido en procesar y generar resultados en comparación a los otros dos. Su rendimiento en comparación fue más bajo, pero aún así posee métricas que superan el 60 porciento.



RESUMEN GENERICO , F1-SCORE Y MATRIZ DE CONFUSIÓN
AL PRIMER EXPERIMENTO REALIZADO CON K=5(FOLDS)

	precision	recall	f1-score	support
0	0.68	0.73	0.70	536
1	0.56	0.49	0.52	369
avg / total	0.63	0.64	0.63	905

Matriz de confusión: [[393 143]
[187 182]]

PROMEDIO DE METRICAS POST 40 EXPERIMENTOS

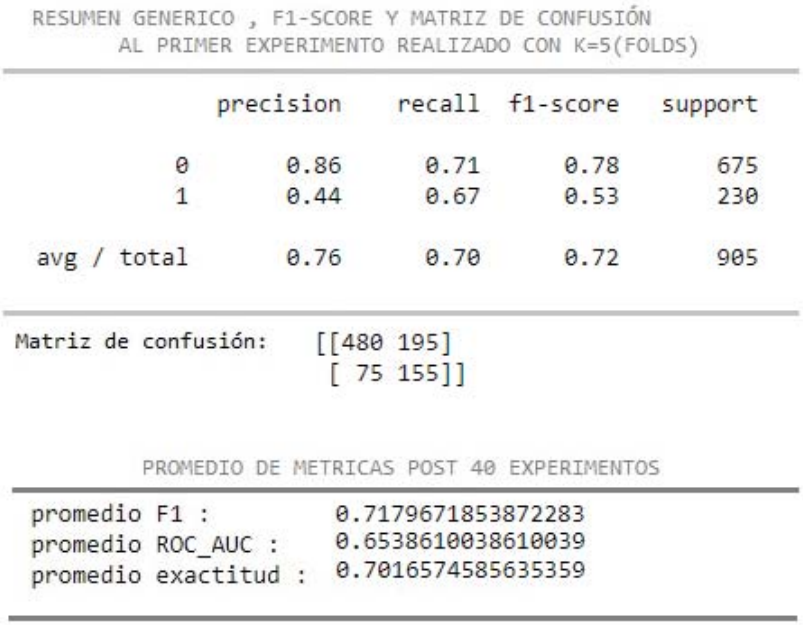
promedio F1 :	0.6309881560698578
promedio ROC_AUC :	0.6187931034482759
promedio exactitud :	0.6353591160220995

Figura 16: Resultado métricas obtenidas por medio de Word2Vec con árbol de decisión

Tras la ejecución del algoritmo de árbol de decisión, se obtuvo que de los 369 elementos que conformaban la clase "Positivo", se tuvieron 182 aciertos y 187 fallos. Por otro lado para la segunda clase "Negativo"que poseía 536 elementos, se pudieron clasificar adecuadamente 393 de ellos, fallando en 143. Promediando todos los resultados de los experimentos se posee un puntaje F1 del 63 porciento. A su vez, un área bajo la curva (roc auc score) del 61 porciento, y una exactitud del 63 porciento.

11.3. Resultados de métricas para Random Forest

Tras su ejecución, el algoritmo arrojó resultados positivos con el conjunto de datos de prueba, obteniendo cifras por sobre el 65 porciento.



RESUMEN GENERICO , F1-SCORE Y MATRIZ DE CONFUSIÓN
AL PRIMER EXPERIMENTO REALIZADO CON K=5(FOLDS)

	precision	recall	f1-score	support
0	0.86	0.71	0.78	675
1	0.44	0.67	0.53	230
avg / total	0.76	0.70	0.72	905

Matriz de confusión: [[480 195]
[75 155]]

PROMEDIO DE METRICAS POST 40 EXPERIMENTOS

promedio F1 : 0.7179671853872283
promedio ROC_AUC : 0.6538610038610039
promedio exactitud : 0.7016574585635359

Figura 17: Resultado métricas obtenidas por medio de Word2Vec con Random Forest

De lo anterior se puede apreciar que bajo los experimentos realizados, se obtuvo que de los 230 elementos que conformaban la clase "Positivo", se tuvieron 155 aciertos y 75 fallos. Por otro lado para la segunda clase "Negativo"que poseía 675 elementos, se pudieron clasificar adecuadamente 480 de ellos, fallando en 195. Promediando todos los resultados de los experimentos se posee un puntaje F1 del 71 porciento. A su vez, un área bajo la curva (roc auc score) del 65 porciento, y una exactitud del 70 porciento.

12. Resumen de resultados

A modo de resumen, se tienen los siguientes valores para cada una de las representaciónes, en donde se puede apreciar el bajo rendimiento de la representación Word2Vec con respecto a TF-IDF, considerando el escenario y configuraciones descritas a lo largo de este informe:

Representación	TF-IDF	Word2Vec
Mejor algoritmo	**SVM**	**SVM**
Accuracy	90.15%	76.02%
ROC_AUC	88.91%	71.72%
F1 Score	90.07%	77.09%
Tweets de prueba	**904**	**905**
Negativos correctos	**525**	**503**
Negativos erróneos	**32**	**159**
Positivos correctos	**290**	**185**
Positivos erróneos	**57**	**85**

Figura 18: Tabla resumen de resultados obtenidos para representaciones probadas durante la investigación.

Se estima que el rendimiento de Word2Vec puede variar según el tipo de Word Embedding que se utilice, ya que para efectos de investigación se utilizó uno basado en los Tweets clasificados de manera inicial. Hoy en día existen Word Embeddings pre-creados que varían en su contexto; Desde Wikipedia, Google u otros sitios. Lo anterior, sumado a que se utilizó un tamaño de vector 'S' predeterminado de 300, resulta en que posiblemente, el rendimiento actual de Word2Vec para el proceso anteriormente expuesto, sea menor al de TF-IDF, no se descarta que este pueda, bajo un adecuado manejo de Word Embeddings de entrenamiento y configuraciones de algoritmo, este pueda superar dichas cifras.

13. Conclusión

A medida que la cantidad de usuarios en las redes sociales aumenta, y que la expresión de opinión está al alcance de un teléfono inteligente, es que han surgido nuevas estrategias para la captación de ésta información, a modo de poder obtener estadísticas o cifras, que permitan conocer respecto a la opinión general de la población de un continente, de un país o de una localidad, respecto a una temática en particular. Chile es uno de los países con mas usuarios en Twitter en los últimos años, y por esta razón es que se ha decidido aplicar esta técnica en él. El acontecer nacional diario, suele atraer a muchas personas que minuto a minuto publican sus opiniones referentes a diversas temáticas, muchas veces twitter es más rápido en difundir cierta información que los mismos otros medios de comunicación masiva; Televisión, Radio, etc.

A través de Analitic PRO, fue posible indagar sobre aspectos socio-demográficos relacionados al perfil de los usuarios que opinaban respecto a ciertos candidatos; Cuantos seguidores tenían (que tan influyente era su opinión para los demás), desde cuál sistema operativo publicaba sus Tweets (Android, IOS, Windows Phone). También, permite observar cuales son las palabras que más se repiten entre los usuarios que opinan respecto a un candidato.

Dentro de los valores rescatados en una primera etapa, los porecentajes obtenidos por el algoritmo SVM (máquina de soporte vectorial) superaron a los de arbol de decisión y Naive Bayes, contribuyendo para esto la mejora a la configuración realizada para el algoritmo. Posteriormente, para las pruebas con el segundo metodo de representación, SVM mantuvo su tendencia a mejor candidato para clasificación.

Debido a que se utilizaron Word Embeddings (texto de entrenamiento para el algoritmo) provenientes de la base de datos de Tweets, obtenida tras la extracción a través de Analitic PRO, es que se puede considerar, dado dicho escenario, a Word2Vec como una alternativa de rendimiento regular para la clasificación de Tweets en el contexto de análisis de sentimiento, debido a que sus cifras en métricas de rendimiento no superaron el 75 porciento, en comparación a las de TF-IDF que bordeaban el 90 porciento de exactitud. Si bien, la técnica ha sido utilizada de manera amplia y sus derivados como doc2vec probablemente obtengan resultados (en relación a lo obtenido con word2vec) cercanos las 77 porciento, Word2Vec resultó como un peor método de clasificación, tras el ambiente dispuesto para la investigación. En base a lo anterior, se estima que el rendimiento de Word2Vec podría mejorar, tras la utilización de otros tipos de Word Embeddings, y bajo otras configuraciones internas, como el largo de vector "s". Se podría considerar en ese caso a Word2Vec como un pontencial candidato para superar o igualar al rendimiento de la representación en base a TF-IDF.

Se comprueba en base a lo anterior, que para la técnica de análisis de sentimientos, la adecuada representación de los datos de entrada a los algoritmos, conforma un papel fundamental en los resultados de la posterior clasificación y obtención de métricas, pero también, su adecuada configuración y pre-entrenamiento constituyen un rol importante para su adecuado rendimiento. Los datos de entrada, también tornan un valor fundamental; El sesgo obtenido hacia la clase negativa por parte de los clasificadores, se debe a que la gente posee una tendencia a Twittear u opinar más sobre aspectos negativos de un candidato que positivos, por lo que naturalmente dicha clase poseerá mayor cantidad de asociaciones, lo que lleva a el error obtenido para ambas representaciones de datos.

Referencias

AnaliticPro (2018). Analitic : Social media listening. *https://www.analitic.cl/productos/analiticPro*.

Andranik Tumasjan, Timm O. Sprenger, P. G. S. I. M. W. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Technische Universität München Lehrstuhl für Betriebswirtschaftslehre Strategie und Organisation Leopoldstraße. Munich, Germany*.

Becerra, C. M. (2016). Análisis de sentimiento en twitter: El bueno, el malo y el >:(. Universidad Nacional de Córdoba, Argentina.

Bermingham, A. and Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results. *CLARITY: Centre for Sensor Web Technologies School of Computing Dublin City University*.

Castro, J. A. P. (2016). Análisis de sentimiento y clasificación de texto mediante adaboost concurrente. Pontificia Universidad Católica de Valparaíso.

Deltell, L. (2015). Estrategias de comunicaciÓn polÍtica en la redes sociales durante la campaña electoral del 2011 en españa: El caso de equo. *Universidad Computense de Madrid*.

Emol (2016). Jóvenes chilenos usuarios de redes sociales. *http://www.emol.com/noticias/Tecnologia/2016/09/26/823693/95-de-los-jovenes-chilenos-se-reconoce-como-usuario-de-las-redes-sociales.html*.

Enrique Alonso, Nerea Blanco, S. C. A. R. (2013). El debate público en las redes sociales. twitter españa como estudio de caso. *Universidad Autónoma de Madrid*.

et al, T. M. (2013). Efficient estimation of word representations in vector space.

G. Paltoglou, M. (2013). More than bag-of-words: Sentence-based document representation for sentiment analysis. *Faculty of Science and Technology University of Wolverhampton*.

García, L. M. (2014). Análisis de sentimientos y prediccion de eventos en twitter. Universidad de Chile.

Go A, Bhayani R, H. L. (2009). Twitter sentiment classification using distant supervision.

Hu M, L. B. (2004). Mining and summarizing customer reviews. *http://doi.acm.org/10.1145/1014052.1014073*, pp 168-177.

LaNación (2017). Encuestas en chile. *http://lanacion.cl/2017/05/22/adimark-y-cadem-las-encuestas-siguen-siendo-el-mejor-instrumento/*.

Marketing4ECommerce (2017). Usuarios de internet en el mundo. *https://marketing4ecommerce.net/usuarios-de-internet-mundo-2017/*.

Murphy, K. P. (2006). Naive bayes classifiers. *https://datajobsboard.com/wp-content/uploads/2017/01/Naive-Bayes-Kevin-Murphy.pdf*.

Sckit-learn (2017). Documentation of sckit-learn. *http://scikit-learn.org/stable/documentation.html*.

SocialBakers (2008). https://www.socialbakers.com/statistics/twitter/profiles/. *Global media analytics system used for marketing proposes*.