Escola Tècnica Superior d'Enginyeria

Electrònica i Informàtica La Salle

Treball Final de Màster

Màster Universitari en Ciència de les Dades

ANÁLISIS DE SENTIMIENTO PARA PREDECIR RESULTADOS DE LA PREMIER LEAGUE

Alumne

Javier Enrique Landaeta Lauria

Professor Ponent

Xavier Vilasís

ACTA DE L'EXAMEN

DEL TREBALL FI DE CARRERA

Reunit el Tribunal qualificador en el dia de la data, l'alumne

D. **Javier Enrique Landaeta Lauria**

va exposar el seu Treball de Fi de Carrera, el qual va tractar sobre el tema següent:

ANÁLISIS DE SENTIMIENTO PARA PREDECIR RESULTADOS DE LA PREMIER LEAGUE

Acabada l'exposició i contestades per part de l'alumne les objeccions formulades pels Srs. membres del tribunal, aquest valorà l'esmentat Treball amb la qualificació de

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENT DEL TRIBUNAL

UNIVERSIDAD RAMON LLULL – LA SALLE
ESTUDIOS DE MÁSTER
MÁSTER UNIVERSITARIO EN CIENCIA DE LOS DATOS

ANÁLISIS DE SENTIMIENTO PARA PREDECIR RESULTADOS DE LA PREMIER LEAGUE

Autor: LANDAETA LAURIA, JAVIER ENRIQUE
Asesor: VILASÍS, XAVIER
AÑO: 2018

RESUMEN

Debido al aumento de la cantidad de apuestas deportivas, predecir los resultados de fútbol ha sido un tema que ha ganado popularidad y se ha convertido en un campo de investigación. Hoy en día, con el aumento de la información disponible en redes sociales, predecir fenómenos sociales, políticos o deportivos con base en las opiniones de las personas que forman parte activa de estas redes sociales, se ha convertido en un área de investigación importante. El objetivo de este proyecto es construir modelos predictivos utilizando las opiniones de las personas en la red social Twitter, resultados de partidos anteriores y las posibilidades (odds) de las casas de apuestas para lograr predecir resultados de partidos de fútbol de la Premier League. Entre los conocimientos más importantes en este trabajo de investigación, se encuentran los conceptos relacionados con Análisis de Sentimiento, Regresiones Lineales, Regresiones Lineales Generalizadas, Naive Bayes, entre otros. La metodología que se utilizará es una investigación aplicada del tipo investigación y desarrollo, y se espera que su aplicación contribuya a la mejora del conocimiento de la industria de las apuestas y de otras industrias que pudieran aplicar los mismos principios para predecir fenómenos en otros campos.

Palabras Clave: Ciencia de Datos, Análisis de Sentimiento, Regresiones Lineales, Naive Bayes, Twitter.
Línea de Trabajo: Modelos Predictivos.

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

INTRODUCCIÓN

Actualmente, las redes sociales se encuentran en un proceso de crecimiento constante de número de usuarios y cantidad de información que es compartida entre sus usuarios. A primera vista, mucha de esa información no pareciera tener mucha relación y no pareciera ser información útil que se pudiese utilizar para algo productivo. Sin embargo, estudios recientes han arrojado que mucha de esa información puede ser utilizada para conocer la opinión de las personas en temas de interés público, social, económico, financiero, entre otros. Se ha encontrado que, en las redes sociales, los usuarios depositan constantemente su opinión en cuanto temas de interés.

Muchos estudios se han realizado en cuanto a conocer la opinión de las personas en ciertos temas utilizando las opiniones de los usuarios de la red social Twitter. Estudios para pronosticar el resultado en elecciones de presidente de un país, para conocer cómo fluctuarán las acciones de una importante empresa, para pronosticar los resultados de la liga de fútbol americano, para pronosticar los resultados de futbol o para conocer cuáles son las principales dolencias y enfermedades de un país en cada temporada, son algunos de los estudios que se han hecho tomando en consideración las opiniones de la red social Twitter.

Muchos de los estudios están basados en la predicción. Esto quiere decir que, utilizan las opiniones generadas en la red social Twitter para poder predecir algún evento. Para ello, se debe realizar un análisis que permita ponderar cada una de esas opiniones y procesarlas para conocer su tendencia. Este análisis se realiza comúnmente con un método de Machine Learning llamado Naive Bayes. Éste es el caso del presente trabajo final de máster.

Para el presente trabajo final de máster, se cogerán las opiniones de la red social Twitter con respecto a la liga de futbol de Inglaterra, se clasificarán por semana y por equipo, se depurarán obteniendo la información relevante para el análisis y luego se entrenará y probará un modelo de predicción. Además, se creará un modelo estadístico basado en los resultados de los partidos anteriores para poder

predecir resultados y se comparará con el modelo de predicción basado en opiniones en Twitter. Finalmente, se creará un método gráfico para medir la competitividad de la liga y se comparará con otras ligas de futbol.

Se espera que la red social Twitter aporte suficientes indicios para poder predecir resultados de futbol con cierta fiabilidad.

CAPITULO I: EL PROBLEMA

En este capítulo se detallan los aspectos más relevantes en relación a la problemática y los objetivos que se buscan alcanzar con este trabajo especial de grado, además de su justificación y su alcance.

1.1 Planteamiento del Problema

El incremento de los mercados de apuestas deportivas en línea ha visto crecer la industria de las apuestas en un 23% desde 2016 donde se registraron un total de apuestas por 448 millones de euros (AZARplus, 2017). Para el 2017, se registraron un total de apuestas por 549 millones de euros de los cuales 331 millones de euros fueron destinados a la industria del futbol. Esto significa que la industria del futbol posee alrededor del 60% del total de apuestas deportivas.

Por otro lado, el aumento de la popularidad de los sitios web de redes sociales ha experimentado un crecimiento exponencial de cantidad de usuarios, con usuarios activos de Facebook que aumentan cada año en un 14% (Facebook, Inc., 2018) y usuarios de Twitter que aumentan cada año en un 4% (Statista, 2018).

Como se ha observado en los antecedentes plasmados en este trabajo de investigación, Twitter se ha utilizado para predecir o explicar una variedad de otros eventos, como el resultado de las elecciones, el mercado de valores, y la propagación de enfermedades. Todas estas son pruebas sólidas de que Twitter puede ser una fuente de información significativa y útil que puede explotarse mediante el uso de métodos estadísticos.

Dado que los 330 millones de usuarios de Twitter crean colectivamente 500 millones de tweets cada día (Aslam, 2018), las redes sociales son claramente un recurso rico en datos que tiene un gran potencial. Una forma en que los datos de texto generados desde Twitter (un tweet) pueden utilizarse para desarrollar su potencial, y que este proyecto pretende indagar más a fondo, es en el campo del análisis de sentimientos. Este análisis pasa por determinar si un tweet puede clasificarse como positivo o negativo para que se pueda entender mejor el “sentimiento” sobre un tema determinado.

Por las razones y argumentos planteados anteriormente, se busca a través de esta investigación, diseñar un modelo de predictivo que permita obtener con fiabilidad los resultados de partidos de fútbol de la asociación de partidos de la Premier League (Wikipedia, s.f.). Esto se logrará a través del desarrollo de diferentes modelos. En primer lugar, mediante la realización de un análisis de opinión de tweets: determinar si un conjunto de tweets que mencionan a un equipo puede predecir si ese equipo ganará, perderá o empatará. En segundo lugar, utilizando los resultados de partidos anteriores, implementar un modelo basado en la Distribución de Poisson, y finalmente, utilizando las posibilidades de las casas de apuestas para cada resultado.

1.1.1 Formulación del Problema

En comparación con la información que nos proporciona las casas de apuestas, ¿Se puede obtener una mejor exactitud en la predicción de resultados de futbol con un modelo creado en base al análisis de sentimientos en la red social Twitter?

1.1.2 Sistemización del Problema

- ¿Qué información es necesaria para construir modelos basados en la red Social Twitter y basados en resultados anteriores de futbol?
- Tomando en consideración las opiniones emitidas en la red social Twitter, ¿se puede construir un modelo que permita predecir los resultados de un partido de futbol? ¿Cuál sería su exactitud? ¿Tendría más exactitud un modelo que contemple varias semanas de análisis en comparación con un modelo que solo contemple una semana de análisis? ¿existirían palabras frecuentes que fuesen únicas para cada clase?
- Tomando en consideración solo los resultados de partidos de futbol anteriores, ¿se podría crear un modelo estadístico basado en una distribución de probabilidades de Poisson para predecir los resultados de un partido de futbol? ¿Cuál sería su exactitud? ¿Cuál sería el umbral de confianza para el modelo?
- ¿Qué modelo posee una mejor exactitud? ¿La competitividad entre los equipos de la Premier League pudiera influir en la exactitud del modelo?

1.2 Objetivos

1.2.1 Objetivo General

Diseñar modelos que permitan predecir los resultados de futbol de la Premier League basados en resultados de partidos anteriores y opiniones en la red social Twitter.

1.2.2 Objetivos Específicos

- Descargar, clasificar, guardar y depurar información referente a Tweets, resultados de partidos anteriores y posibilidades de las casas de apuestas de equipos de futbol de la Premier League.
- Crear un modelo basado en las opiniones emitidas en la red social Twitter.
- Crear un modelo estadístico basados en los resultados de partidos anteriores que tome en consideración una regresión lineal generalizada.
- Comparar los resultados del modelo estadístico basado en resultados anteriores con los datos de casas apuestas y los resultados del modelo basado en opiniones en Twitter.
- Análisis de competitividad de la Premier League.

1.3 Justificación de la Investigación

Teniendo en consideración el tema y los objetivos de este proyecto, se espera que este proyecto y sus hallazgos sean de interés para la industria del juego en general. Los establecimientos de apuestas deportivas pueden estar interesados en adoptar o incorporar una solución que use el sentimiento de las redes sociales en algoritmos que determinen las probabilidades de apuestas. Mientras tanto, los clientes de estos establecimientos de apuestas pueden estar interesados en utilizar los hallazgos de este proyecto para diseñar una estrategia de apuestas rentable contra los corredores de apuestas.

Más genéricamente, las organizaciones de todas las industrias probablemente estén interesadas en los resultados de este proyecto para justificar una implementación de una solución que utilice el análisis de sentimiento de tweets que sea apropiada para sus necesidades. Si bien ya se han realizado investigaciones

para utilizar el sentimiento de tweet para predecir eventos, es posible que el análisis de sentimiento de tweets se pueda aplicar de manera más general en otros campos, como la prevención del delito; por ejemplo, detectar si hay una asociación entre áreas con altas tasas de criminalidad y tweets enviados desde el área, que contienen sentimientos negativos, y diseñar una estrategia apropiada para reducir estas altas tasas de criminalidad.

1.4 Alcance y limitaciones de la Investigación

El presente trabajo de investigación contempla la recopilación, la clasificación, el procesamiento de tweets para crear un modelo de predictivo que, junto con un modelo de distribución de Poisson, sirvan para predecir los resultados del partido de fútbol.

La obtención de los tweets relacionados con los equipos no se realizará con alguna API de Twitter. Las API de Twitter poseen ciertas limitaciones con relación a la antigüedad y la cantidad de información que se desea obtener. En consecuencia, se utilizará una herramienta que permite obtener información antigua en Twitter. Esta herramienta básicamente, funciona de la siguiente manera: cuando se ingresa en la página de Twitter, se inicia un cargador de desplazamiento, si se desplaza hacia abajo, se comienza a visualizar tweets y son descargados en un archivo .csv. Puede que la utilización de esta herramienta impacte la exactitud de la información que se obtiene de Twitter y puede que no contemple toda la información que realmente se generó durante un período tiempo. Sin embargo, luego de realizar algunas pruebas con la herramienta y al compararla con la información que se puede visualizar directamente en la página de Twitter, ésta tiene coherencia y no carece de información.

Las aplicaciones de estos modelos no están contempladas como parte de los objetivos de este trabajo de investigación.

CAPÍTULO II: MARCO TEORICO

En este capítulo se hace mención a otros trabajos de investigación relacionados con el análisis de sentimientos aplicados en diferentes campos.

Este capítulo busca dar a conocer todos los conocimientos previos que son necesarios para comprender el desarrollo de este trabajo de investigación.

2.1 Antecedentes

Para realizar cualquier tipo de investigación es de suma importancia conocer las bases teóricas y fundamentos que serán tratados. Adicionalmente, para todo trabajo final de máster se requiere conocer si el trabajo que se tiene pensado hacer ha sido realizado por otro investigador y que alcances se ha logrado. Es por esto que para realizar este trabajo final de máster se realizó una investigación de antecedentes y se obtuvieron los siguientes resultados:

(Tumasjan, A; Sprenger, T; Sandner, P; Welpe, I., 2010), **Prediciendo elecciones en Twitter: Cómo 140 caracteres revelan el sentimiento de la política**, tuvo como objetivo principal utilizar el contexto de las elecciones federales de Alemania para investigar si la red social Twitter es utilizada como un foro para la deliberación de opiniones de política y si estas opiniones tenían correlación con los resultados electorales, para lo cual, primero se realizó una investigación del ambiente político del país y cuáles son las fuerzas políticas que lo conforman. Luego se extrajeron los tweets relacionados de la red social Twitter y se ejecutó un software de análisis de texto. Para el logro del objetivo principal de este trabajo, se utilizó un software para el análisis de texto llamado LIWC (Linguistic Inquiry and Word Count) con el cual se obtuvieron 100 mil mensajes relacionados con algún partido político o con un político directamente. Como resultado de la investigación, se logró demostrar que la red social Twitter sí es utilizado como palestra para la deliberación de temas políticos. Además, se encontró que el análisis de gran número de mensajes, puede reflejar los resultados de una elección. Este antecedente servirá como referencia para el presente trabajo de investigación, como una de las primeras investigaciones donde se utiliza Twitter para predecir algún evento.

Palabras Clave: Twitter, Elecciones, Análisis de Sentimiento, Predicción.

(Paul, M; Dredze, M., 2011), **Tu eres lo que colocas en Twitter: Analizando datos de Twitter para la salud pública**, tuvo como objetivo principal analizar los mensajes colocados en la red social Twitter para medir características de la población. Particularmente, se consideró una gama amplia de aplicaciones de salud pública para Twitter. Se aplicó el modelo recientemente introducido de Aspectos del Tema de Enfermedades (ATAM) a más de un 1 millón de tweets relacionados con la salud. Este modelo es un nuevo modelo para la red social Twitter que asocia síntomas, tratamientos y palabras generales con enfermedades. ATAM es capaz de identificar enfermedades como la gripe, las infecciones y la obesidad. Sus resultados coinciden con los resultados producidos por Google Flu Trens y con los datos gubernamentales de salud pública. Como parte del proyecto, se tomó la información de los tweets y se eliminaron los signos de puntuación, las palabras vacías, las URLs y los Hashtags. Además, se utilizó el algoritmo SVM de clasificación para identificar mensajes relacionados con salud y los más de 1 millón de Tweets. Como resultado de la investigación, se logró identificar menciones de más de una docena de dolencias, que incluyen alergias, obesidad e insomnio. Este antecedente servirá como referencia para el presente trabajo de investigación, como una de las primeras investigaciones donde se aplicaron métodos de filtrado para elementos que no aportan información al modelo y la aplicación de métodos de Machine Learning para crear un clasificador de palabras.

Palabras Clave: Twitter, Salud Pública, Modelo, Enfermedades, Predicción.

(Sinha, S; Dyer, C; Gimpel, K; Smith, N, 2013), **Prediciendo la NFL utilizando Twitter**, tuvo como objetivo principal estudiar la relación entre los datos de las redes sociales y los partidos de la NFL (National Football League) utilizando un conjunto de datos de mensajes provenientes de Twitter para lo cual, primero se consideraron Tweets pertenecientes a equipos y juegos específicos en temporadas de la NFL y luego se utilizaron, junto con datos estadísticos de juego, para construir modelos predictivos para resultados futuros y apuestas deportivas. Para construir el modelo, se extrajo un promedio de 42 mil tweets por semana durante dos temporadas de la

NFL (2010-2012). Luego se filtró la información para encontrar los Tweets que fuesen relevantes para crear el modelo, y se creó un modelo de predicción que considera un clasificador de regresión logística. Finalmente, se experimentó con varios conjuntos de Features y, como resultado de la investigación, se descubrió que grandes volúmenes de datos de Twitter pueden igualar o superar el rendimiento de las funciones estadísticas más tradicionales. Este antecedente servirá como referencia para el presente trabajo de investigación, como el primer trabajo donde se intentan predecir resultados relacionados con el deporte a través de opiniones en Twitter.

Palabras Clave: Twitter, NFL, Modelo Predictivo, Regresión Logística.

(Venkata, S; Kamal, N; Ganapati, P;, 2016), **Análisis de sentimiento de datos de Twitter para predecir el movimiento del mercado de valores,** tuvo como objetivo principal diseñar un modelo que fuese capaz de predecir los precios en las acciones de una empresa, sus aumentos y caídas, a partir de cómo está correlacionada esta información con las opiniones públicas que se expresan en la red social Twitter para lo cual, primero se realizó una investigación para comprender la opinión del autor a partir de un texto y luego se evaluaron herramientas tecnológicas de Machine Learning que permitieran realizar el análisis de sentimiento. Luego se procedió a la extracción de los datos mediante una API de Twitter, logrando la extracción de 2.5 millones de Tweets en un período de un año (entre agosto 2015 y agosto 2016). Luego se emplearon dos diferentes tipos de representaciones textuales, Word2Vec y Ngram para la extracción de los features Twitter y para el análisis de sentimiento. Como resultado de la investigación, se logró construir un modelo que permite predecir las variaciones del precio de las acciones de una empresa con base en las opiniones en la red social Twitter con una exactitud de casi 70%, lo cual, pudiera alentar a las personas a invertir en acciones en una empresa. Este antecedente servirá como referencia para el presente trabajo de investigación, como una de las primeras investigaciones donde se utiliza Word2Vec. Word2vec es un grupo de modelos relacionados que se utilizan para crear incrustaciones de palabras. Básicamente, Word2vec toma como entrada un gran corpus de texto y produce un

espacio vectorial, típicamente de varios cientos de dimensiones, asignándose a cada palabra única en el corpus un vector correspondiente en el espacio. Los vectores de palabras se ubican en el espacio vectorial de forma tal que las palabras que comparten contextos comunes en el corpus se ubican muy cerca la una de la otra en el espacio.

Palabras Clave: Twitter, Análisis de Sentimiento, Predicción del mercado de valores, Machine Learning.

(Kampakis, 2014), **Utilizando Twitter para predecir resultados de Futbol**, tuvo como propósito estudiar si los datos extraídos de Twitter se pueden usar para este propósito. Desarrollamos un conjunto de modelos predictivos para el resultado de los juegos de fútbol de la Premier League inglesa durante un período de 3 meses basado en tweets y estudiamos si estos modelos pueden superar los modelos predictivos que usan solo datos históricos y estadísticas simples de fútbol. Además, los modelos combinados se construyen utilizando Twitter y datos históricos. Los resultados finales indican que los datos extraídos de Twitter pueden ser una fuente útil para predecir juegos en la Premier League. El modelo final basado en Twitter tiene un desempeño singularmente mejor que el azar. Por lo tanto, este estudio proporciona evidencia de que las características derivadas de Twitter pueden proporcionar información útil para la predicción de los resultados del fútbol (fútbol). Este antecedente servirá como referencia para el presente trabajo de investigación, como una de las primeras investigaciones donde se utiliza Twitter para predecir resultados de Futbol y el primero donde se utiliza Naive Bayes para crear un modelo de Machine Learning.

2.2 Fundamentos Teóricos

A continuación, se presentan las bases teóricas necesarias para soportar desde el punto de vista conceptual el desarrollo de esta investigación. Cada de uno de estos conceptos serán implementados de diversas formas durante el curso de la investigación.

Los siguientes subcapítulos ayudarán al lector al conocer temas relacionados con la teoría de Poisson, Naive Bayes, análisis de sentimiento y regresiones lineales.

2.2.1 Análisis de Sentimiento

A juicio de (Federico Pozzi, Elisabetta Fersini, Enza Messina, Bing Liu, 2017), en su libro de Análisis de Sentimiento en Redes sociales.

> "El objetivo del análisis de sentimiento es definir herramientas automáticas que permitan extraer información subjetiva desde textos creados en lenguaje natural, tal como opiniones y sentimientos, de manera de crear conocimiento accionable y estructurados que puede ser usado en la toma de decisiones o en sistemas de soporte a las decisiones". (p.1)

Además, el mismo autor expresa los siguiente.

> "Hoy en día, el análisis de sentimiento ha ganado mucho más valor con la llegada de las redes sociales. Su excelente difusión y su rol en las sociedades modernas representa una de las más interesantes novedades en los años recientes, capturando el interés en desarrolladores, periodistas, compañías y gobiernos. La densa interconexión que usualmente surge entre los usuarios activos genera un espacio de discusión que permite involucrar y motivar individuos de una extensa ágora, enlazando personas con objetivos comunes y facilitando diversas formas de acción colectiva. Por lo tanto, las redes sociales están creando una revolución digital, permitiendo la expresión y la difusión de emociones y opiniones a través de la red, abriendo una ventana al mundo de otros y a husmear dentro de sus vidas. Los datos de las opiniones en la red, si son recolectados y analizados correctamente, permiten no solo entender y explicar muchos fenómenos sociales complejos, sino que también permiten predecirlos". (p.1 – p.2)

Adicionalmente, el mismo autor expresa que hoy en día, el progreso de las tecnologías facilita el almacenamiento eficiente y la consulta de una gran cantidad de datos, y el gran foco actual se encuentra en los métodos de extracción de información y creación de conocimiento desde fuentes crudas. También establece que las redes sociales representan un sector emergente de reto en el contexto del "big data" donde las expresiones del lenguaje natural de las personas pueden ser fácilmente reportados y analizados para crear conocimiento que permita accionar procesos de toma de decisiones.

Por último, (Federico Pozzi, Elisabetta Fersini, Enza Messina, Bing Liu, 2017) expresan que el principal problema cuando se está realizando un análisis de sentimiento, consiste en distinguir entre las oraciones objetivos y las subjetivas. Si una oración es clasificada con objetiva, ninguna otra tarea se debe realizar, mientras que si una oración se clasifica como subjetiva, se debe estimar su polaridad, es decir, si la oración posee un sentido positivo, negativo o neutral. A continuación, se muestra un flujo de trabajo que explica este proceso.

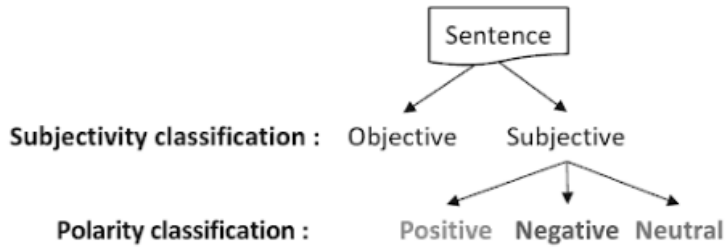


Figura 1. Flujo de trabajo del Análisis de Sentimiento

Tomando otros conceptos, según (Liu, 2012) en su libro Análisis de Sentimiento y Extracción de Opiniones.

> "El análisis de sentimiento, también llamado minería de opiniones, es el campo de estudio que analiza las opiniones, los sentimientos, las evaluaciones, las evaluaciones, las actitudes y las emociones de las personas con relación a entidades como productos, servicios, organizaciones, individuos, problemas, eventos, tópicos y sus atributos." (p.1)

Así mismo, el mismo autor establece las opiniones se encuentran en el centro de casi todas las actividades de los humanos, ya que son influenciadores clave de nuestro comportamiento. Siempre que tenemos que tomar una decisión, deseamos conocer otras opiniones. En el mundo real, las empresas desean siempre conocer las opiniones de sus productos o servicios. Así mismo, los consumidores siempre desean conocer las opiniones de algún usuario sobre un producto antes de adquirirlo, así como también, las opiniones de algún candidato político antes de tomar una decisión de voto en una elección política. El mismo autor también sostiene que con la explosión de los medios sociales (ej.: foros de discusión, blogs, microblogs, comentarios y sitios de redes sociales) en la web, individuos y

organizaciones se encuentran utilizando el contenido en estos medios para tomar decisiones.

Basado en los conceptos los autores anteriormente consultados, se puede decir que existen algunos aspectos clave en los que estos autores coinciden. Esos aspectos son los siguientes:

- El análisis de sentimiento busca crear conocimiento sobre las opiniones desde texto generados en lenguaje natural.
- El conocimiento generado por un análisis de sentimiento es utilizado para la toma de decisiones.
- Hoy en día los medios sociales, especialmente las redes sociales, son un nicho de información importante que, con los métodos de extracción y análisis adecuados, pueden ser utilizados para generar conocimiento.

2.2.2 Clasificación de Datos

De manera de analizar correctamente el sentimiento de la información contenida en los tweets, con el objetivo de detectar cualquier correlación ente el contenido de los tweets y los resultados de un partido de fútbol, se toma en consideración el proceso para la clasificación de datos.

La clasificación de los datos es un problema que radica en poder identificar la clase en la que se deben clasificar los datos y el aprendizaje de la relación entre los datos y las variables de la clase “dado un conjunto de datos de entrenamiento con una clasificación dada para determinar de las instancias que no están clasificadas en el conjunto de datos de prueba” (Aggarwal, 2015).

A juicio de (Aggarwal, 2015), en su libro de Clasificación de Datos: Algoritmos y Aplicaciones.

> “El problema de clasificación segmenta las instancias de prueba en grupos, tal como se define en la etiqueta de la clase. Si bien la segmentación de ejemplos en grupos también se realiza por clústeres, existe una diferencia clave entre los dos problemas. En el caso del “clustering”, la segmentación se realiza utilizando similitudes entre las variables características, sin una comprensión

> previa de la estructura de los grupos. En el caso de la clasificación, la segmentación se realiza sobre la base de un conjunto de datos de capacitación, que codifica el conocimiento sobre la estructura de los grupos en forma de una variable objetivo. Por lo tanto, aunque las segmentaciones de los datos generalmente están relacionadas con nociones de similitud, como en el “clustering”, se pueden lograr desviaciones significativas de la segmentación basada en la similitud en entornos prácticos. Como resultado, el problema de clasificación se denomina aprendizaje supervisado, al igual que el “clustering” se denomina aprendizaje no supervisado. El proceso de supervisión a menudo proporciona una utilidad específica para las aplicaciones porque las etiquetas de clase puedan representar propiedades importantes de interés”. (p.2)

De acuerdo con (Aggarwal, 2015), existen diferentes métodos para la clasificación de datos que funcionan mejor dependiendo de la aplicación, entre los cuales se destacan:

Método de selección de características

La primera fase de prácticamente todos los algoritmos de clasificación es la de selección de características. En la mayoría de los escenarios de minería de datos, una gran variedad de características es recopilada por individuos que a menudo no son expertos en el dominio. Claramente, las características irrelevantes a menudo pueden dar como resultado un modelo pobre, ya que no están bien relacionadas con la etiqueta de la clase. De hecho, estas características empeorarán la precisión de la clasificación debido al sobreajuste, cuando el conjunto de datos de entrenamiento es pequeño y se permite que dichas características sean parte del modelo de entrenamiento. Por lo tanto, es fundamental usar las funciones correctas durante el proceso de entrenamiento.

Métodos probabilísticos (Naive Bayes)

Los métodos probabilísticos son los más fundamentales entre todos los métodos de clasificación de datos. Los algoritmos de clasificación probabilística usan inferencia estadística para encontrar la mejor clase para un ejemplo dado. Además de asignar simplemente la mejor clase como otros algoritmos de clasificación, los algoritmos de clasificación probabilística generarán una probabilidad posterior correspondiente

de que la instancia de prueba sea miembro de cada una de las posibles clases. La probabilidad posterior se define como la probabilidad después de observar las características específicas de la instancia de prueba. Por otro lado, la probabilidad previa es simplemente la fracción de los registros de entrenamiento que pertenecen a cada clase en particular, sin conocimiento de la instancia de prueba. Después de obtener las probabilidades posteriores, utilizamos la teoría de la decisión para determinar la membresía de la clase para cada nueva instancia.

Árboles de decisión

Los árboles de decisión crean una partición jerárquica de los datos, que relaciona las diferentes particiones a nivel de “hoja” con las diferentes clases. La partición jerárquica en cada nivel se crea con el uso de un criterio de división. El criterio de división puede usar una condición en un solo atributo, o puede contener una condición en múltiples atributos. El primero se conoce como división univariada, mientras que el segundo se conoce como división múltiple. El enfoque general es intentar dividir recursivamente los datos de entrenamiento para maximizar la discriminación entre las diferentes clases sobre diferentes nodos.

Métodos basados en reglas

Los métodos basados en reglas están estrechamente relacionados con los árboles de decisión, excepto que no crean una partición jerárquica estricta de los datos de entrenamiento. Por el contrario, se permiten superposiciones para crear una mayor solidez para el modelo de entrenamiento. Cualquier ruta en un árbol de decisión se puede interpretar como una regla, que asigna una instancia de prueba a una etiqueta particular.

Aprendizaje basado en instancias

En el aprendizaje basado en instancias, a menudo se prescinde de la primera fase de la construcción del modelo de capacitación. La instancia de prueba está directamente relacionada con las instancias de entrenamiento para crear un modelo de clasificación. Dichos métodos se denominan métodos de aprendizaje “lazy”, ya

que esperan el conocimiento de la instancia de prueba para crear un modelo optimizado localmente, que es específico de la instancia de prueba.

La ventaja de estos métodos es que se pueden adaptar directamente a la instancia de prueba particular, y pueden evitar la pérdida de información asociada con el carácter incompleto de cualquier modelo de entrenamiento.

Clasificadores SVM

Los métodos SVM usan condiciones lineales para separar las clases entre sí. La idea es usar una condición lineal que separe las dos clases lo mejor posible. En tal caso, la condición de división en el caso multivariante también se puede usar como condición independiente para la clasificación Esto, un clasificador SVM, se puede considerar como un árbol de decisión de un solo nivel con una condición de división multivariada muy cuidadosamente elegida.

Redes Neuronales

Las redes neuronales intentan simular sistemas biológicos, los cuales corresponden al cerebro humano. En el cerebro humano, las neuronas están conectadas entre sí a través de puntos, que se conocen como sinapsis. En los sistemas biológicos, el aprendizaje se realiza cambiando la fuerza de las conexiones sinápticas, en respuesta a los impulsos. Esta analogía biológica se conserva en una red neuronal artificial. La unidad de cálculo básica en una red neuronal artificial es una neurona o unidad. Estas unidades se pueden organizar en diferentes tipos de arquitecturas mediante conexiones entre ellas. La arquitectura más básica de la red neuronal es un perceptron, que contiene un conjunto de nodos de entrada y un nodo de salida. La unidad de salida recibe un conjunto de entradas de las unidades de entrada.

Se supone que los datos son numéricos. Los datos categóricos pueden necesitar transformarse en representaciones binarias y, por lo tanto, el número de entradas puede ser mayor. El nodo de salida está asociado con un conjunto de ponderaciones W, que se utilizan para calcular una función f() de sus entradas.

Cada componente del vector de ponderación está asociado a una conexión desde la unidad de entrada a la unidad de salida.

Basado en los conceptos de los autores anteriormente consultados, se puede decir que existen algunos aspectos clave. Esos aspectos son los siguientes:

- Para realizar una clasificación de datos deben existir un conjunto de datos de entrenamiento y un conjunto de datos de prueba para entrenar el modelo y para ponerlo a prueba respectivamente.
- Dado cualquier método de clasificación, siempre debe realizarse un primer paso de selección de características.
- Existen diversos métodos de clasificación de datos y se debe elegir el que más se adapte al problema.

2.2.3 Distribución de Poisson

De acuerdo con (Mónica Martínez; Manuel Marí), la distribución de Poisson es una distribución binomial y es una de las distribuciones de probabilidad más importantes donde cada variable aleatoria representa el número total de ocurrencias de un fenómeno durante un período de tiempo fijo. Además, expresa la probabilidad de un número k de ocurrencias acaecidas en un tiempo fijo, si estos eventos ocurren con una frecuencia media conocida y son independientes del tiempo discurrido desde la última ocurrencia o suceso.

Adicionalmente, el mismo autor expresa que las características de la distribución de Poisson son las siguientes:

- Sea una población de tamaño ∞.
- Sea una muestra de tamaño n bastante elevado (se suele hablar de que tiende a ∞)
- Los sucesos son independientes entre sí.
- Sea A un suceso que tiene una probabilidad p de suceder durante un periodo de tiempo, siendo esta probabilidad de ocurrencia durante un periodo de tiempo concreto muy pequeña (se suele hablar de que tiende a 0).

- El producto n*p, tiende a aproximarse a un valor promedio o número medio, al que llamaremos λ. Por ejemplo, promedio de llamadas recibidas en una central telefónica por minuto o número medio de accidentes producidos en una carretera durante el fin de semana.
- X: número de individuos de la muestra que cumplen A.
- El conjunto de posibles valores de A es, E = {0,1,2,3,4....}

Entonces, la función de probabilidad viene definida de la siguiente manera.

$$F(X = x) = \frac{e^{-\lambda} * \lambda^{x}}{x!}$$

2.2.4 Regresiones Lineales

De acuerdo con (Academy Khan, 2010), la regresiones lineales parten de una cantidad “n” puntos donde se desea encontrar una línea que minimice el error cuadrático entre esta línea y los “n” puntos. A continuación, se muestra una figura donde se puede observar una representación de los “n” puntos y la recta que se desea encontrar.

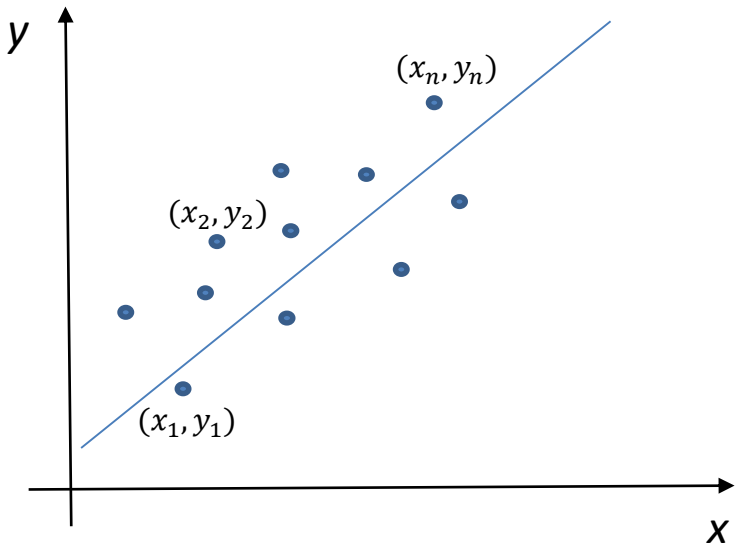


Figura 2. Conjunto de puntos que se desean aproximar por una recta

Dada ecuación de la recta que se desea encontrar tiene la siguiente estructura:

$$y = mx + b$$

$$donde, m = pendiente\ de\ la\ recta$$

$$b = al\ punto\ de\ corte\ en\ el\ eje\ de\ las\ y$$

Se desea encontrar los valores de “m” y de “b” que minimizan el error cuadrático. El error cuadrático (SE) se puede escribir de la siguiente manera.

$$SE_{recta} = (y_1 - (m * x_1 \ + b))^2 + \ (y_2 - (m * x_2 \ + b))^2 + \cdots + \ (y_n - (m * x_n \ + b))^2$$

A continuación, se rescribe la función anterior de la siguiente forma.

$$SE_{recta} = \ {y_1}^2 - 2 * y_1 * (m * x_1 \ + b) + \ (m * x_1 \ + b)^2$$

$$+ {y_2}^2 - \ 2 * y_2 * (m * x_2 \ + b) + \ (m * x_2 \ + b)^2$$

$$+ {y_n}^2 - 2 * y_n * (m * x_n \ + b) + \ (m * x_n \ + b)^2$$

A continuación, nuevamente se rescribe la función anterior de la siguiente forma.

$$SE_{recta} = \ {y_1}^2 - 2 * y_1 * \ m * x_1 - 2 * y_1 * b + m^2 * {x_1}^2 + 2 * m * x_1 * b + \ b^2$$

$$+ {y_2}^2 - 2 * y_2 * \ m * x_2 - 2 * y_2 * b + m^2 * {x_2}^2 + 2 * m * x_2 * b + \ b^2$$

$${y_n}^2 - 2 * y_n * \ m * x_n - 2 * y_n * b + m^2 * {x_n}^2 + 2 * m * x_n * b + \ b^2$$

A continuación, nuevamente se rescribe la función anterior de la siguiente forma.

$$SE_{recta} = \ ({y_1}^2 + {y_2}^2 + \cdots + {y_n}^2)$$

$$+2 * m * (y_1 * x_1 + \ y_2 * x_2 + \cdots + \ y_n * x_n)$$

$$-2 * b * (y_1 + \ y_2 + \cdots + \ y_n)$$

$$+ m^2 * ({x_1}^2 + \ {x_2}^2 + \cdots + \ {x_n}^2)$$

$$+ 2 * m * b\ (x_1 + \ x_2 + \cdots + \ x_n) + n * b^2$$

Si se toma en consideración las siguientes ecuaciones.

$$\frac{({y_1}^2 + {y_2}^2 + \cdots + {y_n}^2)}{n} = \ \overline{y^2} \rightarrow \ ({y_1}^2 + {y_2}^2 + \cdots + {y_n}^2) = \ \ n * \overline{y^2}$$

$$\frac{(y_1 * x_1 + \ y_2 * x_2 + \cdots + \ y_n * x_n)}{n} = \overline{x * y}$$

$$\rightarrow (y_1 * x_1 + y_2 * x_2 + \cdots + y_n * x_n) = n * \overline{x * y}$$

Se puede rescribir la ecuación del error cuadrático de la siguiente manera

$$SE_{recta} = n * \overline{y^2} - 2 * m * n * \overline{x * y} - 2 * b * n * \bar{y}$$

$$+m^2 * n * \overline{x^2} + 2 * m * b * n * \bar{x} + n * b^2$$

Para minimizar esta función, se debe calcular $\frac{\partial SE}{\partial m}$ y $\frac{\partial SE}{\partial b}$ e igualar ambas a cero.

$$\frac{\partial SE}{\partial m} = \ -2 * n * \overline{x * y} + 2 * n * \overline{x^2} \ * m + 2 * b * n * \bar{x} = 0$$

$$\frac{\partial SE}{\partial b} = \ -2 * n * \bar{y} + 2 * m * n * \bar{x} + 2 * b * n = 0$$

Como se puede ver en la figura anterior, ambas ecuaciones son divisibles por un factor de $2 * n$. Por lo tanto, las ecuaciones se pueden rescribir de la siguiente manera.

$$-\overline{x * y} + \overline{x^2} \ * m + b * \bar{x} = 0$$

$$-\bar{y} + m * \bar{x} + b = 0$$

Con este sistema de ecuaciones en función de “b” y “m”, se pueden obtener las siguientes ecuaciones.

$$m = \frac{\bar{x} * \bar{y} - \overline{x * y}}{\bar{x}^2 - \overline{x^2}}$$

$$b = \bar{y} - m * \bar{x}$$

Si se toma en consideración más de dos dimensiones, se puede decir que la recta $y = mx + b$, puede ser escrita de la siguiente forma.

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \epsilon$$

Donde, β_n son los coeficientes, x_n son variables independientes y ϵ es el error que puede existir entre el valor real y el que se intenta predecir.

Si Y está distribuida de acuerdo a una distribución $N(\mu, \sigma^2)$ y además que la función explicativa es $\eta = \mu = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$, la función de conexión entre es $\mu = E[Y] = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$.

2.2.5 Regresiones Lineales Generalizadas

A diferencia de las regresiones lineales donde se calcula directamente la relación entre la variable Y y la variable X, las regresiones lineales generales buscan calcular la relación entre $f(y)$ y la variable.

Esta función $f(y)$ será comúnmente una función logarítmica basada en el tipo de distribución que posea Y. Estas distribuciones pertenecen a la familia de distribuciones exponenciales, entre las que se destacan las siguiente.

- Binomial
- Multinomial
- Poisson

Los componentes de una regresión lineal generalizada serán los siguientes.

- **Distribución:** Familia de distribuciones exponenciales
- **Función explicativa:** $\eta_i = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$.
- **Función de conexión:** $\eta_i = ln(\mu_i) \rightarrow \mu_i = e^{\eta_i}$

Regresiones lineales: Bernoulli

Para el caso donde los valores de $Y_i \sim Bernulli(\mu_i)$, las componentes de una regresión lineal generalizada serán las siguientes.

- **Distribución:** $Y_i \sim Bernulli(\mu_i)$
- **Función explicativa:** $\eta_i = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$.
- **Función de conexión:** $f(\mu) = \eta = log(\frac{\mu}{1-\mu})$

Regresiones lineales: Poisson

Para el caso donde los valores de $Y_i \sim Poisson(\mu_i)$, las componentes de una regresión lineal generalizada serán las siguientes.

- **Distribución:** $Y_i \sim Poisson(\mu_i)$
- **Función explicativa:** $\eta_i = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$.
- **Función de conexión:** $f(\mu) = \eta = log(\mu)$

Estimación de parámetros basados en la máxima similitud

Si Y_i son observaciones independientes con los valores correspondientes x_i de las variables predictoras, entonces θ se puede estimar por máxima verosimilitud. Las estimaciones de máxima verosimilitud carecen de una expresión de forma cerrada y deben encontrarse por métodos numéricos.

Dado un conjunto de parámetros θ y un vector de entrada x, la media de la distribución de Poisson prevista, como se indicó anteriormente, viene dada por la siguiente expresión.

$$\lambda := \mathrm{E}(Y \mid x) = e^{\theta' x}$$

De esta manera, la función de probabilidad de la distribución de Poisson está dada por la siguiente expresión.

$$p(y \mid x; \theta) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{e^{y\theta' x} e^{-e^{\theta' x}}}{y!}$$

Si se considera un conjunto de datos de m vectores donde $x_i \epsilon \mathbb{R}^{n+1}, i = 1, \dots, m$ junto con un conjunto de m valores $y_{1,\dots,} y_m \epsilon \mathbb{R}$. Luego, para un conjunto de parámetros θ, la probabilidad de alcanzar este conjunto particular de datos viene dada por la siguiente expresión.

$$p(y_1, \dots, y_m \mid x_1, \dots, x_m; \theta) = \prod_{i=1}^{m} \frac{e^{y_i \theta' x_i} e^{-e^{\theta' x_i}}}{y_i!}.$$

Por el método de máxima verosimilitud, se desea encontrar el conjunto de parámetros θ que hace que esta probabilidad sea lo más grande posible. Para hacer esto, la ecuación se puede rescribir primero como una función de verosimilitud en términos de θ:

$$L(\theta \mid X, Y) = \prod_{i=1}^{m} \frac{e^{y_i \theta' x_i} e^{-e^{\theta' x_i}}}{y_i!}$$

Una fórmula en esta forma es típicamente difícil de trabajar; en cambio, se usa la log-verosimilitud.

$$\ell(\theta \mid X, Y) = \log L(\theta \mid X, Y) = \sum_{i=1}^{m} \left(y_i \theta' x_i - e^{\theta' x_i} - \log(y_i!) \right)$$

Los parámetros θ solo aparecen en los primeros dos términos de cada término en la suma. Por lo tanto, dado que solo interesa encontrar el mejor valor para θ, se puede descartar parte de la ecuación.

$$\ell(\theta \mid X, Y) = \sum_{i=1}^{m} \left(y_i \theta' x_i - e^{\theta' x_i} \right)$$

Finalmente, se debe calcular la derivada de la ecuación e igualarla a cero para encontrar los parámetros θ.

2.2.6 Naive Bayes

Naive Bayes es un algoritmo de aprendizaje que tiene la capacidad de resolver problemas de clasificación. Comúnmente, este algoritmo es utilizado para resolver problemas de clasificación de texto. Para explicar el algoritmo, se desarrollarán estas ideas bajo el contexto de un algoritmo de clasificación de texto.

Uno de las principales aplicaciones de NaiveBayes como clasificador de texto, es identificar si un correo electrónico es Spam o no. Otras de las aplicaciones del algoritmo es clasificar los correos electrónicos en cada una de las carpetas del correo según el contenido del correo electrónico, clasificar productos de acuerdo a su descripción o realizar clasificación de sentimientos (i.e. comentarios positivos o negativos). De manera de explicar el algoritmo, se utilizará como ejemplo la aplicación de detección de correos spam.

Si llega un correo electrónico a la bandeja de entrada de un correo electrónico que dice “Viagra, cómprala ya con el mayor descuento”, ¿esto es un correo Spam o no? NaiveBayes intenta resolver este problema al clasificar el correo sin tener que analizar el significado concreto del correo.

Unos de los primeros pasos para poder identificar si un correo electrónico es spam o no, es tomar un conjunto de datos de entrenamiento y calcular las palabras que han tenido mayor frecuencia de aparición (a esto se le denomina diccionario). Luego, se crea un vector de características o Feature Vector que describe cuáles de las palabras aparecen en el correo electrónico que se quiere analizar. Bajo este concepto, se pretende calcular P(C|x) donde:

$$x_i = \begin{cases} 1 \; si \; la \; palabra \; i \; aparece \; en \; el \; texto \\ 0 \; en \; los \; demás \; casos \end{cases}$$

De acuerdo con (Jurafsky, Manning, 2012),gracias a la regla de Bayes, se puede escribir P(C|x) como:

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)}$$

La mejor clase será, entre todas las clases, la clase que maximice la probabilidad de esa clase dada una lista de Features (map – Maximun a posteriori).

$$C_{map} = argmax_{c \varepsilon C} P(c|x)$$

Gracias a la regla de Bayes:

$$C_{map} = argmax_{c \varepsilon C} \frac{P(x|c) * P(c)}{P(x)}$$

Ahora, cualquier clase que maximice la ecuación anterior, también maximizará la siguiente ecuación:

$$C_{map} = argmax_{c \varepsilon C} P(x|c) * P(c)$$

$$Donde: P(x|c)\ es\ la\ verosimilitud\ y\ P(c)\ es\ "a\ priori"$$

La razón de porqué se puede eliminar el denominador, es que el denominador será igual para todas las clases. Es decir, la probabilidad del documento (la lista de Features) será la misma para todo el documento.

Ahora bien, si se tiene en consideración que X es un vector de observaciones, la formula anterior se puede representar de la siguiente manera.

$$C_{map} = argmax_{c \varepsilon C} P(x_1, x_2, \ldots . x_n|c) * P(c)$$

Se analiza esta fórmula, se puede observar que la probabilidad de una clase (P(c)), es básicamente calcular que tan frecuente es una clase de que ocurra. Esta variable se puede calcular contando las frecuencias relativas en nuestro dataset. Ahora, ¿cómo calcular la verosimilitud? Para responder esta pregunta, se deben realizar las siguientes asunciones:

- Las posiciones en la lista de Features (Bag of Words) no es importante. Lo que importará es calcular cuál palabra se encuentra en el texto a analizar.
- Independencia: Se asume que las probabilidades de los Features son independientes para una clase.

Es importante resaltar que estas asunciones están erradas y no son ciertas, pero serán de utilidad para simplificar el problema y como resultado de estas asunciones, se podrá representar la verosimilitud como:

$$P(x_1, x_2, \ldots . x_n|c) = P(x_1|c) * P(x_2|c) * \ldots * P(x_n|c)$$

Teniendo en consideración esto, se puede definir NaiveBayes como:

$$C_{NB} = argmax_{c_j \varepsilon C} P(c_j) \prod_{i \varepsilon posiciones} P(x_i|c_j)$$

CAPITULO III: MARCO METODOLOGICO

En el presente capítulo de esta investigación se expone la metodología que se utilizará para realizar la investigación. Esta metodología busca definir el tipo de la investigación, las fases de la investigación, entre otros aspectos importantes.

3.1 Tipo de Investigación

Para comenzar con la metodología que dará curso a la investigación, primero se debe definir el concepto de Investigación. A juicio de Hernández, Fernandez y Baptista (2010), “La investigación es un conjunto de procesos sistemáticos, críticos y empíricos que se aplican al estudio de un fenómeno” (p.4).

De acuerdo con Valarino, Yáber y Cemborain (2010), la investigación cumple con dos propósitos fundamentales: Resolver los problemas prácticos y producir conocimientos y teorías.

De acuerdo con el objetivo general de esta investigación, se busca diseñar modelos que permitan predecir los resultados de futbol de la Premier League, los cuales pretenderán generar una propuesta ante un problema planteado y aplicar nuevos conocimientos para su desarrollo. Por tal sentido, este trabajo de investigación se emarcará en el contexto de una investigación aplicada ya que para Valarino, et al (2010) , “la investigación aplicada además de generar conocimiento, busca soluciones aceptables y pertinentes a un fenómeno social determinado” (p.4).

3.2 Fases de la Investigación

Para el siguiente trabajo investigación, el proceso de investigación se enmarcará en una primera etapa de obtención de información y una segunda de Desorrollo de los modelos las cuales poseen las siguientes características:

- **Etapa I – Obtención de información:** En esta etapa se realizará un proceso de descarga, análisis y depuración de la información que interviene en la elaboración de los modelos de predicción. Esta información contempla los Tweets, la información de los resultados de partidos anteriores y la

información de las posibilidades de las casa de apuestas. Esto para dar cumplimiento al objetivo 1 de este trabajo de investigación.

- **Etapa II – Desarrollo de los modelos:** En esta etapa de diseñarán los modelos predictivos con base en la información recopilada en la Etapa I. Todo esto para dar cumplimiento a los objetivos 2, 3, 4 y 5 planteados en este trabajo de investigación.

3.3 Procedimiento por Objetivos

A continuación se muestra una tabla que contiene el procedimiento por objetivos que será realizado durante la ejecución de las fases de la investigación.

Tabla 1 Procedimiento por objetivos según las fases de la investigación.

		Procedimiento			
Fase del Proyecto	**Objetivo**	**Paso 1**	**Paso 2**	**Paso 3**	**Paso 4**
Etapa I - Obtención de información y duración:	1) Descargar, clasificar, guardar y depurar información referente a Tweets, resultados de partidos anteriores y posibilidades de las casas de apuestas de equipos de futbol de la Premier League.	Levantamiento, recolección y almacenamiento de información proveniente de Twitter.	Levantamiento, recolección y almacenamiento de información relacionada con resultados de partidos anteriores.	Levantamiento, recolección y almacenamiento de información relacionada con posibilidades de casas de puestas.	Análisis y depuración de la información que será utilizada para crear los modelos predictivos.
Etapa II - Desarrollo de los modelos	2) Crear un modelo basado en las opiniones emitidas en la red social Twitter.	Obtener los Features(bag words).	Entrenar y probar el modelo con base en datos de cada semana y acumulado de semanas.	Probar el modelo con resultados de partidos individuales.	N/A
	3) Crear un modelo estadístico basados en los resultados de partidos anteriores que tome en consideración una regresión lineal generalizada.	Análisis de la información de partidos anteriores	Aplicar un modelo de regresión lineal generalizada.	Cálculo de la exactitud del modelo y análisis de Resultados.	N/A
	4) Comparar los resultados del modelo estadístico basado en resultados anteriores con los datos de casas apuestas y los resultados del modelo basado en opiniones en Twitter .	Tomar la exactitud del modelo basado en resultados anteriores y compararlo con las posibilidades (odds) de las casas de apuesta.	Tomar la exactitud del modelo basado en opiniones en twitter y comparárlo con el modelo basado en resultados anteriores y los datos de las casas de apuestas.	N/A	N/A
	5) Análisis de competitividad de la Premier League.	Crear método de medición de la competitividad	Aplicar el método de medición en la Premier League	Comparar la competitividad de la Premier Legue con otra liga de futbol europea	N/A

N/A: No aplica

A continuación, se detallan los pasos que comprende el procedimiento de ejecución para el cumplimiento de cada uno de los objetivos.

Objetivo #1: *Descargar, clasificar, guardar y depurar información referente a Tweets, resultados de partidos anteriores y posibilidades de las casas de apuestas de equipos de futbol de la Premier League*

Paso 1 - Levantamiento, recolección y almacenamiento de información proveniente de Twitter: Se recopilan los datos de Twitter utilizando palabras claves que tengan relación con los 20 equipos que participan en la Premier League.

Los datos que se recopilaron tienen relación directa con cada una de las cuentas oficiales de twitter de cada uno de los equipos que participa en la competición, de esta manera, se minimiza el riesgo a recopilar tweets que tengan relación con otros tópicos diferentes al del equipo de fútbol.

Debido a que la API oficial de Twitter tiene limitaciones de tiempo, no se pueden obtener tweets anteriores a una semana. Algunas herramientas brindan acceso a tweets más antiguos, pero en la mayoría de ellos se tiene que gastar dinero antes. Para obtener tweets que tenga una antigüedad mayor a una semana, básicamente, cuando se ingresa en la página de Twitter, se inicia un cargador de desplazamiento, si se desplaza hacia abajo, se comienza a visualizar tweets.

Finalmente, se espera recuperar los tweets de 10 semanas, partiendo del 06 de septiembre de 2017. Luego, estas 10 semanas de datos se dividirán para así crear el dataset de entrenamiento y el de prueba.

Paso 2 - Levantamiento, recolección y almacenamiento de información relacionada con resultados de partidos anteriores. Para recolectar los datos relacionados con los partidos de futbol de la Premier League, se utilizará un paquete de R llamado FootballR el cual utiliza algunas APIs de futbol que son gratuitas.

Se espera recurar los resultados de los partidos de futbol para la temporada 2017-2018. Se considera que, para efectos de predictivos, los partidos de las temporadas anteriores pueden atentar contra la exactitud del modelo tomando en consideración que las características de los equipos y su rendimiento puede cambiar considerablemente de una temporada a otra.

Paso 3 - Levantamiento, recolección y almacenamiento de información relacionada con posibilidades de casas de apuestas. Para recolectar los datos relacionados con las posibilidades de las casas de apuestas de los partidos de futbol de la Premier League, se descargará un archivo .csv desde (football-data.co.uk, 2018).

Paso 4 - Análisis y depuración de la información que será utilizada para crear los modelos predictivos. Antes de realizar el análisis de sentimiento, se debe realizar un prepocesamiento de los datos que permita colocar todas las palabras en minúscula, quitar signos de puntuación, quitar palabras que no aportan información (ej.: artículos, “http”, https”, etc.) y separar las palabras resultantes entre ellas. Adicionalmente, Para entrenar el modelo, los datos se deben clasificar según el

resultado que hayan obtenido en cada resultado (ganar, perder, empatar). Luego de que los tweets estén clasificados de acuerdo al grupo en el que pertenecen, se debe almacenar todos los tweets en un archivo para luego ser procesado y obtener los Features.

Objetivo #2: *Crear un modelo basado en las opiniones emitidas en la red social Twitter.*

Paso 1 - Obtener los Features (bag words) que serán utilizados para el modelo. Luego de que los datos se encuentran clasificados de acuerdo al grupo en el que pertenecen y luego de que son depurados, se debe calcular las palabras más frecuentes para crear el Bag of Words que luego se utilizan para entrenar el modelo.

Paso 2 - Entrenar y probar el modelo con base en datos de cada semana y acumulado de semanas. Luego de obtener las palabras más frecuentes, se debe entrenar el modelo utilizando los datos de cada semana y utilizando el acumulado de los datos de las semanas. De esta manera, se puede realizar una comparación entre los modelos que surgen producto de los datos de cada semana con los modelos que toman en consideración los datos acumulados de las semanas. Para realizar esta comparación, se prueban los modelos con un conjunto de datos de prueba y se comparan las exactitudes de cada modelo.

Paso 3 - Probar el modelo con resultados de partidos individuales. De manera de seguir probando el modelo, se toma el modelo ya entrenado con la mayor cantidad de datos (10 semanas acumuladas) y se prueba con datos de Tweets que solo pertenezcan a un solo partido. Esta prueba se realiza con partidos que pertenezcan a una misma fecha de competición. Luego, se analizan los resultados y se calcula el nivel de exactitud.

Objetivo #3: *Crear un modelo estadístico basado en los resultados de partidos anteriores que tome en consideración una regresión lineal generalizada.*

Paso 1 – Análisis de la información de partidos anteriores. Una vez depurada la información de partidos anteriores, se analiza qué función de distribución de

probabilidades tiene los datos, y en base a eso, se aplica una regresión lineal generalizada que permite predecir los resultados de un partido de futbol.

Paso 2 - Aplicar un modelo de regresión lineal generalizada. Luego de analizar las características de los datos de los resultados de partidos de futbol anteriores, se aplica una regresión lineal generalizada sobre los datos.

Paso 3 - Cálculo de la exactitud del modelo y análisis de Resultados. Luego de generar el modelo de regresión lineal, se ejecuta el modelo con un conjunto de datos de prueba para calcular la exactitud del modelo para los partidos de las fechas 32, 33, 34 y 35 de competición.

Objetivo #4: *Comparar los resultados del modelo estadístico basado en resultados anteriores con los datos de casas apuestas y los resultados del modelo basado en opiniones en Twitter .*

Paso 1 - Tomar la exactitud del modelo basado en resultados anteriores y compararlo con las posibilidades (odds) de las casas de apuesta. Una vez calculada la exactitud del modelo basados en resultados anteriores, se comprará con las posibilidades de las casas de apuestas. Para ello, se tomarán las posibilidades de dos casas de apuestas, Bet365 y Blue Square, y se compararán con la exactitud del modelo para las fechas de competición 32, 33, 34, y 35. El objetivo es conocer si el modelo tiene mejores beneficios para predecir partidos de la Premier League que solo tomar en consideración las posibilidades de las casas de apuestas.

Paso 2 - Tomar la exactitud del modelo basado en opiniones en twitter y compáralo con el modelo basado en resultados anteriores y los datos de las casas de apuestas. Una vez comparada la exactitud del modelo basado en resultados anteriores con las posibilidades de las casas de apuesta, se compararán con la exactitud del modelo basado en opiniones en Twitter. Para ello, se tomará el modelo basado en opiniones en Twitter de 10 semanas acumuladas (ya entrenado), se tomarán los Tweets de las fechas 32, 33, 34 y 35, y se utilizará el modelo para predecir los resultados partido a partido para cada una de estas fechas. Luego se

compararán los resultados con la exactitud del modelo basado en partidos anteriores y con las posibilidades de las casas de apuestas.

Objetivo #5: *Análisis de competitividad de la Premier League.*

Paso 1 – Crear método de medición de la competitividad. Se creará un método de medición para representar el nivel de competitividad para un sistema de competición donde todos los equipos se enfrentan entre ellos con la misma cantidad de partidos en casa y de visita. El objetivo será representar gráficamente el nivel de competitividad que existe entre los equipos que conforman la competencia.

Paso 2 – Aplicar el método de medición en la Premier League. Se utilizará el método de medición de la competitividad en la Premier League y se observará qué tan competitiva puede ser esta liga.

Paso 3 – Comparar la competitividad de la Premier Legue con otra liga de futbol europea. Se cogerá La Liga de España y se aplicará el método de medición para comparar con la Premier League.

3.4 Aspectos Éticos

Para este trabajo de investigación se hace constar que se respetará los derechos de autor y que cada una de las informaciones, conceptos y puntos de vista que no son genereadas por el autor de este trabajo especial de grado, serán debidamente referenciadas de acuerdo a las normas APA de manera de mantener su autoria.

3.5 Cronograma

El presente trabajo de investigación posee un cronograma con fecha de inicio de actividades al 1 de Marzo de 2018 y una fecha de culminación al 7 de Junio del 2018 con alrededor de 13 semanas.

Tabla 2 Cronograma de actividades

Objetivo	Fecha de Inicio	Fecha Fin	Semanas
1) Descargar, clasificar, guardar y depurar información referente a Tweets, resultados de partidos anteriores y posibilidades de las casas de apuestas de equipos de futbol de la Premier League.	1/3/2018	1/5/2018	8
2) Crear un modelo basado en las opiniones emitidas en la red social Twitter.	1/5/2018	15/5/2018	2
3) Crear un modelo estadístico basados en los resultados de partidos anteriores que tome en consideración una regresión lineal generalizada.	15/5/2018	22/5/2018	1
4) Comparar los resultados del modelo estadístico basado en resultados anteriores con los datos de casas apuestas y los resultados del modelo basado en opiniones en Twitter .	22/5/2018	1/6/2018	1
5) Análisis de competitividad de la Premier League.	1/6/2018	7/6/2018	1

CAPITULO IV: DESARROLLO DE LOS OBJETIVOS ESPECÍFICOS

4.1 Objetivo No.1: Descargar, clasificar, guardar y depurar información referente a Tweets, resultados de partidos anteriores y posibilidades de las casas de apuestas de equipos de futbol de la Premier League

4.1.1 Levantamiento, recolección y almacenamiento de información proveniente de Twitter

Para realizar la descarga de los tweets, no se utilizaron las APIs de Twitter ya que las APIs de uso gratuito poseen limitaciones que podían poner en peligro el desarrollo de este trabajo final de máster. La API de búsqueda de Twitter solo permite capturar tweets con una antigüedad no mayor a 7 días desde el momento en que se hace la consulta. La API de Streaming permite solo capturar tweets en línea y no permite realizar una consulta de data histórica. En su lugar, se utilizó una herramienta programada en Python para obtener tweets antiguos que, básicamente, funciona de la siguiente forma: cuando se ingresa en la página de Twitter, se inicia un cargador de desplazamiento, si se desplaza hacia abajo, se comienza a visualizar tweets. En el anexo 2, se muestra el código de la herramienta que se utilizó para descargar los tweets. Luego de realizar la descarga de los tweets referentes a un período de 10 semanas, partiendo del 06 de septiembre de 2017 (jornada 4 de la competición), se recopilaron 1661322 tweets. Esta información fue almacenada en archivos separados por cada jornada y por cada equipo. A continuación, se muestra una figura que muestra un ejemplo de la estructura de los archivos que se descargaron.

	username	date	retweets	favorites	text	geo	mentions	hashtags
2	Rajas40	19/10/2017 01:59	0	0	Win is all that matters in games like this. Can't always score four			
3	Rofloveer	19/10/2017 01:59	0	0	Así no más quedo. @ManUtd		@ManUtd	
4	bay_krizMUw	19/10/2017 01:59	0	0	Retweeted Indonesian @ManUtd (@indomanutd): Rashford nend		@ManUtd @indomanutd	
5	OleMU	19/10/2017 01:58	0	0	@ManUtd Found our new LB		@ManUtd	
6	gratped	19/10/2017 01:58	0	0	Oh man you're the best thanks for that goal			
7	MetaGoles	19/10/2017 01:58	0	0	#ChampionsLeague Grupo A: Hoy en Lisboa, @SLBenfica cayó ante		@SLBenfica @ManUtd	#ChampionsLeague
8	MayangoArku1	19/10/2017 01:58	0	0	Rashford			
9	deanhope09	19/10/2017 01:56	0	0	How's the leg kid? Will miss ya			
10	Jason_Bugatti	19/10/2017 01:56	0	1	The Gaffer! @ManUtd #RedArmy #ChampionsLeague #EstadiodaLu		@ManUtd	#RedArmy #Champic
11	MayangoArku1	19/10/2017 01:56	0	0	Job well done			
12	madridistamil	19/10/2017 01:56	0	0	vaya cantada del portero,,, en las ferias de pueblo seguro los encontrais mejores			
13	Dean_Burns10	19/10/2017 01:56	0	1	@GavStew99 youre good at filling voids aren't you? With your may		@GavStew99	
14	Aghost6969	19/10/2017 01:56	0	0	Koyo jagoku ngono lo @MafiaWasit .. @persisofficial mbi @ManUt		@MafiaWasit @persisofficial @ManUtd	
15	Thedublinbeast	19/10/2017 01:55	0	5	Stop playing like Tom Cleverly yeah			
16	UtdIndonesiaSRA	19/10/2017 01:55	0	0	regranned from @manutd - FT: Benfica 0 #MUFC 1. marcusrashfor		@manutd	#MUFC
17	Dave999Sutty	19/10/2017 01:55	1	0	Technically he didn't score. The goalie did by taking it over the line.		#OwnGoal #svilar	#OwnGoal #svilar
18	futbolecuador	19/10/2017 01:55	0	0	(VIDEO) El @ManUtd de Antonio #Valencia venció por la mínima al		@ManUtd	#Valencia #Benfica #
19	201405024	19/10/2017 01:54	0	0	Lets bounce back to the 4:0 trend we getting weak Boke Herrera!			

Figura 3 Estructura de archivo contenedores de tweets

Como se pueden observar en la figura anterior, existen algunos campos de esta estructura que no serán utilizados para el presente trabajo final de máster. Solo se utilizará la fecha en la que se produjo el tweet (“date”) y el contenido del tweet (“text”).

Es importante resaltar que la captura de los tweets de cada equipo se hizo tres días antes del comienzo de la jornada. Es decir, para la primera jornada que se tomó en consideración para la recolección, se consideró un período de recolección desde el 6 de Septiembre de 2017 debido a que la jornada empezó el 9 de Septiembre de 2017.

4.1.2 Levantamiento, recolección y almacenamiento de información relacionada con resultados de partidos anteriores

Para realizar la descarga de los resultados de los partidos anteriores de la Premier League, se utilizó un paquete de R llamado FootballR, el cual utiliza algunas APIs que son de uso gratuito.

Utilizando este paquete de R, se obtuvieron los resultados de los partidos de futbol de la Premier League correspondientes a la temporada 2017-2018 hasta la fecha de la consulta (1 de Mayo de 2018).

La siguiente figura, muestra los campos que contienen las tablas que conforman la información de los resultados de los partidos de la Premier League de la temporada 2017-2018.

	home	away	homeGoals	awayGoals	date
1	Arsenal	Leicester	4	3	2017-08-11T18:45:00Z
2	Watford	Liverpool	3	3	2017-08-12T11:30:00Z
3	C Palace	Huddersfield Town	0	3	2017-08-12T14:00:00Z
4	West Brom	Bournemouth	1	0	2017-08-12T14:00:00Z
5	Chelsea	Burnley	2	3	2017-08-12T14:00:00Z
6	Everton	Stoke	1	0	2017-08-12T14:00:00Z
7	Southampton	Swansea	0	0	2017-08-12T14:00:00Z
8	Brighton & Hove Albion	Man City	0	2	2017-08-12T16:30:00Z

Figura 4 Campos de las tablas que contienen los resultados de los partidos de la Premier League de la temporada 2017-2018

4.1.3 Levantamiento, recolección y almacenamiento de información relacionada con posibilidades de casas de puestas

La información de las posibilidades de las casas de apuestas es pública y gratuita. Esta información se pudo obtener desde (football-data.co.uk, 2018) a través de la descarga de un archivo .csv. Este archivo contiene información diversa entre la cual se destaca las posibilidades de casas de apuestas como (bet365, 2018) y (Blue Square, 2018). A continuación, se muestra una figura que muestra la estructura del dataset.

Div	Date	HomeTeam	AwayTeam	B365H	B365D	B365A	BWH	BWD	BWA
E0	11/8/2017	Arsenal	Leicester	1.53	4.5	6.5	1.5	4.6	6.75
E0	12/8/2017	Brighton	Man City	11	5.5	1.33	11	5.25	1.3
E0	12/8/2017	Chelsea	Burnley	1.25	6.5	15	1.22	6.5	12.5
E0	12/8/2017	Crystal Palace	Huddersfield	1.83	3.6	5	1.8	3.5	4.75
E0	12/8/2017	Everton	Stoke	1.7	3.8	5.75	1.7	3.6	5.5
E0	12/8/2017	Southampton	Swansea	1.62	4	6.5	1.57	4	6
E0	12/8/2017	Watford	Liverpool	6	4.2	1.62	6	4.2	1.55
E0	12/8/2017	West Brom	Bournemouth	2.4	3.3	3.3	2.4	3.2	3.1
E0	13/8/2017	Man United	West Ham	1.3	5.75	12	1.28	5.5	11
E0	13/8/2017	Newcastle	Tottenham	5.5	4	1.7	5.25	3.8	1.67
E0	19/8/2017	Bournemouth	Watford	2	3.6	4	1.95	3.5	4

Figura 5 Campos de las tablas que contienen las posibilidades de las casas de apuestas de los partidos de la Premier League de la temporada 2017-2018

4.1.4 Análisis y depuración de la información que será utilizada para crear los modelos predictivos

Es importante resaltar que los tweets que se utilizan para crear el modelo no deben comprender el tiempo en el que transcurre el partido o luego de que haya finalizado. Por esta razón, se realizó un filtrado de la información con respecto a la fecha y hora exacta a la que inició cada partido. De manera de llevar un control, se realizaron unos archivos de control que contienen información de la fecha y hora exacta para luego realizar el filtrado.

```
Notes:

#Fecha y hora para filtrar los datos
    From 20170906 000000 to 20170909 123000 - 1

        Man City, Liverpool

    From 20170906 000000 to 20170909 150000 - 2

        Arsenal,Bournemouth, Brighton & Hove Albion, West Bromwich Albion, Everton, Tottenham, Leicester, Chelsea, Southampton, Watford

    From  20170906 000000 to  20170909 173000 - 3

        Stoke City, Manchester United

    From 20170906 000000 to 20170910 133000 - 4

        Burnley, Crystal Palace

    From 20170906 000000 to 20170910 160000 - 5

        Swansea, Newcastle United

    From 20170906 000000 to 20170911 200000 - 6

        West Ham United, Huddersfield Town
```

Figura 6 Ejemplo de archivo de control para la Jornada No.1 de la Premier League

El primer paso para la combinación de todos los archivos fue importar los dataset.

Luego se quitaron las columnas del dataset que no se utilizarán para este trabajo final de máster, se eliminaron las filas que estuviesen vacías y, por último, se agregó una categoría que tiene relación con los resultados propios de cada partido para la fecha de competición (ganar, empatar, perder).

En un segundo paso, se realizó la unión de todos los datasets de cada equipo en un solo dataset por semana y luego se realizó la unión en un dataset final pero no sin antes borrar las filas duplicadas.

Se ejecutó un tercer paso que consiste en dividir la información para crear un conjunto de entrenamiento y otro de prueba. Para realizar esta división, se utilizó el modelo de validación de división por porcentaje. En este caso, se destinó el 60% del dataset para el conjunto de entrenamiento y 40% para el conjunto de prueba.

Se tomó el conjunto de datos de entrenamiento y se verificó que no se mencionara a más de un equipo en un mismo tweet. De esta manera, se reduce el riesgo de interpretar un tweet para cualquiera de los dos o más equipos mencionados en el tweet. Luego de realizar este filtrado, la cantidad de tweets se redujo hasta alcanzar 1661322 tweets.

Luego, se tomó el conjunto de datos de entrenamiento resultante y se ejecutaron los siguientes de pasos para la depuración de las palabras que contienen los tweets:

- **Eliminar mayúsculas:** De manera de que no hubiese diferenciación entre las palabras que contienen mayúsculas y las que no, se eliminaron todas las letras que estuviesen en mayúscula y se sustituyeron por palabras en minúscula.
- **Eliminar valores numéricos:** Debido a que los números no aportan información adicional para poder categorizar los tweets entre las clases, se decidió eliminar esta información por completo del conjunto de datos.
- **Eliminar palabras vacías:** Las palabras vacías son palabras que no poseen significado como artículos, pronombres y preposiciones. Para este conjunto de datos, se decidió eliminar este tipo de palabras ya que tampoco aportan ninguna información para poder categorizar los tweets en las clases.
- **Eliminar signos de puntuación:** Debido a que los signos de puntuación no aportan información adicional para poder categorizar los tweets entre las clases, se decidió eliminar esta información por completo del conjunto de datos.
- **Eliminar los espacios en blanco:** Debido a que los espacios en blanco no aportan información adicional para poder categorizar los tweets entre las clases, se decidió eliminar esta información por completo del conjunto de datos.
- **Cambiar palabras por palabras madres:** Las palabras madres son las palabras raíz de una palabra que, para el caso de análisis de sentimiento, tiene el mismo impacto. Por ejemplo, es vez de tomar 3 palabras como “winner”, “winning” y “win”, se toma una sola palabra madre. En este caso la palabra raíz es “win”.
- **Eliminar las palabras menos frecuentes:** Finalmente, para obtener el “bag of words” que será utilizado en el modelo, se eliminan las palabras menos frecuentes del conjunto de datos.

Para el caso de la depuración del conjunto de datos de los resultados de los partidos anteriores, primero se eliminó la columna correspondiente a la fecha del partido. Luego, de manera de realizar una regresión lineal que tomara en consideración la

cantidad de goles en función de si el equipo se encuentra jugando en casa o no, se agregó una nueva columna que ofrece información de si el equipo estaba jugando en casa (valor 1) o no (valor 0). Adicionalmente de esta columna y de la columna del equipo, también se encuentra la columna del oponente y de la cantidad de goles. De esta manera se crea una relación entre el equipo de casa y la cantidad de goles anotados. A continuación, se muestra la estructura del conjunto de datos depurados de resultados de partidos anteriores.

	goals	team	opponent	home
1	4	Arsenal	Leicester	1
2	3	Watford	Liverpool	1
3	0	C Palace	Huddersfield Town	1
4	1	West Brom	Bournemouth	1
5	2	Chelsea	Burnley	1
6	1	Everton	Stoke	1
7	0	Southampton	Swansea	1
8	0	Brighton & Hove Albion	Man City	1
9	0	Newcastle	Tottenham	1

Figura 7 Estructura del conjunto de datos depurados de resultados de partidos anteriores de la Premier League

Finalmente, para el caso de las posibilidades de las casas de apuestas, se tomó el conjunto de datos y se filtraron algunas columnas hasta quedar solo dos casas de apuestas. A continuación, se muestra la estructura del conjunto de datos de las posibilidades de las casas de apuestas.

Date	HomeTeam	AwayTeam	B365H	B365D	B365A	BWH	BWD	BWA
11/8/2017	Arsenal	Leicester	1.53	4.5	6.5	1.5	4.6	6.75
12/8/2017	Brighton	Man City	11	5.5	1.33	11	5.25	1.3
12/8/2017	Chelsea	Burnley	1.25	6.5	15	1.22	6.5	12.5
12/8/2017	Crystal Palace	Huddersfield	1.83	3.6	5	1.8	3.5	4.75
12/8/2017	Everton	Stoke	1.7	3.8	5.75	1.7	3.6	5.5
12/8/2017	Southampton	Swansea	1.62	4	6.5	1.57	4	6
12/8/2017	Watford	Liverpool	6	4.2	1.62	6	4.2	1.55
12/8/2017	West Brom	Bournemouth	2.4	3.3	3.3	2.4	3.2	3.1
13/8/2017	Man United	West Ham	1.3	5.75	12	1.28	5.5	11
13/8/2017	Newcastle	Tottenham	5.5	4	1.7	5.25	3.8	1.67
19/8/2017	Bournemouth	Watford	2	3.6	4	1.95	3.5	4

Figura 8 Estructura del conjunto de datos de posibilidades de casas de apuestas para los partidos de la Premier League

4.2 Objetivo No.2: Crear un modelo basado en las opiniones emitidas en la red social Twitter.

4.2.1 Obtener los Features (Bag of Words)

Una vez ya depurados los datos provenientes de la red social Twitter y colocada la categoría de “ganar”, “perder”, “empatar” según sea el resultado de cada partido, se procedió a calcular cuáles son las palabras más frecuentes que se utilizan en el conjunto de datos. Estas palabras más frecuentes se utilizarán posteriormente para entrenar el modelo.

De manera de realizar una primera prueba, solo se tomaron los datos de la primera semana de análisis y se realizó una nube de palabras con las 70 palabras más frecuentes que se repiten en esa semana. A continuación, se muestran las 70 palabras más frecuentes en el conjunto de datos de la primera semana de competición.

Figura 9 70 palabras más frecuentes en el conjunto de datos de la primera semana de competición.

Adicionalmente, se realizó una nube de palabras más frecuentes para cada una de las categorías de manera de observar si existen palabras que identifiquen plenamente a una clase. A continuación, se muestran las palabras más frecuentes para cada una de las categorías en el conjunto de datos de la primera semana de competición.

Figura 10 70 palabras más frecuentes para cada una de las categorías en el conjunto de datos de la primera semana de competición

Como se puede ver en las dos figuras anteriores, existen palabras que tienen relación con el contexto y que se repiten con bastante frecuencia, como por ejemplo “win”, “match”, “game”, “football”, “player”, “score”, “need”, “better”, entre otras, pero también se puede observar que, dentro de las nubes de palabras más frecuentes, aparecen los nombres de los equipos como unas de las palabras que más se repiten en el conjunto de datos.

Adicionalmente, se puede observar que para la categoría “Win”, existen palabras con frecuencias de repetición bien establecidas, mientras que para las categorías “Draw” y “Lose”, no existen palabras que tengan una frecuencia de repetición bien marcada. En general, esto podría llevar a pensar que el modelo de dos clases (“ganar” y “no ganar”) puede que sea más exacto.

Siguiendo con la primera semana de competición, se calcularon las palabras más frecuentes eliminando los nombres de los equipos. A continuación, se muestran las nubes de las palabras más frecuentes eliminando los nombres de los equipos.

Figura 11 70 palabras más frecuentes para cada una de las categorías en el conjunto de datos de la primera semana de competición eliminando el nombre de los equipos

Como se puede ver en la figura anterior, si se quitan los nombres de los equipos, la nube de palabras aparenta ser muy parecida para cada una las categorías. Con este resultado, se puede llegar a pensar que los nombres de los equipos poseen un rol importante dentro del “Bag of Words” y por ende en la exactitud que puede llegar a tener el modelo.

Es importante resaltar que para entrenar el modelo de cada semana y cada uno de los modelos con los datos de las semanas acumuladas, se observaron las nubes con las palabras más frecuentes en cada caso. Por motivos de simplicidad, se mostrarán cada una de las nubes de datos en la siguiente sección junto con los resultados de las pruebas del modelo.

4.2.2 Entrenar y probar el modelo con base en datos de cada semana y acumulado de semanas.

De manera de comprobar qué tan importante pueden ser los nombres de los equipos para la exactitud del modelo, se entrenó y se probó un modelo con un “Bag of Words” que cogiera los nombres de los equipos y otro modelo con otro “Bag of Words” que no cogiera los nombres de los equipos.

Primera Semana

Primero se entrenó y se probó el modelo tomando en consideración los nombres de los equipos. Cabe destacar que los datos de la primera semana de competición se encuentran distribuidos entre las categorías de la siguiente manera.

Draw	Lose	Win
0.2703436	0.3013716	0.4282848

Figura 12 Distribución de los datos de la primera semana de competición según su categoría

Como se puede observar en la figura anterior, para la primera semana de competición, hay un 27% de Tweets que se encuentran en la categoría “Draw”, 30% de Tweets que se encuentran en la categoría “Lose” 43% de Tweets que se encuentran en la categoría “Win”.

Luego de entrenar y probar el modelo tomando en consideración los nombres de los equipos, se obtuvieron los siguientes resultados.

	True			
Prediction	Draw	Lose	Win	
Draw	8496	515	728	87.23%
Lose	1474	8206	5099	55.52%
Win	7438	10685	21751	54.54%

Overall Statistics

Accuracy : 0.5972

Figura 13 Resultados de probar el modelo con la primera semana de competición y tomando en consideración los nombres de los equipos

Como se puede observar en la figura anterior, tomando en consideración los nombres de los equipos y la primera semana de competición, se obtuvo un modelo con una exactitud del 60%, consiguiendo la mejor exactitud para la categoría “Draw”.

Luego, se filtraron los nombres de los equipos del “Bag of Words” y se procedió a entrenar y a probar nuevamente el modelo. A continuación, se muestran los resultados del modelo con la primera semana de competición sin considerar los nombres de los equipos.

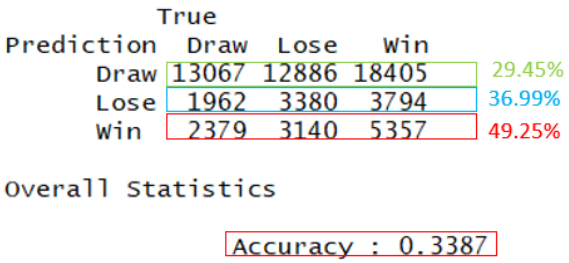


Figura 14 Resultados de probar el modelo con la primera semana de competición sin considerar los nombres de los equipos

Como se puede observar en la figura anterior, la exactitud del modelo descendió considerablemente hasta alcanzar un 34% aproximadamente. Con estos resultados, podemos concluir que los nombres de los equipos en el "Bag of Word" sí poseen relevancia para el modelo y puede afectar su la exactitud.

Debido a que los nombres de los equipos en el "Bag of Word" sí producen un impacto en la exactitud del modelo, no se considera eliminar el nombre de los equipos para el entrenamiento del modelo en las siguientes semanas de competición.

Segunda Semana

Para analizar la segunda semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la segunda semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la segunda semana de competición para cada una de las categorías.

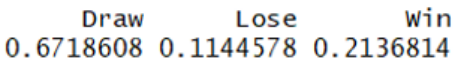


Figura 15 Distribución de los datos de la segunda semana de competición según su categoría

Como se puede observar en la figura anterior, para la segunda semana de competición, hay un 67% de Tweets que se encuentran en la categoría "Draw", 12% de Tweets que se encuentran en la categoría "Lose" 21% de Tweets que se encuentran en la categoría "Win".

Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la segunda semana de competición.

Figura 16 70 palabras más frecuentes en el conjunto de datos de la segunda semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la segunda semana de competición.

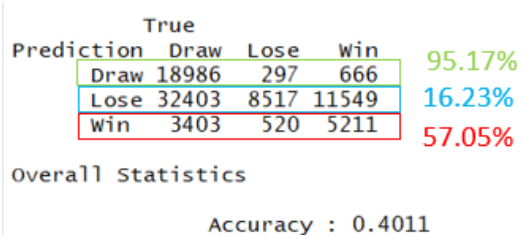


Figura 17 Resultados de probar el modelo con la segunda semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 40%. La exactitud del modelo para la categoría “Draw” es de 95.17%, para la categoría “Lose” es de 16.23% y para la categoría “Win” es de 57.05%.

En cuanto a los datos acumulados hasta la segunda semana, a continuación, se muestra la distribución de los datos hasta la segunda semana de competición para cada una de las categorías.

Draw	Lose	Win
0.4947091	0.1969255	0.3083655

Figura 18 Distribución de los datos hasta la segunda semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la segunda semana de competición.

Figura 19 70 palabras más frecuentes en el conjunto de datos hasta la segunda semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la segunda semana de competición.

	True			
Prediction	Draw	Lose	Win	
Draw	12053	987	3822	71.48%
Lose	47449	26720	32610	25.02%
Win	12699	1033	8573	38.43%

Overall Statistics

Accuracy : 0.3244

Figura 20 Resultados de probar el modelo hasta la segunda semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 32.4%. La exactitud del modelo para la categoría “Draw” es de 71.48%, para la categoría “Lose” es de 25.02% y para la categoría “Win” es de 38.43%.

Tercera Semana

Para analizar la tercera semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la tercera semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la tercera semana de competición para cada una de las categorías.

```
     Draw      Lose       Win
0.0150546 0.1369785 0.8479669
```

Figura 21 Distribución de los datos de la tercera semana de competición según su categoría

Como se puede observar en la figura anterior, para la tercera semana de competición, hay un 1.5% de Tweets que se encuentran en la categoría “Draw”, 13.7% de Tweets que se encuentran en la categoría “Lose” 84.8% de Tweets que se encuentran en la categoría “Win”.

Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la tercera semana de competición.

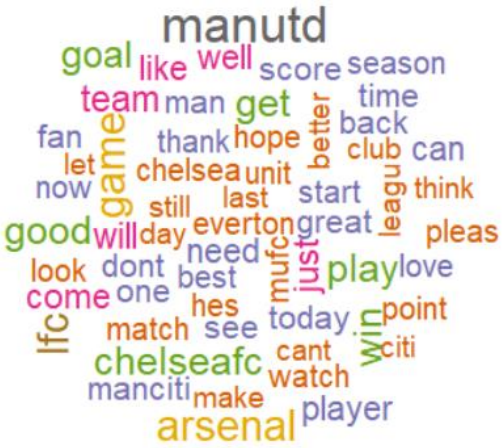


Figura 22 70 palabras más frecuentes en el conjunto de datos de la tercera semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la tercera semana de competición.

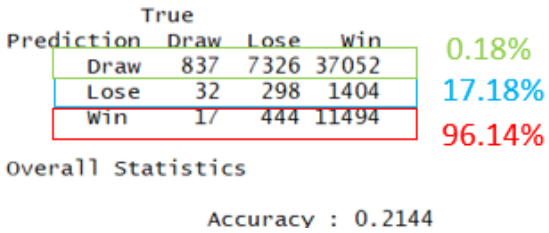


Figura 23 Resultados de probar el modelo con la tercera semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 21.44%. La exactitud del modelo para la categoría “Draw” es de 0.18%, para la categoría “Lose” es de 17.18% y para la categoría “Win” es de 96.14%.

En cuanto a los datos acumulados hasta la tercera semana, a continuación, se muestra la distribución de los datos hasta la tercera semana de competición para cada una de las categorías.

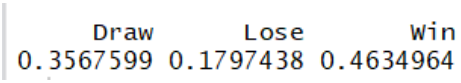


Figura 24 Distribución de los datos hasta la tercera semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la tercera semana de competición.

Figura 25 70 palabras más frecuentes en el conjunto de datos hasta la tercera semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la tercera semana de competición.

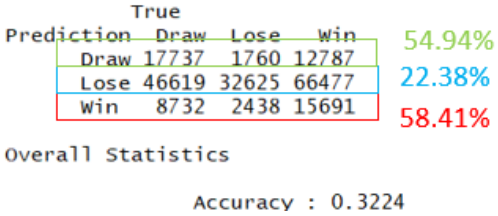


Figura 26 Resultados de probar el modelo hasta la tercera semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 32.24%. La exactitud del modelo para la categoría “Draw” es de 54.94 %, para la categoría “Lose” es de 22.38% y para la categoría “Win” es de 58.41%.

Cuarta Semana

Para analizar la cuarta semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la cuarta semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la cuarta semana de competición para cada una de las categorías.

```
      Draw       Lose        Win
0.07265155 0.37994637 0.54740208
```

Figura 27 Distribución de los datos de la cuarta semana de competición según su categoría

Como se puede observar en la figura anterior, para la cuarta semana de competición, hay un 7.3% de Tweets que se encuentran en la categoría “Draw”, 37.99% de Tweets que se encuentran en la categoría “Lose” 54.7% de Tweets que se encuentran en la categoría “Win”.

Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la cuarta semana de competición.

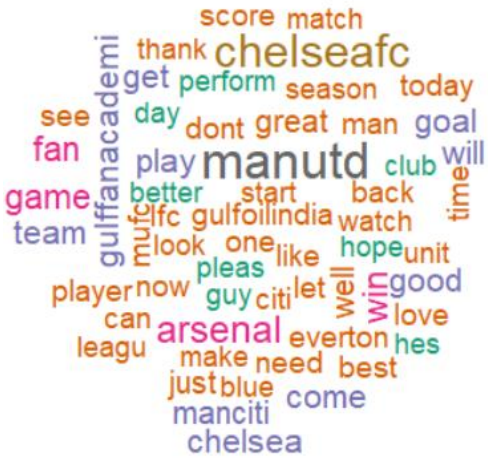


Figura 28 70 palabras más frecuentes en el conjunto de datos de la cuarta semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la cuarta semana de competición.

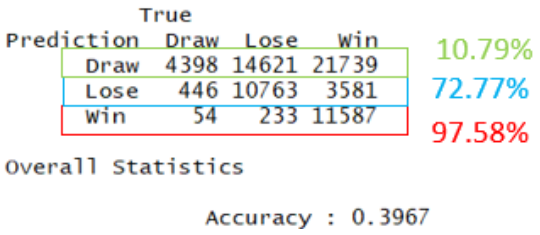


Figura 29 Resultados de probar el modelo con la cuarta semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 39.6%. La exactitud del modelo para la categoría “Draw” es de 10.8%, para la categoría “Lose” es de 72.77% y para la categoría “Win” es de 97.58%.

En cuanto a los datos acumulados hasta la cuarta semana, a continuación, se muestra la distribución de los datos hasta la cuarta semana de competición para cada una de las categorías.

```
     Draw      Lose       Win
0.2864095 0.2293163 0.4842742
```

Figura 30 Distribución de los datos hasta la cuarta semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la cuarta semana de competición.

Figura 31 70 palabras más frecuentes en el conjunto de datos hasta la cuarta semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la cuarta semana de competición.

```
                True
Prediction  Draw  Lose   Win
      Draw 14653  3883 14994   43.70%
      Lose 54609 54831 92055   27.21%
      Win   8724  3726 24814   66.58%
Overall Statistics

               Accuracy : 0.3463
```

Figura 32 Resultados de probar el modelo hasta la cuarta semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 34.63%. La exactitud del modelo para la categoría “Draw” es de 43.70%, para la categoría “Lose” es de 27.21% y para la categoría “Win” es de 66.58%.

Quinta Semana

Para analizar la quinta semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la quinta semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la quinta semana de competición para cada una de las categorías.

```
     Draw      Lose       Win
0.4437715 0.3692830 0.1869454
```

Figura 33 Distribución de los datos de la quinta semana de competición según su categoría

Como se puede observar en la figura anterior, para la quinta semana de competición, hay un 44.37% de Tweets que se encuentran en la categoría “Draw”, 36.92% de Tweets que se encuentran en la categoría “Lose” 18.69% de Tweets que se encuentran en la categoría “Win”. Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la quinta semana de competición.

Figura 34 70 palabras más frecuentes en el conjunto de datos de la quinta semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la quinta semana de competición.

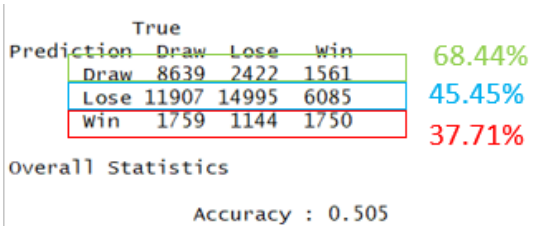


Figura 35 Resultados de probar el modelo con la quinta semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 50.5%. La exactitud del modelo para la categoría “Draw” es de 68.44%, para la categoría “Lose” es de 45.45% y para la categoría “Win” es de 37.71%.

En cuanto a los datos acumulados hasta la quinta semana, a continuación, se muestra la distribución de los datos hasta la quinta semana de competición para cada una de las categorías.

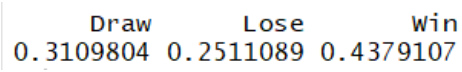


Figura 36 Distribución de los datos hasta la quinta semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la quinta semana de competición.

Figura 37 70 palabras más frecuentes en el conjunto de datos hasta la quinta semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la quinta semana de competición.

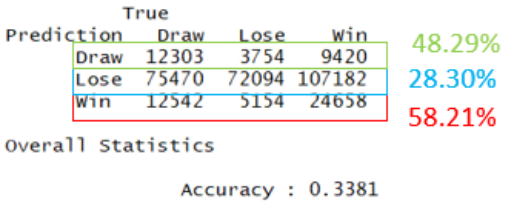


Figura 38 Resultados de probar el modelo hasta la quinta semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 33.81%. La exactitud del modelo para la categoría “Draw” es de 48.29%, para la categoría “Lose” es de 28.30% y para la categoría “Win” es de 58.21%.

Sexta Semana

Para analizar la sexta semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la sexta semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la sexta semana de competición para cada una de las categorías.

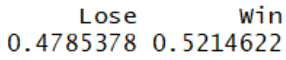


Figura 39 Distribución de los datos de la sexta semana de competición según su categoría

Como se puede observar en la figura anterior, para la sexta semana de competición, hay un 47.85% de Tweets que se encuentran en la categoría “Lose” y 52.15% de Tweets que se encuentran en la categoría “Win”. Para esta semana no existieron datos para la categoría “Draw”.

Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la sexta semana de competición.

Figura 40 70 palabras más frecuentes en el conjunto de datos de la sexta semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la sexta semana de competición.

ACC	TPR1	TPR2
59.73768	54.68011	81.44781

Figura 41 Resultados de probar el modelo con la sexta semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 59.73%. La exactitud del modelo para la categoría “Lose” es de 54.68%, y para la categoría “Win” es de 81.44 %.

En cuanto a los datos acumulados hasta la sexta semana, a continuación, se muestra la distribución de los datos hasta la sexta semana de competición para cada una de las categorías.

Draw	Lose	Win
0.2406083	0.3026130	0.4567786

Figura 42 Distribución de los datos hasta la sexta semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la sexta semana de competición.

Figura 43 70 palabras más frecuentes en el conjunto de datos hasta la sexta semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la sexta semana de competición.

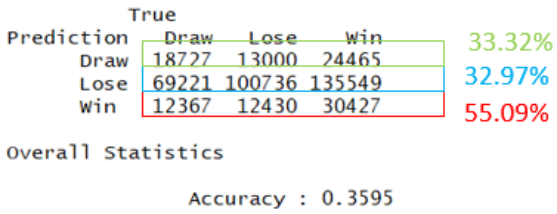


Figura 44 Resultados de probar el modelo hasta la sexta semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 35.95%. La exactitud del modelo para la categoría “Draw” es de 33.32%, para la categoría “Lose” es de 32.97% y para la categoría “Win” es de 55.09%.

Séptima Semana

Para analizar la séptima semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la séptima semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la séptima semana de competición para cada una de las categorías.

```
      Draw      Lose       Win
0.1044025 0.2163410 0.6792565
```

Figura 45 Distribución de los datos de la séptima semana de competición según su categoría

Como se puede observar en la figura anterior, para la séptima semana de competición, hay un 10.44% de Tweets que se encuentran en la categoría “Draw”, 21.63% de Tweets que se encuentran en la categoría “Lose” y 67.92 % de Tweets que se encuentran en la categoría “Win”. Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la séptima semana de competición.

Figura 46 70 palabras más frecuentes en el conjunto de datos de la séptima semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la séptima semana de competición.

```
          True
Prediction  Draw  Lose   Win
      Draw  5208  9436 28524   12.06%
      Lose   234  1662  1966   43.03%
      Win    166   522  5996   89.70%

Overall Statistics

               Accuracy : 0.2395
```

Figura 47 Resultados de probar el modelo con la séptima semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 23.95%. La exactitud del modelo para la categoría “Draw” es de 12.06%, para la categoría “Lose” es de 43.03% y para la categoría “Win” es de 89.7%.

En cuanto a los datos acumulados hasta la séptima semana, a continuación, se muestra la distribución de los datos hasta la séptima semana de competición para cada una de las categorías.

```
     Draw      Lose       Win
0.2250760 0.2927616 0.4821624
```

Figura 48 Distribución de los datos hasta la séptima semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la séptima semana de competición.

Figura 49 70 palabras más frecuentes en el conjunto de datos hasta la séptima semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la séptima semana de competición.

```
          True
Prediction  Draw  Lose   Win
      Draw 16719 12165 24529  31.30%
      Lose 70448 106562 153989  32.19%
      Win  18764 19060 48410  56.13%

Overall Statistics

               Accuracy : 0.3648
```

Figura 50 Resultados de probar el modelo hasta la séptima semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 36.48%. La exactitud del modelo para la categoría “Draw” es de 31.3%, para la categoría “Lose” es de 32.19% y para la categoría “Win” es de 56.13%.

Octava Semana

Para analizar la octava semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la octava semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la octava semana de competición para cada una de las categorías.

```
      Draw       Lose        Win
0.02130248 0.38677666 0.59192086
```

Figura 51 Distribución de los datos de la octava semana de competición según su categoría

Como se puede observar en la figura anterior, para la octava semana de competición, hay un 2.13 % de Tweets que se encuentran en la categoría “Draw”, 38.67 % de Tweets que se encuentran en la categoría “Lose” y 59.19 % de Tweets que se encuentran en la categoría “Win”. Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la octava semana de competición.

Figura 52 70 palabras más frecuentes en el conjunto de datos de la octava semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la octava semana de competición.

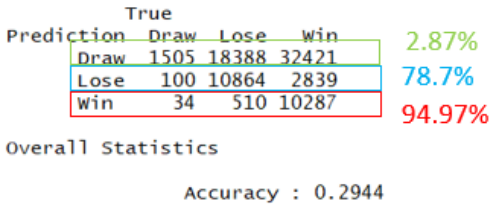


Figura 53 Resultados de probar el modelo con la octava semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 29.44%. La exactitud del modelo para la categoría “Draw” es de 2.87%, para la categoría “Lose” es de 78.7% y para la categoría “Win” es de 94.97%.

En cuanto a los datos acumulados hasta la octava semana, a continuación, se muestra la distribución de los datos hasta la octava semana de competición para cada una de las categorías.

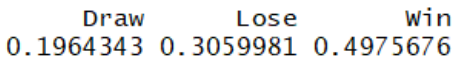


Figura 54 Distribución de los datos hasta la octava semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la octava semana de competición.

Figura 55 70 palabras más frecuentes en el conjunto de datos hasta la octava semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la octava semana de competición.

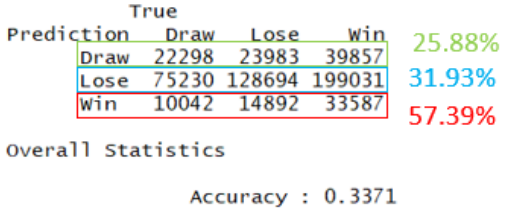


Figura 56 Resultados de probar el modelo hasta la octava semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 33.71%. La exactitud del modelo para la categoría “Draw” es de 25.88%, para la categoría “Lose” es de 31.93% y para la categoría “Win” es de 57.39%.

Novena Semana

Para analizar la novena semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la novena semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la distribución de los datos de la novena semana de competición para cada una de las categorías.

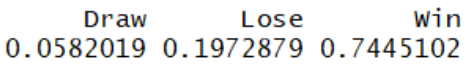


Figura 57 Distribución de los datos de la novena semana de competición según su categoría

Como se puede observar en la figura anterior, para la novena semana de competición, hay un 5.82 % de Tweets que se encuentran en la categoría “Draw”, 19.72% de Tweets que se encuentran en la categoría “Lose” y 74.45% de Tweets que se encuentran en la categoría “Win”.

Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la novena semana de competición.

Figura 58 70 palabras más frecuentes en el conjunto de datos de la novena semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la novena semana de competición.

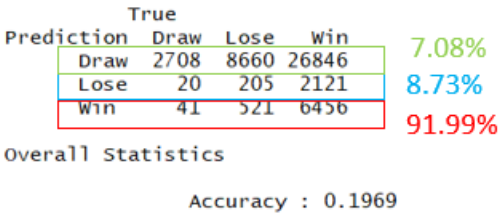


Figura 59 Resultados de probar el modelo con la novena semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 19.69%. La exactitud del modelo para la categoría “Draw” es de 7.08%, para la categoría “Lose” es de 8.73% y para la categoría “Win” es de 91.99%.

En cuanto a los datos acumulados hasta la novena semana, a continuación, se muestra la distribución de los datos hasta la novena semana de competición para cada una de las categorías.

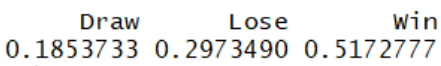


Figura 60 Distribución de los datos hasta la novena semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la novena semana de competición.

Figura 61 70 palabras más frecuentes en el conjunto de datos hasta la novena semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la novena semana de competición.

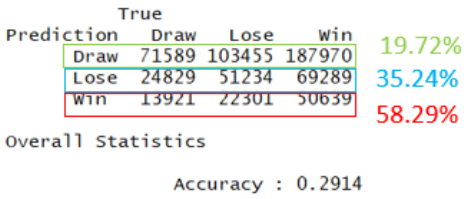


Figura 62 Resultados de probar el modelo hasta la novena semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 29.14%. La exactitud del modelo para la categoría “Draw” es de 19.72%, para la categoría “Lose” es de 35.24% y para la categoría “Win” es de 58.29%.

Décima Semana

Para analizar la décima semana de competición, primero se cogieron los datos de los partidos correspondientes solo a la décima semana y se observó la distribución de los datos en cada una de las categorías. A continuación, se muestra la

distribución de los datos de la décima semana de competición para cada una de las categorías.

```
     Draw      Lose       Win
0.4838730 0.1390494 0.3770776
```

Figura 63 Distribución de los datos de la décima semana de competición según su categoría

Como se puede observar en la figura anterior, para la décima semana de competición, hay un 48.38 % de Tweets que se encuentran en la categoría “Draw”, 13.9% de Tweets que se encuentran en la categoría “Lose” y 37.7% de Tweets que se encuentran en la categoría “Win”.

Adicionalmente, se calcularon las 70 palabras más frecuentes dentro del conjunto de datos. A continuación, se muestra la nube con las 70 palabras más frecuentes de la décima semana de competición.

Figura 64 70 palabras más frecuentes en el conjunto de datos de la décima semana de competición.

A continuación, se muestra el resultado de entrenar y probar el modelo para la décima semana de competición.

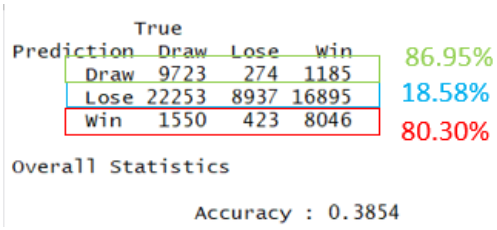


Figura 65 Resultados de probar el modelo con la décima semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 38.54%. La exactitud del modelo para la categoría “Draw” es de 86.95%, para la categoría “Lose” es de 18.58% y para la categoría “Win” es de 80.30%.

En cuanto a los datos acumulados hasta la décima semana, a continuación, se muestra la distribución de los datos hasta la décima semana de competición para cada una de las categorías.

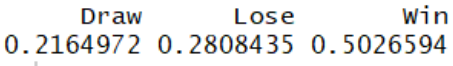


Figura 66 Distribución de los datos hasta la décima semana de competición según su categoría

A continuación, se muestra la nube con las 70 palabras más frecuentes para los datos hasta la décima semana de competición.

Figura 67 70 palabras más frecuentes en el conjunto de datos hasta la décima semana de competición

A continuación, se muestra el resultado de entrenar y probar el modelo hasta la décima semana de competición.

	True			
Prediction	Draw	Lose	Win	
Draw	93215	107685	196122	23.47%
Lose	27641	49748	65902	34.71%
Win	23010	29192	72001	57.97%

Overall Statistics

Accuracy : 0.3235

Figura 68 Resultados de probar el modelo hasta la décima semana de competición

Como se puede observar en la figura anterior, la exactitud del modelo es de 32.35%. La exactitud del modelo para la categoría “Draw” es de 23.47%, para la categoría “Lose” es de 34.71% y para la categoría “Win” es de 57.97%.

Resumen de resultados

De manera de conocer cuál es la influencia de la distribución de los datos alrededor de las tres categorías en la exactitud de los modelos, se construyó una gráfica de la exactitud del modelo versus el porcentaje de datos por cada una de las categorías para cada una de las semanas de análisis y las semanas acumuladas.

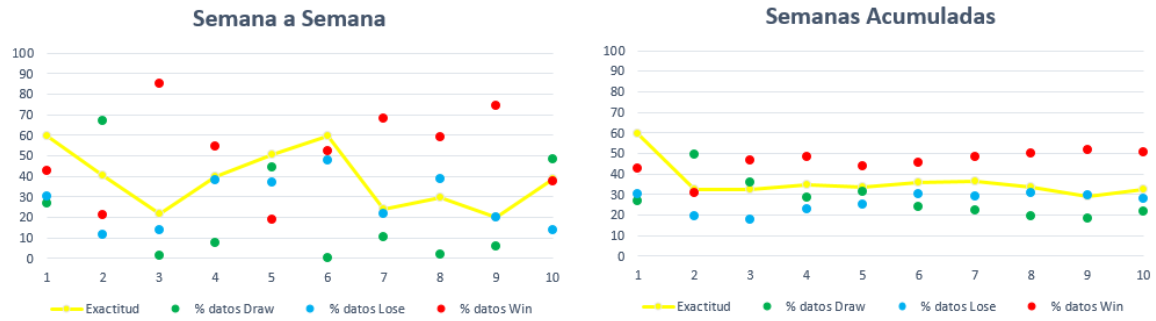


Figura 69 Exactitud del modelo vs. Porcentaje de datos por cada una de las categorías

Como se puede observar en la figura anterior, existe una relación entre la distribución de datos entre cada una de las categorías y la exactitud del modelo. En el caso del análisis de semana a semana, para la semana 1 se puede observar que el porcentaje de datos para las tres categorías estaba repartido casi uniformemente entre las categorías y cómo la exactitud del modelo arrojó casi un 60%. Un caso parecido se puede observar para la semana 5 en donde la distribución de los datos entre las categorías también se encuentra distribuido de manera uniforme y en donde la exactitud del modelo arrojó un 50%. Casos contrarios se pueden observar

para las semanas 3, 7 y 9 donde la distribución de los datos no es equitativa entre las categorías y donde la exactitud del modelo llega al 20% aproximadamente en los tres casos. Es importante resaltar que la semana 6 es un caso especial, ya que para esta semana no existieron datos para la clase “Draw”.

En el caso del análisis de los datos de las semanas acumuladas, se puede observar que, a mayor diferencia entre el porcentaje de datos para cada una de las categorías, menor es la exactitud del modelo.

Como complemento del análisis, se construyó una gráfica de la exactitud del modelo versus el máximo porcentaje de datos entre las tres categorías.

Figura 70 Exactitud del modelo vs. Máximo porcentaje de datos entre las categorías

Como se puede ver en la figura anterior, para ambos tipos de análisis, se puede observar que existe una relación inversa entre el máximo porcentaje de datos entre las tres categorías y la exactitud del modelo. De hecho, se puede observar que para un porcentaje máximo de datos menor al 50%, la exactitud del modelo supera el 50% (Semana 1 y Semana 5).

Adicionalmente, de manera de ver si el porcentaje de datos para cada categoría influye en la exactitud del modelo para esa categoría, se analizaron las siguientes figuras.

Figura 71 Exactitud del modelo para la categoría “Draw” vs. Porcentaje de datos para la categoría “Draw”

Figura 72 Exactitud del modelo para la categoría “Lose” vs. Porcentaje de datos para la categoría “Lose”

Figura 73 Exactitud del modelo para la categoría “Win” vs. Porcentaje de datos para la categoría “Win”

Como se puede observar en las tres figuras anteriores, existe una relación directa entre el porcentaje de datos de una categoría y la exactitud que tiene el modelo para esa misma categoría. Luego, se construyó la siguiente figura para conocer la relación entre la exactitud del modelo y la exactitud para cada una de las categorías.

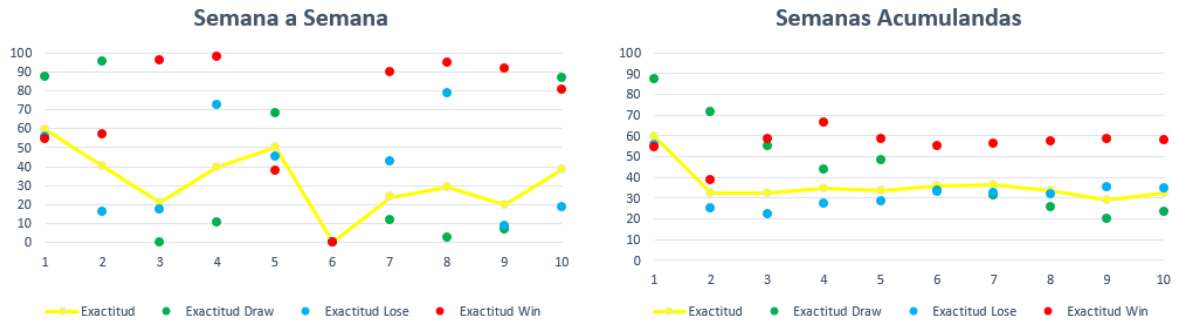


Figura 74 Exactitud del modelo vs. Exactitud del modelo para cada una de las categorías

Como se pueden ver en la figura anterior, la exactitud del modelo está directamente relacionada con la exactitud del modelo para cada una de las categorías. Para el análisis semana a semana se pude observar que, para la mayoría de las semanas, el modelo es muy bueno clasificando datos que pertenecen a la categoría “Win”. Sin embargo, el modelo no es muy bueno para clasificar datos de las otras dos categorías, sobre todo para los datos de la categoría “Draw”, donde a partir de la semana 3 la cantidad de datos para esta categoría no se incrementa en la misma medida que los demás datos de las otras dos categorías.

En cuanto al análisis de semanas acumuladas, al igual que para el análisis semana a semana, se puede observar que la categoría que mejor se puede clasificar es la categoría “Win”. En general, también se puede observar que la exactitud del modelo se ve afectada por la exactitud del modelo para la categoría “Draw” y esta, a su vez, se ve afectada por la poca velocidad de incremento de datos de esa categoría con respecto a las demás. Un escenario parecido ocurre con la categoría “Lose”. Durante las primeras 6 semanas, se puede observar que la categoría con menor cantidad de datos es la categoría “Lose”. A medida que transcurren las semanas, se puede observar que la cantidad de datos para la categoría “Lose” empieza a incrementar, mientras que la cantidad de datos para la categoría “Draw” empieza a disminuir hasta llegar a punto de encuentro en la semana 6 donde la categoría con menos cantidad de datos es la categoría “Draw”. Ni la categoría “Lose” ni la categoría “Draw” crecen en cantidad de datos a la misma velocidad que lo hace la categoría “Win”. Por esta razón, el modelo obtiene mucha más exactitud para clasificar datos pertenecientes a la categoría “Win”.

En definitiva, se puede decir que la exactitud del modelo se ve afectada por la poca exactitud del modelo para clasificar datos que pertenecen a las categorías “Draw” y “Lose”, y esta se ve afecta por la poca cantidad de datos que existen en ambas categorías durante las 10 semanas de estudio.

Por otro lado, se experimentó el impacto que puede tener la cantidad de palabras dentro del “Bag of Words” para entrenar el modelo. En principio, este estudio tomó en consideración las 70 palabras más frecuentes en el conjunto de datos. Sin embargo, para evaluar el impacto que puede tener el incremento en la cantidad de palabras del “Bag of Words”, se tomó en consideración las 117 palabras más frecuentes para los datos acumulados hasta la semana 10. A continuación, se muestran los resultados de un modelo que toma en consideración las 117 palabras más frecuentes para los datos acumulados hasta la semana 10.

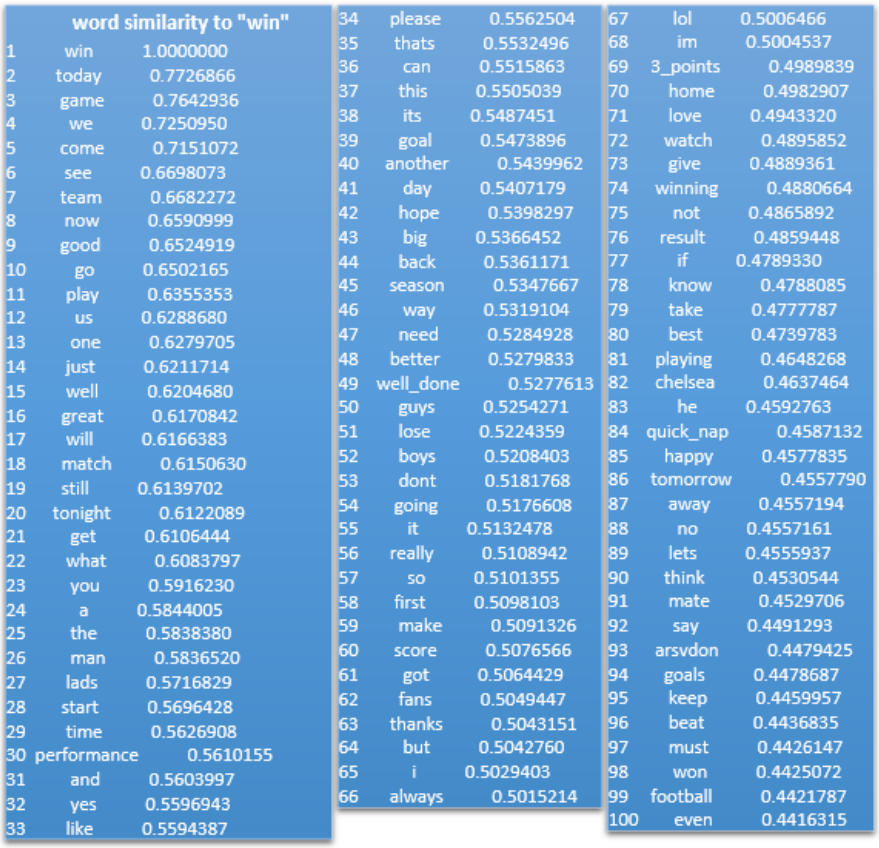
```
            True
Prediction   Draw   Lose    Win
      Draw  28526  24735  37888
      Lose  34624  59425  79717
      Win   80716 102465 216420

Overall Statistics

                Accuracy : 0.458
```

Figura 75 Resultados de probar el modelo hasta la décima semana de competición con 117 palabras más frecuentes

Como se puede observar en la figura anterior, el modelo obtiene una exactitud de 45.8%, incrementando así la exactitud en unos 13 puntos porcentuales (al compararlo con el 32.35% del modelo que tomó en consideración las 70 palabras más frecuentes del conjunto de datos).

De manera de explorar metodologías que pudieran mejorar el modelo, se tomó en consideración la utilización de Word2Vec para coger palabras del “Bag of Words” que ayudaran a mejorar el modelo. Word2Vec es un método que ayuda a conocer palabras que estén más relacionadas con el contexto. Para este trabajo final de máster, se utilizó el Word2Vec para conocer cuáles son las palabras que están relacionadas con las palabras “win”, “draw” y “lose” dentro del conjunto de datos de

semanas acumuladas hasta la semana 10. A continuación, se muestran las palabras más relacionadas con las categorías “Win”, “Draw” y “Lose”.

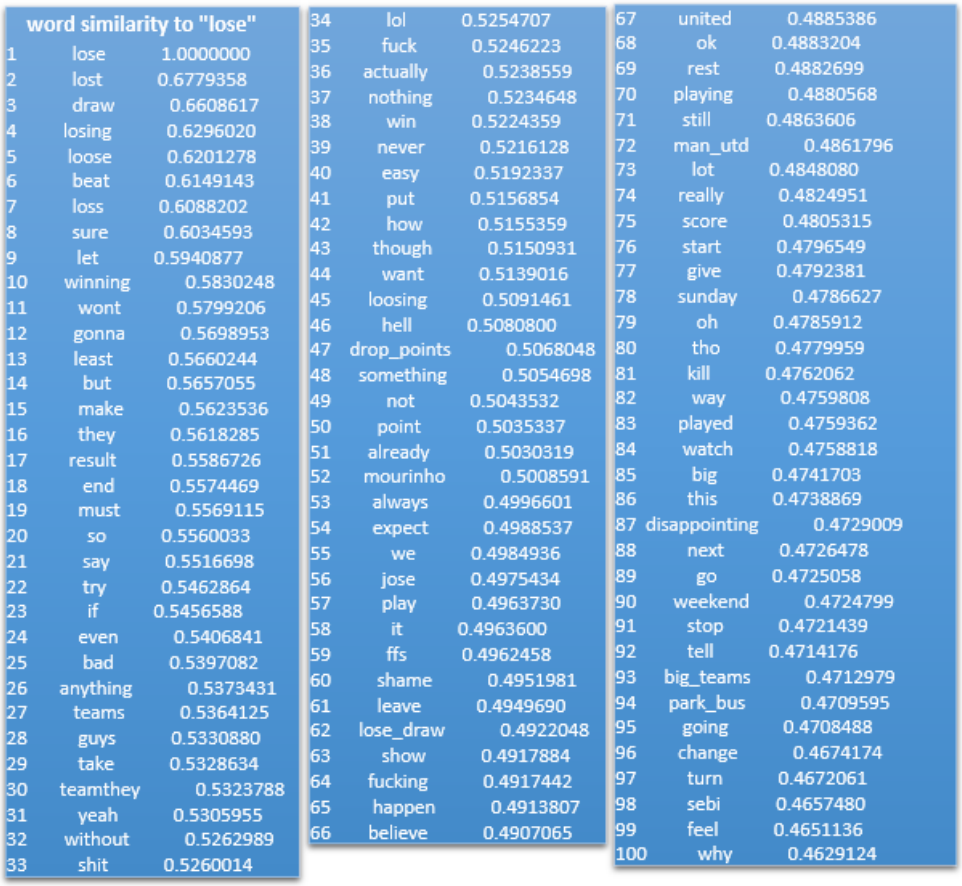


word similarity to "win"		
1	win	1.0000000
2	today	0.7726866
3	game	0.7642936
4	we	0.7250950
5	come	0.7151072
6	see	0.6698073
7	team	0.6682272
8	now	0.6590999
9	good	0.6524919
10	go	0.6502165
11	play	0.6355353
12	us	0.6288680
13	one	0.6279705
14	just	0.6211714
15	well	0.6204680
16	great	0.6170842
17	will	0.6166383
18	match	0.6150630
19	still	0.6139702
20	tonight	0.6122089
21	get	0.6106444
22	what	0.6083797
23	you	0.5916230
24	a	0.5844005
25	the	0.5838380
26	man	0.5836520
27	lads	0.5716829
28	start	0.5696428
29	time	0.5626908
30	performance	0.5610155
31	and	0.5603997
32	yes	0.5596943
33	like	0.5594387
34	please	0.5562504
35	thats	0.5532496
36	can	0.5515863
37	this	0.5505039
38	its	0.5487451
39	goal	0.5473896
40	another	0.5439962
41	day	0.5407179
42	hope	0.5398297
43	big	0.5366452
44	back	0.5361171
45	season	0.5347667
46	way	0.5319104
47	need	0.5284928
48	better	0.5279833
49	well_done	0.5277613
50	guys	0.5254271
51	lose	0.5224359
52	boys	0.5208403
53	dont	0.5181768
54	going	0.5176608
55	it	0.5132478
56	really	0.5108942
57	so	0.5101355
58	first	0.5098103
59	make	0.5091326
60	score	0.5076566
61	got	0.5064429
62	fans	0.5049447
63	thanks	0.5043151
64	but	0.5042760
65	i	0.5029403
66	always	0.5015214
67	lol	0.5006466
68	im	0.5004537
69	3_points	0.4989839
70	home	0.4982907
71	love	0.4943320
72	watch	0.4895852
73	give	0.4889361
74	winning	0.4880664
75	not	0.4865892
76	result	0.4859448
77	if	0.4789330
78	know	0.4788085
79	take	0.4777787
80	best	0.4739783
81	playing	0.4648268
82	chelsea	0.4637464
83	he	0.4592763
84	quick_nap	0.4587132
85	happy	0.4577835
86	tomorrow	0.4557790
87	away	0.4557194
88	no	0.4557161
89	lets	0.4555937
90	think	0.4530544
91	mate	0.4529706
92	say	0.4491293
93	arsvdon	0.4479425
94	goals	0.4478687
95	keep	0.4459957
96	beat	0.4436835
97	must	0.4426147
98	won	0.4425072
99	football	0.4421787
100	even	0.4416315

Figura 76 Palabras relacionadas con la categoría “Win” utilizando Word2Vec

	word similarity to "draw"	
1	draw	1.0000000
2	point	0.6840737
3	lose	0.6608617
4	result	0.6520249
5	easy	0.5511304
6	lucky	0.5446678
7	beat	0.5382437
8	fair_result	0.5345872
9	lost	0.5315900
10	loss	0.5246682
11	end	0.5013578
12	big_teams	0.5010319
13	mourinho	0.5003067
14	total_fix	0.4914623
15	teams	0.4887046
16	tho	0.4879721
17	disappointing	0.4854352
18	disappointed	0.4834402
19	home	0.4817684
20	poor	0.4817359
21	second_half	0.4798913
22	kept_apart	0.4774491
23	so	0.4755917
24	jose	0.4736030
25	least	0.4722716
26	not	0.4716836
27	side	0.4710265
28	draws	0.4708722
29	away	0.4695928
30	they	0.4694976
31	boring	0.4687779
32	performance	0.4622919
33	drawing_33	0.4609235
34	dropping_points	0.4607331
35	goalless_draw	0.4604145
36	park_bus	0.4599338
37	drawing	0.4583855
38	deserved	0.4580039
39	man_utd	0.4543019
40	drawn	0.4540561
41	utd	0.4538111
42	acquire_credibility	0.4529608
43	dropped_points	0.4491013
44	battling_display	0.4490893
45	ill_take	0.4490521
46	losing	0.4487616
47	3_points	0.4468341
48	poor_performance	0.4463693
49	take	0.4461860
50	though	0.4453180
51	actually	0.4444791
52	sure	0.4443565
53	bcfctweets_vs	0.4438611
54	winning	0.4436975
55	but	0.4416463
56	bristol_city	0.4414627
57	drop_points	0.4406124
58	points	0.4387539
59	draw_bcfctweets	0.4378564
60	gonna	0.4370322
61	drew	0.4340488
62	point_gained	0.4331716
63	bristol	0.4299819
64	2nd_half	0.4284575
65	it	0.4279083
66	anfield	0.4277309
67	united	0.4269911
68	expected	0.4263721
69	short_animation	0.4263143
70	00_draw	0.4260243
71	didnt	0.4244539
72	parking_bus	0.4233040
73	didnt_deserve	0.4201423
74	victory	0.4200811
75	last_season	0.4192480
76	very_disappointed	0.4187201
77	shame	0.4159624
78	pre_recorded	0.4145406
79	first_half	0.4138386
80	should	0.4123567
81	how	0.4116865
82	nothing	0.4091623
83	fixed	0.4083577
84	teamthey	0.4078307
85	points_dropped	0.4053468
86	that	0.4046299
87	spoils	0.4039601
88	poor_display	0.4034640
89	anything	0.4029270
90	man_city	0.4022105
91	bristolcity	0.4021586
92	fifth_round	0.4020553
93	u18s_kieran	0.4011427
94	defeat	0.4008223
95	carabaocup	0.4006051
96	liverpool	0.4002166
97	mourinhio	0.3978593
98	dull	0.3971977
99	bcfctweets	0.3971869
100	mourinho_coward	0.3967913

Figura 77 Palabras relacionadas con la categoría “Draw” utilizando Word2Vec



word similarity to "lose"		
1	lose	1.0000000
2	lost	0.6779358
3	draw	0.6608617
4	losing	0.6296020
5	loose	0.6201278
6	beat	0.6149143
7	loss	0.6088202
8	sure	0.6034593
9	let	0.5940877
10	winning	0.5830248
11	wont	0.5799206
12	gonna	0.5698953
13	least	0.5660244
14	but	0.5657055
15	make	0.5623536
16	they	0.5618285
17	result	0.5586726
18	end	0.5574469
19	must	0.5569115
20	so	0.5560033
21	say	0.5516698
22	try	0.5462864
23	if	0.5456588
24	even	0.5406841
25	bad	0.5397082
26	anything	0.5373431
27	teams	0.5364125
28	guys	0.5330880
29	take	0.5328634
30	teamthey	0.5323788
31	yeah	0.5305955
32	without	0.5262989
33	shit	0.5260014
34	lol	0.5254707
35	fuck	0.5246223
36	actually	0.5238559
37	nothing	0.5234648
38	win	0.5224359
39	never	0.5216128
40	easy	0.5192337
41	put	0.5156854
42	how	0.5155359
43	though	0.5150931
44	want	0.5139016
45	loosing	0.5091461
46	hell	0.5080800
47	drop_points	0.5068048
48	something	0.5054698
49	not	0.5043532
50	point	0.5035337
51	already	0.5030319
52	mourinho	0.5008591
53	always	0.4996601
54	expect	0.4988537
55	we	0.4984936
56	jose	0.4975434
57	play	0.4963730
58	it	0.4963600
59	ffs	0.4962458
60	shame	0.4951981
61	leave	0.4949690
62	lose_draw	0.4922048
63	show	0.4917884
64	fucking	0.4917442
65	happen	0.4913807
66	believe	0.4907065
67	united	0.4885386
68	ok	0.4883204
69	rest	0.4882699
70	playing	0.4880568
71	still	0.4863606
72	man_utd	0.4861796
73	lot	0.4848080
74	really	0.4824951
75	score	0.4805315
76	start	0.4796549
77	give	0.4792381
78	sunday	0.4786627
79	oh	0.4785912
80	tho	0.4779959
81	kill	0.4762062
82	way	0.4759808
83	played	0.4759362
84	watch	0.4758818
85	big	0.4741703
86	this	0.4738869
87	disappointing	0.4729009
88	next	0.4726478
89	go	0.4725058
90	weekend	0.4724799
91	stop	0.4721439
92	tell	0.4714176
93	big_teams	0.4712979
94	park_bus	0.4709595
95	going	0.4708488
96	change	0.4674174
97	turn	0.4672061
98	sebi	0.4657480
99	feel	0.4651136
100	why	0.4629124

Figura 78 Palabras relacionadas con la categoría “Lose” utilizando Word2Vec

Como se puede ver en las tres figuras anteriores, por cada una de las categorías existe una lista de palabras acompañadas de un número. Este número representa el nivel de relacionamiento que tiene esa palabra con la categoría. A medida que el número se acerca al valor uno, se encuentra más fuertemente relacionada con la categoría.

Para visualizar mejor y entender si realmente estas nuevas palabras pueden diferenciar entre cada una de las categorías, se realizó un análisis entre las palabras de las categorías dos a dos.

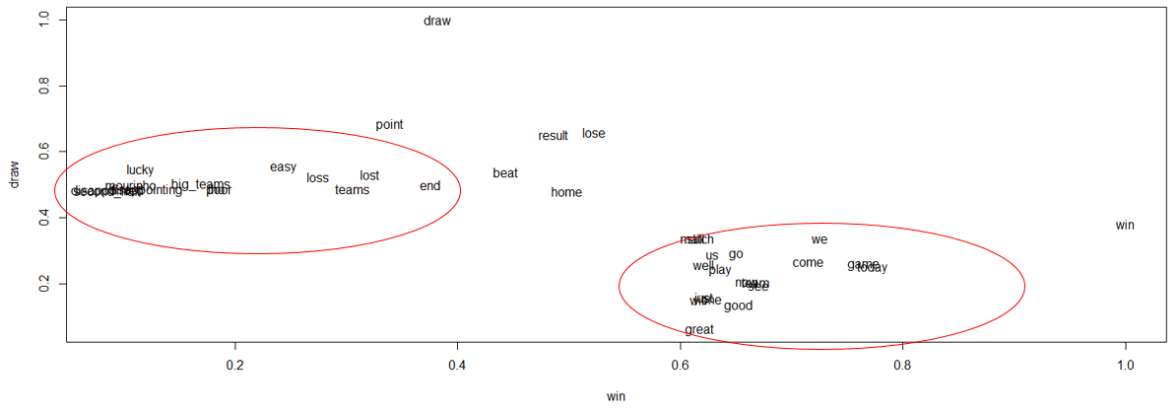


Figura 79 Relación entre palabras de la categoría Draw y la categoría Win

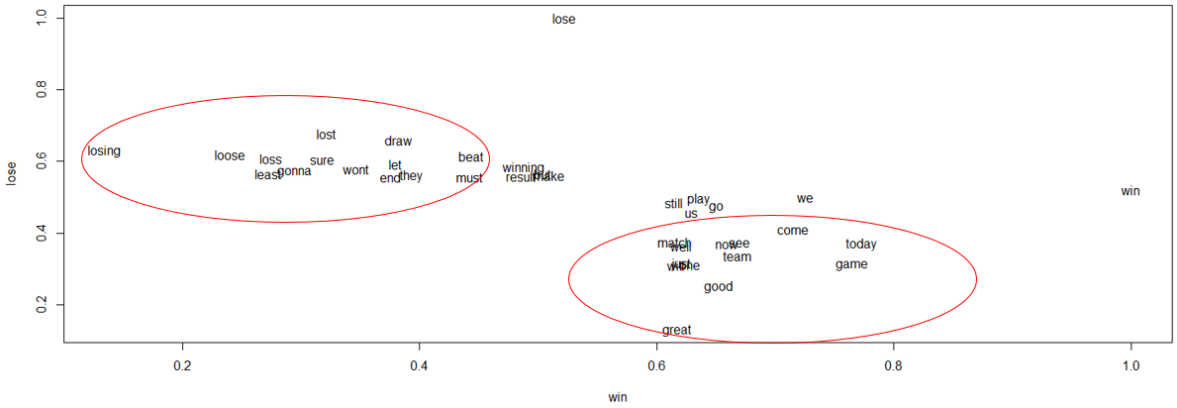


Figura 80 Relación entre palabras de la categoría Lose y la categoría Win

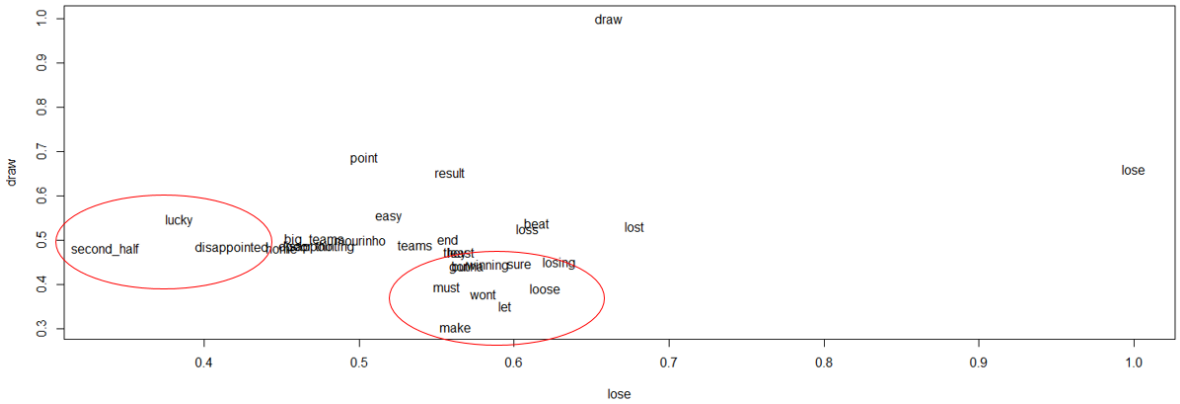


Figura 81 Relación entre palabras de la categoría Draw y la categoría Lose

Como se puede observar en las tres figuran anteriores, el conjunto de palabras entre las categorías “Draw” y “Win” y entre las categorías “Lose” y “Win” se encuentran

suficientemente espaciadas entre ellas como para concluir que la categoría “Win” puede ser diferenciada con el resto de categorías. Sin embargo, el conjunto de palabras entre las categorías “Draw” y “Lose” se encuentran muy cerca entre ellas. Esta cercanía puede influir en el modelo y puede perjudicarlo al momento de poder diferencias Tweets entre estas dos categorías.

Luego, con estas nuevas palabras, se creó un nuevo “Bag of Words” y se entrenó el modelo con el conjunto de datos pertenecientes a las semanas acumuladas hasta la semana 10. A continuación, se muestra el resultado de probar el modelo con el nuevo “Bag of Words” y con semanas acumuladas hasta la semana 10.

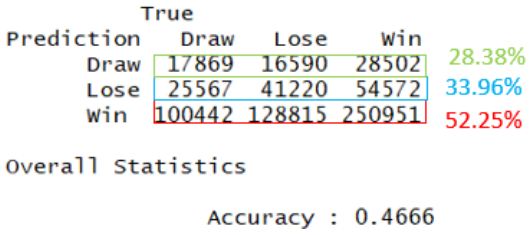


Figura 82 Resultados de probar el modelo con un nuevo “Bag of Words” y con un conjunto de datos de semanas acumuladas hasta la semana 10.

Como se puede observar en la figura anterior, la mejor categoría que logra clasificar el modelo es la categoría “Win” con un 52.25%. En general, el modelo alcanza una exactitud del 46.66% que, en comparación con el modelo anterior que considera un “Bag of Words” con las categorías más frecuentes y el mismo conjunto de datos (obtuvo una exactitud del 45.8%), tiene una diferencia en exactitud de poco menos del 1%. En líneas generales, el método no causa una mejoría sustancial en la exactitud del modelo.

4.2.3 Probar el modelo con resultados de partidos individuales

De manera de continuar probando el modelo, se cogió el modelo de las semanas acumuladas hasta la semana 10 y se probaron con datos (Tweets) que pertenecieran a partidos individuales. Se cogió la semana 10 con 117 palabras en

el “Bag of Words” (exactitud del 45.8%) ya que este modelo es el que más exactitud posee entre todos los modelos de las semanas acumuladas que se calcularon.

Para probar el modelo, se tomaron las semanas 9 y 10 de análisis. A continuación, se muestran los resultados.

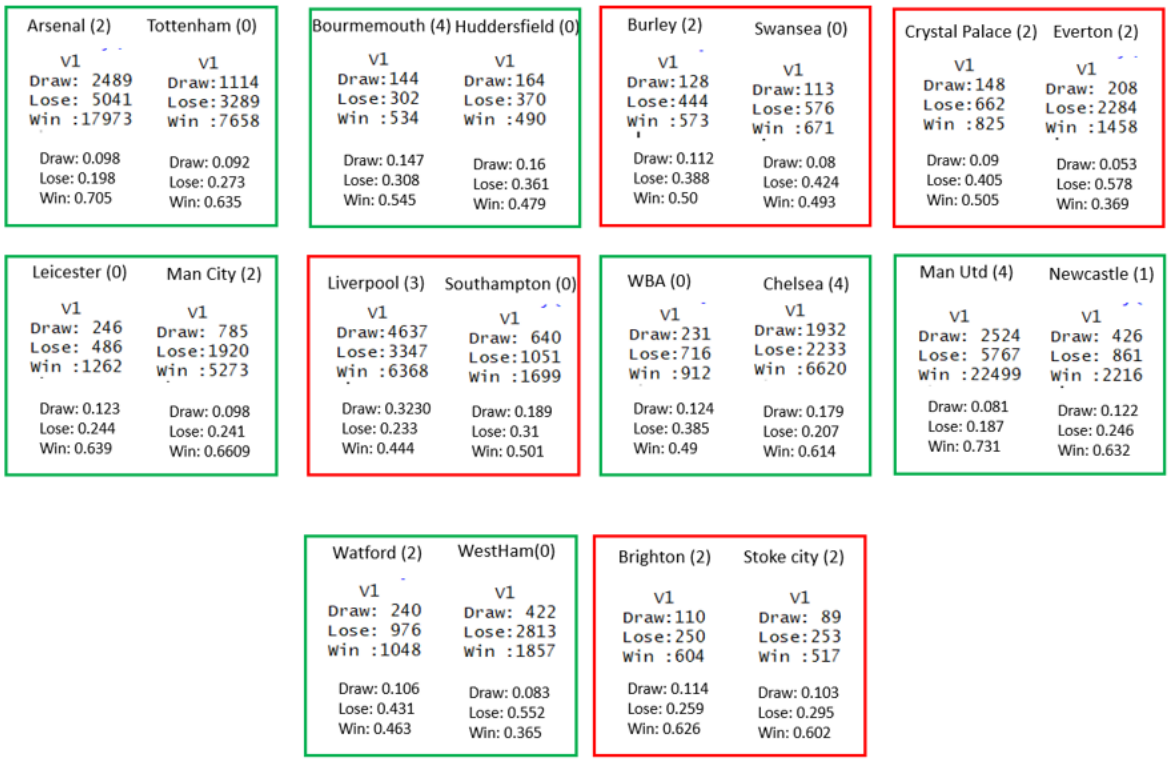


Figura 83 Prueba partido a partido del modelo de semanas acumuladas hasta la semana 9 con 117 palabras

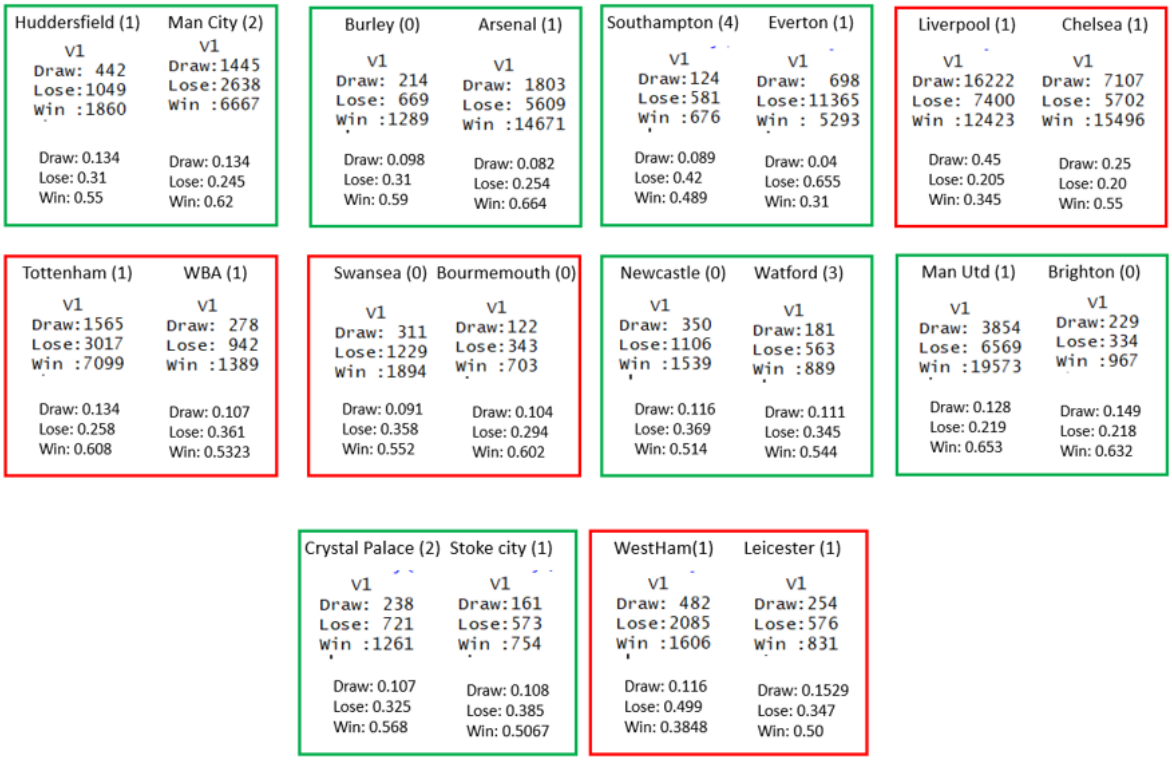


Figura 84 Prueba partido a partido del modelo de semanas acumuladas hasta la semana 10 con 117 palabras

Como se puede observar en las dos figuras anteriores, los recuadros que se encuentran en color verde representan los partidos en los que el resultado del partido coincidió con la tendencia de las opiniones en Twitter, mientras que los recuadros de color rojo representan los partidos en los que el resultado del partido no coincidió con la tendencia de las opiniones en Twitter.

Es importante resaltar que, de los 20 partidos analizados, el modelo pudo acertar en 12 ocasiones y de las 8 ocasiones en las que no acertó, 6 de ellos fueron empates. Tomando en consideración los resultados presentados en la sección anterior, estos resultados coinciden con el hecho de que el modelo posee inconvenientes al clasificar Tweets que se encuentran en la categoría “Draw” y “Lose”.

4.3 Objetivo No.3: Crear un modelo estadístico basados en los resultados de partidos anteriores que tome en consideración una regresión lineal generalizada.

4.3.1 Análisis de la información de partidos anteriores

Antes de aplicar cualquier modelo, se debe analizar la información para corroborar a qué modelo se puede ajustar mejor. Adicionalmente, antes de aplicar el modelo, se sugiere que se analice si los equipos de casa son más propensos o no a anotar goles en comparación a los equipos que juegan fuera de casa. Para ello, se debe calcular el promedio de goles anotados en casa y fuera de casa. A continuación, se muestra el promedio de goles anotados en casa y fuera de casa.

	avg_home_goals	avg_away_goals
1	1.547619	1.136054

Figura 85 Promedio de goles anotados en casa y fuera de casa

Como se puede observar en la figura anterior, en promedio, el equipo local marca más goles que el equipo visitante. Esta es la llamada “ventaja de jugar en casa” (Gallagher, 2017) y no es específica para el fútbol. Este es un momento conveniente para recordar la distribución de Poisson. Poisson, es una distribución de probabilidad discreta que describe la probabilidad del número de eventos dentro de un período de tiempo específico (por ejemplo, 90 minutos) con una tasa promedio conocida de ocurrencia. Una suposición clave es que la cantidad de eventos es independiente del tiempo. En el contexto del presenta trabajo final de máster, esto significa que los goles no se vuelven más o menos probables por el número de goles anotados en el partido. En cambio, el número de goles se expresa simplemente como función de una tasa promedio de goles.

La siguiente figura muestra la proporción de goles anotados en comparación con el número de goles estimados por las distribuciones de Poisson correspondientes.

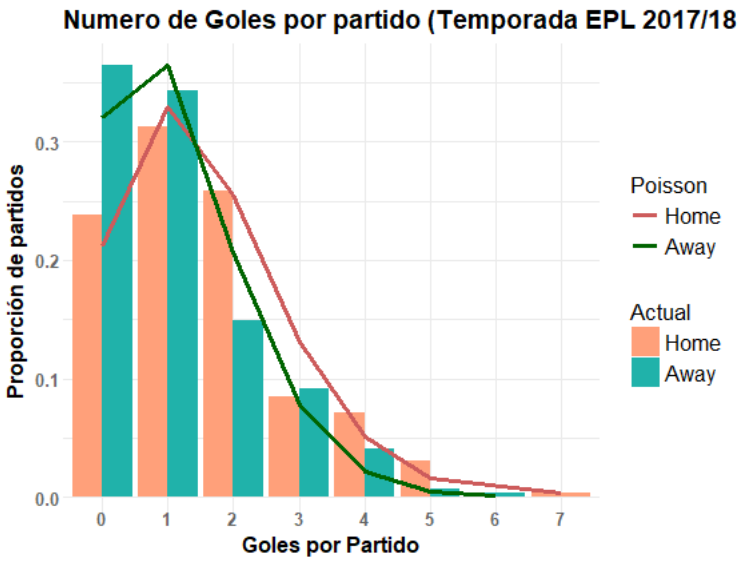


Figura 86 Comparación entre goles anotados y goles estimados por la distribución de Poisson

Como se puede observar en la figura anterior, la cantidad de goles anotados por el equipo local y visitante se pueden aproximar como distribuciones de Poisson.

Analizando este resultado, se puede decir que es un resultado esperado ya que es mucho más probable que los equipos anoten una cantidad de goles entre 0 y 2 goles que anotar 3 o más goles. Por esta razón, se intentará realizar una regresión lineal generalizada que permita modelar este comportamiento para poder predecir la cantidad de goles que pueden anotar en partido específicos y poder predecir su resultado.

4.3.2 Aplicar un modelo de regresión lineal generalizada

Para el momento de la elaboración del presente trabajo final de máster, la API que se utilizó para la obtención de los resultados de los partidos, arrojó los resultados correspondientes a 350 partidos hasta el 09/05/2018.

Antes de elaborar el modelo de regresión lineal generalizada, se decidió crear un conjunto de datos para crear el modelo y otro para probarlo. Para este caso, el conjunto de datos para crear el modelo no tomó en consideración los últimos 40 partidos para la fecha mencionada.

Luego, se construyó el modelo de regresión lineal generalizada de Poisson con los datos depurados de partidos de resultados de partidos anteriores. A continuación, se muestran los resultados del modelo.

```
Call:
glm(formula = goals ~ home + team + opponent, family = poisson(link = log),
    data = .)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.48802  -1.09361  -0.09366   0.53897   2.39833

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.47447    0.21377   2.220 0.026451 *
home                        0.28541    0.07055   4.045 5.22e-05 ***
teamBournemouth            -0.45626    0.21384  -2.134 0.032869 *
teamBrighton & Hove Albion -0.73601    0.23315  -3.157 0.001595 **
teamBurnley                -0.73592    0.23577  -3.121 0.001800 **
teamC Palace               -0.62062    0.22824  -2.719 0.006546 **
teamChelsea                -0.09813    0.19431  -0.505 0.613530
teamEverton                -0.43137    0.21361  -2.019 0.043438 *
teamHuddersfield Town      -0.83270    0.24201  -3.441 0.000580 ***
teamLeicester              -0.21307    0.20152  -1.057 0.290375
teamLiverpool               0.23995    0.18011   1.332 0.182772
teamMan City                0.39238    0.17422   2.252 0.024310 *
teamMan Utd                 0.02980    0.18929   0.157 0.874897
teamNewcastle              -0.61946    0.22779  -2.719 0.006539 **
teamSouthampton            -0.66509    0.23024  -2.889 0.003869 **
teamStoke                  -0.68554    0.23012  -2.979 0.002891 **
teamSwansea                -0.80843    0.24224  -3.337 0.000846 ***
teamTottenham               0.02726    0.18867   0.144 0.885136
teamWatford                -0.39136    0.21053  -1.859 0.063040 .
teamWest Brom              -0.90839    0.24567  -3.698 0.000218 ***
teamWest Ham               -0.41837    0.21489  -1.947 0.051546 .
opponentBournemouth         0.14647    0.21279   0.688 0.491242
opponentBrighton & Hove Albion  0.02368    0.22339   0.106 0.915596
opponentBurnley            -0.48809    0.25127  -1.942 0.052078 .
opponentC Palace            0.09685    0.21379   0.453 0.650551
opponentChelsea            -0.42130    0.24883  -1.693 0.090428 .
opponentEverton             0.16526    0.21182   0.780 0.435271
opponentHuddersfield Town   0.22375    0.20977   1.067 0.286139
opponentLeicester           0.03690    0.21887   0.169 0.866114
opponentLiverpool          -0.19988    0.23339  -0.856 0.391775
opponentMan City           -0.66119    0.27387  -2.414 0.015766 *
opponentMan Utd            -0.53706    0.26126  -2.056 0.039817 *
opponentNewcastle          -0.02552    0.22297  -0.114 0.908885
opponentSouthampton         0.09969    0.21781   0.458 0.647159
opponentStoke               0.32678    0.20484   1.595 0.110645
opponentSwansea             0.02827    0.22070   0.128 0.898078
opponentTottenham          -0.46599    0.25493  -1.828 0.067556 .
opponentWatford             0.24984    0.20731   1.205 0.228143
opponentWest Brom           0.13000    0.21263   0.611 0.540920
opponentWest Ham            0.37346    0.20538   1.818 0.069000 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 834.18  on 607  degrees of freedom
Residual deviance: 627.82  on 568  degrees of freedom
AIC: 1755.6

Number of Fisher Scoring iterations: 5
```

Figura 87 Modelo de regresión lineal generalizada de Poisson

Primero que todo, se debe analizar el coeficiente “Home” con una estimación de 0.28541. Este valor representa el hecho de que los equipos que juegan en casa generalmente anotan más goles que los equipos visitantes (específicamente, $e^{0.28541} = 1.33$ más probable). Pero no todos los equipos se comportan de la misma

manera. Por ejemplo, el Manchester City posee una estimación de 0.39238 mientras que el West Bromwich posee una estimación de -0.90839. Esto, de cierta forma, quiere decir que el Manchester City es más propenso a anotar goles en comparación al promedio. En cambio, el West Bromwich es mucho menos propenso a anotar goles.

Por otro lado, los valores de los "oponentes" penalizan o premian a los equipos en base en la calidad de su oposición. Por ejemplo, el Manchester City posee un valor de -0.6619 y el West Bromwich posee un valor de 0.13. Esto quiere decir que, los equipos son menos propensos a anotar un gol cuando juegan en contra del Manchester City, mientras que cuando juegan en contra del West Bromwich, los equipos tienen más oportunidad de anotar goles.

4.3.3 Cálculo de la exactitud del modelo y análisis de resultados

De manera calcular la exactitud del modelo, se utilizarán los 40 últimos partidos que fueron sacados del conjunto de datos para crear el modelo. A continuación, se muestran los resultados de los primeros 10 partidos utilizando el modelo de regresión que fue creado en el punto anterior.

Crystal Palace vs. Liverpool

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.0410926462	0.0924890717	0.1040846620	0.0780893473	0.0439398038	0.0197794595
2	0.0386756182	0.0870489577	0.0979625071	0.0734962106	0.0413553088	0.0186160516
3	0.0182003786	0.0409644129	0.0461002256	0.0345866188	0.0194614156	0.0087605371
4	0.0057099502	0.0128516425	0.0144628856	0.0108507562	0.0061055716	0.0027484170
5	0.0013435242	0.0030239304	0.0034030483	0.0025531314	0.0014366120	0.0006466895
6	0.0002528999	0.0005692132	0.0006405769	0.0004805918	0.0002704224	0.0001217304

Figura 88 Matriz de probabilidades de anotar de 1 a 5 goles – Crystal Palace vs. Liverpool

Esta matriz simplemente muestra las probabilidades de que Crystal Palace (filas de la matriz) y el Liverpool (columnas de la matriz) anoten un número específico de

goles. Por ejemplo, a lo largo de la diagonal, ambos equipos obtienen el mismo número de goles (por ejemplo, P (0-0) = 0.041). Por lo tanto, se puede calcular las probabilidades de sorteo al sumar todas las entradas diagonales. Todo debajo de la diagonal representa una victoria de Crystal Palace (por ejemplo, P (3-0) = 0.018), mientras que puedes estimar P (más de 2.5 goles) al sumar todas las entradas excepto los cuatro valores en la esquina superior izquierda. Afortunadamente, podemos usar funciones básicas de manipulación de matrices para realizar estos cálculos. A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.1434022
>
> #Empate
> sum(diag(teamsVs))
[1] 0.1866509
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.6421217
```

Figura 89 Probabilidades de Ganar, Empatar y Perder – Crystal Palace vs. Liverpool

Como se puede observar en la figura anterior, el equipo visitante (el Liverpool) tiene 0.64 probabilidades de ganar el partido, mientras que el equipo de casa (el Crystal Palace) posee solo 0.14 probabilidades de ganar y ambos poseen un 0.19 de empatar.

West Ham United vs. Southampton

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.063597267	0.076356001	0.04583718	0.01834431	0.0055061254	0.0013221502
2	0.098866221	0.118700530	0.07125698	0.02851746	0.0085596415	0.0020553713
3	0.076847089	0.092263971	0.05538688	0.02216615	0.0066532687	0.0015976064
4	0.039821320	0.047810179	0.02870087	0.01148626	0.0034476510	0.0008278621
5	0.015476229	0.018581033	0.01115436	0.00446404	0.0013399012	0.0003217418
6	0.004811767	0.005777093	0.00346804	0.00138793	0.0004165933	0.0001000338

Figura 90 Matriz de probabilidades de anotar de 1 a 5 goles – West Ham United vs. Southampton

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.4498467
>
> #Empate
> sum(diag(teamsVs))
[1] 0.2506109
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.2927695
```

Figura 91 Probabilidades de Ganar, Empatar y Perder – West Ham United vs. Southampton

Como se puede observar en la figura anterior, el equipo visitante (el Southampton) tiene 0.29 probabilidades de ganar el partido, mientras que el equipo de casa (el West Ham United) posee un 0.45 probabilidades de ganar y ambos poseen un 0.25 de empatar.

Brighton & Hove Albion vs. Leicester City

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.091399870	0.121548774	0.0808212555	0.0358269092	0.0119111682	3.168031e-03
2	0.097126440	0.129164295	0.0858850333	0.0380716097	0.0126574510	3.366520e-03
3	0.051605902	0.068628480	0.0456330386	0.0202284749	0.0067252457	1.788723e-03
4	0.018279740	0.024309444	0.0161640442	0.0071652901	0.0023822032	6.335980e-04
5	0.004856260	0.006458132	0.0042941967	0.0019035561	0.0006328644	1.683239e-04
6	0.001032105	0.001372552	0.0009126491	0.0004045643	0.0001345032	3.577401e-05

Figura 92 Matriz de probabilidades de anotar de 1 a 5 goles – Brighton & Hove Albion vs. Leicester City

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.2974826
>
> #Empate
> sum(diag(teamsVs))
[1] 0.2740311
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.4251833
```

Figura 93 Probabilidades de Ganar, Empatar y Perder – Brighton & Hove Albion vs. Leicester City

Como se puede observar en la figura anterior, el equipo visitante (el Leicester City) tiene 0.42 probabilidades ganar el partido, mientras que el equipo de casa (el Brighton & Hove Albion) posee un 0.29 probabilidades de ganar y ambos poseen un 0.27 de empatar.

Manchester United vs. Swansea City

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.06826444	0.02857021	0.005978640	0.0008340653	8.726879e-05	7.304793e-06
2	0.15467655	0.06473563	0.013546663	0.0018898615	1.977374e-04	1.655152e-05
3	0.17523644	0.07334041	0.015347310	0.0021410655	2.240211e-04	1.875158e-05
4	0.13235280	0.05539264	0.011591536	0.0016171066	1.691989e-04	1.416271e-05
5	0.07497269	0.03137777	0.006566152	0.0009160277	9.584457e-05	8.022624e-06
6	0.03397528	0.01421942	0.002975575	0.0004151151	4.343377e-05	3.635603e-06

Figura 94 Matriz de probabilidades de anotar de 1 a 5 goles – Manchester United vs. Swansea City

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.7680518
>
> #Empate
> sum(diag(teamsVs))
[1] 0.150064
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.05370352
```

Figura 95 Probabilidades de Ganar, Empatar y Perder – Manchester United vs. Swansea City

Como se puede observar en la figura anterior, el equipo visitante (el Swansea City) tiene 0.053 probabilidades ganar el partido, mientras que el equipo de casa (el Manchester United) posee un 0.77 probabilidades de ganar y ambos poseen un 0.15 de empatar.

Newcastle United vs. Huddersfield Town

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.119958052	0.081727751	0.027840671	0.006322642	1.076908e-03	1.467401e-04
2	0.172656873	0.117631603	0.040071368	0.009100244	1.550005e-03	2.112046e-04
3	0.124253417	0.084654195	0.028837568	0.006549038	1.115469e-03	1.519945e-04
4	0.059613079	0.040614555	0.013835404	0.003142033	5.351689e-04	7.292241e-05
5	0.021450431	0.014614238	0.004978360	0.001130590	1.925686e-04	2.623950e-05
6	0.006174766	0.004206885	0.001433081	0.000325454	5.543318e-05	7.553356e-06

Figura 96 Matriz de probabilidades de anotar de 1 a 5 goles – Newcastle United vs. Huddersfield Town

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.5499968
>
> #Empate
> sum(diag(teamsVs))
[1] 0.2697694
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.1764984
```

Figura 97 Probabilidades de Ganar, Empatar y Perder – Newcastle United vs. Huddersfield Town

Como se puede observar en la figura anterior, el equipo visitante (el Huddersfield Town) tiene 0.18 probabilidades ganar el partido, mientras que el equipo de casa (el Newcastle United) posee un 0.54 probabilidades de ganar y ambos poseen un 0.27 de empatar.

Watford FC vs. AFC Bournemouth

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.05074041	0.066338517	0.043365822	0.018898971	0.0061771762	0.0016152205
2	0.08492029	0.111025646	0.072578023	0.031629746	0.0103382621	0.0027032696
3	0.07106226	0.092907519	0.060734113	0.026468130	0.0086511749	0.0022621266
4	0.03964380	0.051830710	0.033881996	0.014765887	0.0048262675	0.0012619821
5	0.01658719	0.021686263	0.014176419	0.006178131	0.0020193377	0.0005280205
6	0.00555214	0.007258924	0.004745195	0.002067972	0.0006759219	0.0001767414

Figura 98 Matriz de probabilidades de anotar de 1 a 5 goles – Watford FC vs. AFC Bournemouth

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.4531747
>
> #Empate
> sum(diag(teamsVs))
[1] 0.2394621
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.2976427
```

Figura 99 Probabilidades de Ganar, Empatar y Perder – Watford FC vs. AFC Bournemouth

Como se puede observar en la figura anterior, el equipo visitante (el Bournemouth) tiene 0.3 probabilidades ganar el partido, mientras que el equipo de casa (el Watford) posee un 0.45 probabilidades de ganar y ambos poseen un 0.24 de empatar.

West Bromwich Albion vs. Burnley FC

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	2.451448e-01	2.149478e-01	9.423523e-02	2.754244e-02	6.037438e-03	1.058749e-03
2	1.297028e-01	1.137260e-01	4.985859e-02	1.457233e-02	3.194327e-03	5.601697e-04
3	3.431200e-02	3.008544e-02	1.318975e-02	3.855011e-03	8.450375e-04	1.481891e-04
4	6.051338e-03	5.305932e-03	2.326173e-03	6.798781e-04	1.490326e-04	2.613495e-05
5	8.004203e-04	7.018243e-04	3.076867e-04	8.992858e-05	1.971279e-05	3.456912e-06
6	8.469831e-05	7.426514e-05	3.255857e-05	9.515999e-06	2.085954e-06	3.658011e-07

Figura 100 Matriz de probabilidades de anotar de 1 a 5 goles – West Bromwich Albion vs. Burnley FC

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.2098867
>
> #Empate
> sum(diag(teamsVs))
[1] 0.3727605
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.4170339
```

Figura 101 Probabilidades de Ganar, Empatar y Perder – West Bromwich Albion vs. Burnley FC

Como se puede observar en la figura anterior, el equipo visitante (el Burnley FC) tiene 0.42 probabilidades ganar el partido, mientras que el equipo de casa (el West Bromwich Albion) posee un 0.20 probabilidades de ganar y ambos poseen un 0.37 de empatar.

Everton vs. Manchester City

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	2.948257e-02	0.0827565093	0.116147276	0.1086737445	0.0762608250	4.281228e-02
2	2.113877e-02	0.0593357559	0.083276669	0.0779182065	0.0546784022	3.069606e-02
3	7.578163e-03	0.0212716314	0.029854353	0.0279333657	0.0196019887	1.100442e-02
4	1.811161e-03	0.0050838633	0.007135111	0.0066760000	0.0046848231	2.630026e-03
5	3.246469e-04	0.0009112723	0.001278954	0.0011966597	0.0008397452	4.714269e-04
6	4.655385e-05	0.0001306750	0.000183400	0.0001715991	0.0001204181	6.760187e-05

Figura 102 Matriz de probabilidades de anotar de 1 a 5 goles – Everton vs. Manchester City

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.06838287
>
> #Empate
> sum(diag(teamsVs))
[1] 0.126256
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.739546
```

Figura 103 Probabilidades de Ganar, Empatar y Perder – Everton vs. Manchester City

Como se puede observar en la figura anterior, el equipo visitante (el Manchester City) tiene 0.74 probabilidades ganar el partido, mientras que el equipo de casa (el Everton) posee un 0.06 probabilidades de ganar y ambos poseen un 0.12 de empatar.

Arsenal vs. Stoke City

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.02295858	0.01858988	0.007526238	0.002031366	0.0004112061	6.659183e-05
2	0.06805728	0.05510690	0.022310402	0.006021679	0.0012189587	1.974015e-04
3	0.10087278	0.08167806	0.033067918	0.008925181	0.0018067100	2.925835e-04
4	0.09967407	0.08070745	0.032674959	0.008819119	0.0017852401	2.891066e-04
5	0.07386720	0.05981128	0.024215001	0.006535738	0.0013230190	2.142533e-04
6	0.04379365	0.03546031	0.014356347	0.003874843	0.0007843782	1.270243e-04

Figura 104 Matriz de probabilidades de anotar de 1 a 5 goles – Arsenal vs. Stoke City

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.7263633
>
> #Empate
> sum(diag(teamsVs))
[1] 0.1214026
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.0716868
```

Figura 105 Probabilidades de Ganar, Empatar y Perder – Arsenal vs. Stoke City

Como se puede observar en la figura anterior, el equipo visitante (el Stoke City) tiene 0.072 probabilidades ganar el partido, mientras que el equipo de casa (el Arsenal) posee un 0.73 probabilidades de ganar y ambos poseen un 0.12 de empatar.

Chelsea FC vs. Tottenham Hotspur

Se realizó el cálculo de las posibilidades que tenía cada equipo de anotar de 1 a 5 goles y se obtuvo una matriz que probabilidades de cada uno de los dos equipos.

	V1	V2	V3	V4	V5	V6
1	0.100262825	0.108658728	0.058878847	0.0212697664	0.0057627185	1.249056e-03
2	0.121941789	0.132153065	0.071609711	0.0258687440	0.0070087413	1.519129e-03
3	0.074154104	0.080363689	0.043546630	0.0157310594	0.0042620904	9.237987e-04
4	0.030062601	0.032580012	0.017654114	0.0063774834	0.0017278818	3.745146e-04
5	0.009140694	0.009906126	0.005367827	0.0019391079	0.0005253717	1.138731e-04
6	0.002223422	0.002409608	0.001305693	0.0004716769	0.0001277937	2.769899e-05

Figura 106 Matriz de probabilidades de anotar de 1 a 5 goles – Chelsea FC vs. Tottenham Hotspur

A continuación, se muestran las probabilidades de que gane el equipo de casa, de que empaten y de que gane el equipo visitante.

```
> #Gane home
> sum(teamsVs[lower.tri(teamsVs)])
[1] 0.3896483
>
> #Empate
> sum(diag(teamsVs))
[1] 0.2828931
>
> #Gane Away
> sum(teamsVs[upper.tri(teamsVs)])
[1] 0.3249587
```

Figura 107 Probabilidades de Ganar, Empatar y Perder – Chelsea FC vs. Tottenham Hotspur

Como se puede observar en la figura anterior, el equipo visitante (el Tottenham Hotspur) tiene 0.32 probabilidades ganar el partido, mientras que el equipo de casa (el Chelsea FC) posee un 0.39 probabilidades de ganar y ambos poseen un 0.28 de empatar.

En las siguientes tablas se pueden observar todas las predicciones hechas por el modelo junto con el resultado real de cada partido para los últimos 40 partidos desde la fecha de obtención de los datos (fecha 32, 33, 34 y 35 de la competición).

Tabla 3 Predicciones del modelo de regresión lineal vs. Resultados de cada partido – fecha 32 de la competición

	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	Resultado real
1	Crystal Palace	Liverpool	0.14	0.64	0.19	2
2	West Ham United	Southampton	0.45	0.29	0.25	1
3	Brighton & Hove Albion	Leicester City	0.29	0.42	0.27	2
4	Manchester United	Swansea City	0.77	0.0.53	0.15	1
5	Newcastle United	Huddersfield Town	0.54	0.18	0.27	1
6	Watford FC	AFC Bournemouth	0.45	0.3	0.24	0
7	West Bromwich Albion	Burnley FC	0.2	0.42	0.37	2
8	Everton	Manchester City	0.06	0.74	0.12	2
9	Arsenal	Stoke City	0.73	0.072	0.12	1
10	Chelsea FC	Tottenham Hotspur	0.39	0.32	0.28	2

2	Gana visita
1	Gana casa
0	Empatan

Al igual que para la fecha 32, se utilizó el mismo modelo para calcular las predicciones de los siguientes 30 partidos (fecha 33, fecha 34 y fecha 35). A continuación, se muestran los resultados para estas fechas.

Tabla 4 Predicciones del modelo de regresión lineal vs. Resultados de cada partido – fecha 33 de la competición

	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	Resultado real
1	Everton	Liverpool	0.16	0.617	0.18	0
2	AFC Bournemouth	Crystal Palace	0.48	0.25	0.26	0
3	Brighton & Hove Albion	Huddersfield Town	0.48	0.22	0.3	0
4	Leicester City	Newcastle United	0.55	0.19	0.24	2
5	Stoke City	Tottenham Hotspur	0.084	0.72	0.16	2
6	Watford FC	Burnley FC	0.31	0.38	0.31	2
7	West Bromwich Albion	Swansea City	0.35	0.34	0.31	0
8	Manchester City	Manchester United	0.59	0.17	0.22	2
9	Arsenal	Southampton	0.7	0.15	0.16	1
10	Chelsea FC	West Ham United	0.77	0.08	0.14	0

2	gana visita
1	gana casa
0	empatan

Tabla 5 Predicciones del modelo de regresión lineal vs. Resultados de cada partido – fecha 34 de la competición

	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	Resultado real
1	Southampton	Chelsea FC	0.17	0.59	0.24	2
2	Burnley FC	Leicester City	0.41	0.27	0.31	1
3	Crystal Palace	Brighton & Hove Albion	0.45	0.3	0.25	1
4	Huddersfield Town	Watford FC	0.32	0.27	0.41	1
5	Swansea City	Everton	0.36	0.29	0.34	0
6	Liverpool	AFC Bournemouth	0.72	0.07	0.11	1
7	Tottenham Hotspur	Manchester City	0.3	0.44	0.26	2
8	Newcastle United	Arsenal	0.27	0.47	0.25	1
9	Manchester United	West Bromwich Albion	0.8	0.04	0.12	2
10	West Ham United	Stoke City	0.56	0.21	0.21	0

2	gana visita
1	gana casa
0	empatan

Tabla 6 Predicciones del modelo de regresión lineal vs. Resultados de cada partido – fecha 35 de la competición

	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	Resultado real
1	Brighton & Hove Albion	Tottenham Hotspur	0.12	0.64	0.22	0
2	AFC Bournemouth	Manchester United	0.15	0.63	0.21	2
3	Leicester City	Southampton	0.63	0.14	0.21	0
4	West Bromwich Albion	Liverpool	0.09	0.71	0.16	0
5	Watford FC	Crystal Palace	0.48	0.27	0.25	0
6	Arsenal	West Ham United	0.73	0.12	0.14	1
7	Stoke City	Burnley FC	0.21	0.46	0.32	0
8	Manchester City	Swansea City	0.8	0.02	0.18	1
9	Everton	Newcastle United	0.44	0.28	0.28	1
10	Chelsea FC	Huddersfield Town				-

2	gana visita
1	gana casa
0	empatan

Como parte de los resultados, de los 39 partidos analizados en las 4 fechas de competición, el modelo pudo acertar 20. Esto le otorga al modelo un porcentaje de exactitud del 51.2%.

Además, para la utilización de este modelo, se ha creado un umbral de confianza. El nivel de confianza para la utilización de este modelo consta en determinar cuál es el rango de probabilidades para el cual el modelo puede alcanzar un porcentaje de exactitud aceptable. Para lograr visualizar este umbral de confianza, a continuación, se muestra una figura donde se muestran tres histogramas para las tres categorías (gana casa, gana visita y empate) en relación a los rangos de probabilidades.

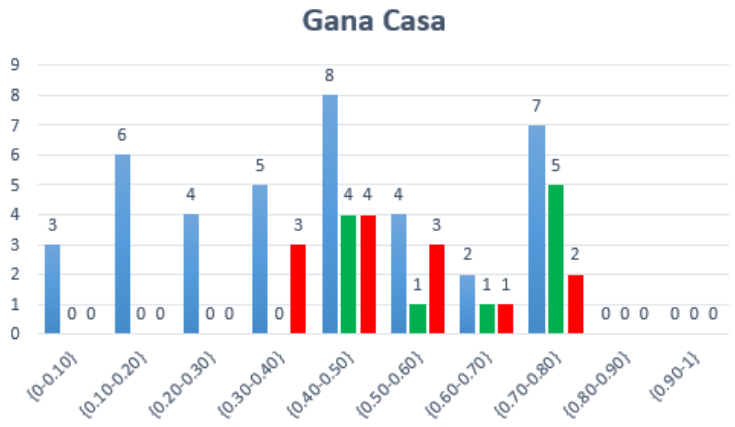


Figura 108 histograma de las probabilidades del modelo para la categoría Gana Casa. En azul la cantidad total de partido en las 4 fechas de competición (32, 33, 34 y 35). En rojo la cantidad de partidos que no fueron acertados por el modelo. En verde la cantidad de partidos que sí fueron acertados por el modelo.

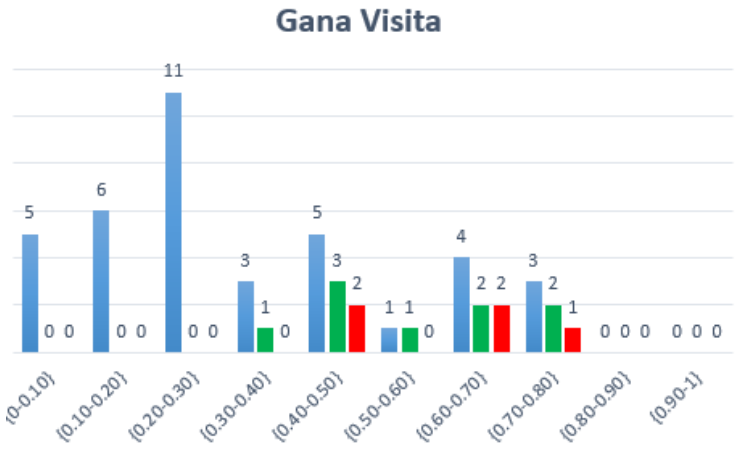


Figura 109 histograma de las probabilidades del modelo para la categoría Gana Visita. En azul la cantidad total de partido en las 4 fechas de competición (32, 33, 34 y 35). En rojo la cantidad de

partidos que no fueron acertados por el modelo. En verde la cantidad de partidos que sí fueron acertados por el modelo.

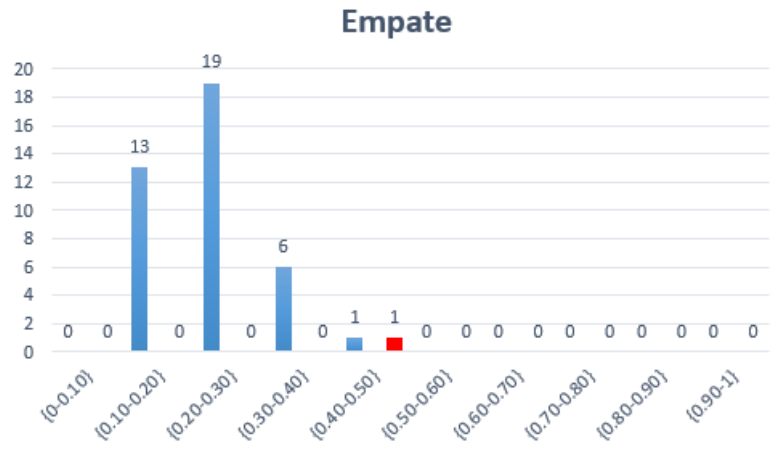


Figura 110 histograma de las probabilidades del modelo para la categoría Empate. En azul la cantidad total de partido en las 4 fechas de competición (32, 33, 34 y 35). En rojo la cantidad de partidos que no fueron acertados por el modelo. En verde la cantidad de partidos que sí fueron acertados por el modelo.

Con los resultados de las tres figuras anteriores se puede calcular el umbral de confianza para cada uno de los rangos de probabilidades del modelo. A continuación, se muestra una figura del umbral de confianza acumulado versus los rangos de probabilidades del modelo.

Figura 111 Umbral de confianza acumulado vs. Rangos de probabilidades del modelo.

Como se puede observar en la figura anterior, para obtener un umbral de exactitud del 70%, se deben considerar probabilidades de 0.7 o más.

4.4 Objetivo No.4: Comparar los resultados del modelo estadístico basado en resultados anteriores con los datos de casas apuestas y los resultados del modelo basado en opiniones en Twitter.

4.4.1 Tomar la exactitud del modelo basado en resultados anteriores y compararlo con las posibilidades (odds) de las casas de apuesta.

Una vez calculada la exactitud del modelo basado en resultados anteriores, se compararon con las posibilidades de las casas de apuestas. Las casas de apuestas publican esta información justo antes de cada partido para que los usuarios tengan conocimiento de las tendencias que posee la casa de apuestas para ese partido y, en base a eso, el beneficio en caso de ganar la apuesta.

Para este punto, se tomaron en consideración las casas de apuestas que se tomaron en consideración fueron Bet365 y Blue Square. Ambas manejan posibilidades decimales. Éstas se pueden aproximar como el inverso de las probabilidades. Esto hace que las posibilidades tengan una relación inversa con las probabilidades.

A continuación, se muestra una tabla para las fechas de competición 32, 33, 34 y 35 junto con las correspondientes posibilidades de las casas de apuestas.

Tabla 7 Comparación de probabilidades del modelo basado en resultados de partidos anteriores con las posibilidades de las casas de apuestas para la fecha 32.

			Probabilidades			Decimal Odds						
	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	B365H	B365A	B365D	BWH	BWA	BWD	Resultado real
1	Crystal Palace	Liverpool	0.14	0.64	0.19	7	1.44	4.33	6.75	1.5	4.33	2
2	West Ham United	Southampton	0.45	0.29	0.25	2.7	2.6	3.25	2.85	2.6	3.2	1
3	Brighton & Hove Albion	Leicester City	0.29	0.42	0.27	2.7	2.62	3.2	2.7	2.8	3.1	2
4	Manchester United	Swansea City	0.77	0.0.53	0.15	1.25	10	6	1.22	13	6.5	1
5	Newcastle United	Huddersfield Town	0.54	0.18	0.27	1.75	5	3.39	1.75	5.25	3.5	1
6	Watford FC	AFC Bournemouth	0.45	0.3	0.24	2.29	3.1	3.29	2.3	3.1	3.4	0
7	West Bromwich Albion	Burnley FC	0.2	0.42	0.37	2.5	3.1	3	2.55	3	3.1	2
8	Everton	Manchester City	0.06	0.74	0.12	9	1.33	5	8.5	1.36	5	2
9	Arsenal	Stoke City	0.73	0.072	0.12	1.28	10	5.5	1.3	9.75	5.5	1
10	Chelsea FC	Tottenham Hotspur	0.39	0.32	0.28	2.29	3.1	3.29	2.35	3.2	3.25	2

2	gana visita
1	gana casa
0	empatan

Tabla 8 Comparación de probabilidades del modelo basado en resultados de partidos anteriores con las posibilidades de las casas de apuestas para la fecha 33.

			Probabilidades			Decimal Odds						
	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	B365H	B365A	B365D	BWH	BWA	BWD	Resultado real
1	Everton	Liverpool	0.16	0.617	0.18	4.2	1.9	3.75	4.1	1.9	3.6	0
2	AFC Bournemouth	Crystal Palace	0.48	0.25	0.26	2.5	3	3.39	2.45	2.9	3.4	0
3	Brighton & Hove Albion	Huddersfield Town	0.48	0.22	0.3	1.85	5	3.5	1.85	4.75	3.4	0
4	Leicester City	Newcastle United	0.55	0.19	0.24	1.95	4.2	3.6	1.95	4.1	3.4	2
5	Stoke City	Tottenham Hotspur	0.084	0.72	0.16	11	1.33	5.5	9.5	1.33	5.25	2
6	Watford FC	Burnley FC	0.31	0.38	0.31	2.39	3.39	3.2	2.25	3.4	3.2	2
7	West Bromwich Albion	Swansea City	0.35	0.34	0.31	2.37	3.39	3.2	2.4	3.2	3.1	0
8	Manchester City	Manchester United	0.59	0.17	0.22	1.89	4.2	3.79	1.85	4.2	3.6	2
9	Arsenal	Southampton	0.7	0.15	0.16	1.57	6	4.5	1.53	6	4.4	1
10	Chelsea FC	West Ham United	0.77	0.08	0.14	1.28	12	6	1.28	11	5.75	0

2	gana visita
1	gana casa
0	empatan

Tabla 9 Comparación de probabilidades del modelo basado en resultados de partidos anteriores con las posibilidades de las casas de apuestas para la fecha 34.

			Probabilidades			Decimal Odds						
	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	B365H	B365A	B365D	BWH	BWA	BWD	Resultado real
1	Southampton	Chelsea FC	0.17	0.59	0.24	4.5	1.85	3.75	4.5	1.85	3.5	2
2	Burnley FC	Leicester City	0.41	0.27	0.31	2.62	3	3.2	2.65	2.85	3.1	1
3	Crystal Palace	Brighton & Hove Albion	0.45	0.3	0.25	1.9	4.75	3.5	1.9	4.33	3.4	1
4	Huddersfield Town	Watford FC	0.32	0.27	0.41	2.62	3	3.25	2.6	2.9	3.1	1
5	Swansea City	Everton	0.36	0.29	0.34	2.54	3.1	3.25	2.55	3	3.1	0
6	Liverpool	AFC Bournemouth	0.72	0.07	0.11	1.25	12	7	1.25	11.5	6.25	1
7	Tottenham Hotspur	Manchester City	0.3	0.44	0.26	2.62	2.7	3.6	2.6	2.6	3.6	2
8	Newcastle United	Arsenal	0.27	0.47	0.25	3.29	2.25	3.6	3.1	2.2	3.6	1
9	Manchester United	West Bromwich Albion	0.8	0.04	0.12	1.18	19	8	1.18	18.5	6.75	2
10	West Ham United	Stoke City	0.56	0.21	0.21	2	4	3.6	2	4	3.3	0

2	gana visita
1	gana casa
0	empatan

Tabla 10 Comparación de probabilidades del modelo basado en resultados de partidos anteriores con las posibilidades de las casas de apuestas para la fecha 34.

			Probabilidades			Decimal Odds						
	Equipo de Casa	Equipo Visitante	P Gana Casa	P Gana Visita	P Empate	B365H	B365A	B365D	BWH	BWA	BWD	Resultado real
1	Brighton & Hove Albion	Tottenham Hotspur	0.12	0.64	0.22	6.5	1.55	4.33	6	1.55	4.2	0
2	AFC Bournemouth	Manchester United	0.15	0.63	0.21	5	1.75	3.89	5	1.7	3.8	2
3	Leicester City	Southampton	0.63	0.14	0.21	2.25	3.2	3.29	2.3	3.1	3.4	0
4	West Bromwich Albion	Liverpool	0.09	0.71	0.16	7.5	1.5	4.5	6.75	1.5	4.25	0
5	Watford FC	Crystal Palace	0.48	0.27	0.25	2.89	2.54	3.39	2.95	2.5	3.2	0
6	Arsenal	West Ham United	0.73	0.12	0.14	1.55	6	4.59	1.53	5.75	4.5	1
7	Stoke City	Burnley FC	0.21	0.46	0.32	2.29	3.5	3.29	2.3	3.3	3.2	0
8	Manchester City	Swansea City	0.8	0.02	0.18	1.14	23	9	1.13	21	8.75	1
9	Everton	Newcastle United	0.44	0.28	0.28	2.29	3.39	3.39	2.25	3.4	3.2	1
10	Chelsea FC	Huddersfield Town				-	-	-	-	-	-	-

2	gana visita
1	gana casa
0	empatan

Como se puede observar en las tablas anteriores, cada una de las probabilidades del modelo y cada una de las posibilidades son comparada con el resultado real de cada partido. El color rojo significa que la predicción del modelo o de la casa de apuesta no acertó el resultado final, mientras que el color verde significa que sí lo acertaron. Recordando los resultados del objetivo anterior, el modelo basado en resultados anteriores acertó 20 de los 39 resultados, obteniendo así una exactitud de 51.2%. En el caso de la casa de apuestas Bet365, acertó 18 de los 39 resultados, obteniendo así una exactitud de 46.2%. Por último, la casa de apuestas Blue Square, acertó 16 de 38 resultados (uno de los partidos no fue considerado porque la casa de apuestas les otorgó las mismas posibilidades a dos categorías) obteniendo así una exactitud de 42.2%.

De manera de visualizar mejor los resultados del modelo en comparación con las posibilidades, se crearon las siguientes figuras. A continuación, se muestra la exactitud del modelo para la categoría “Gana Casa” versus las posibilidades de las casas de apuesta.

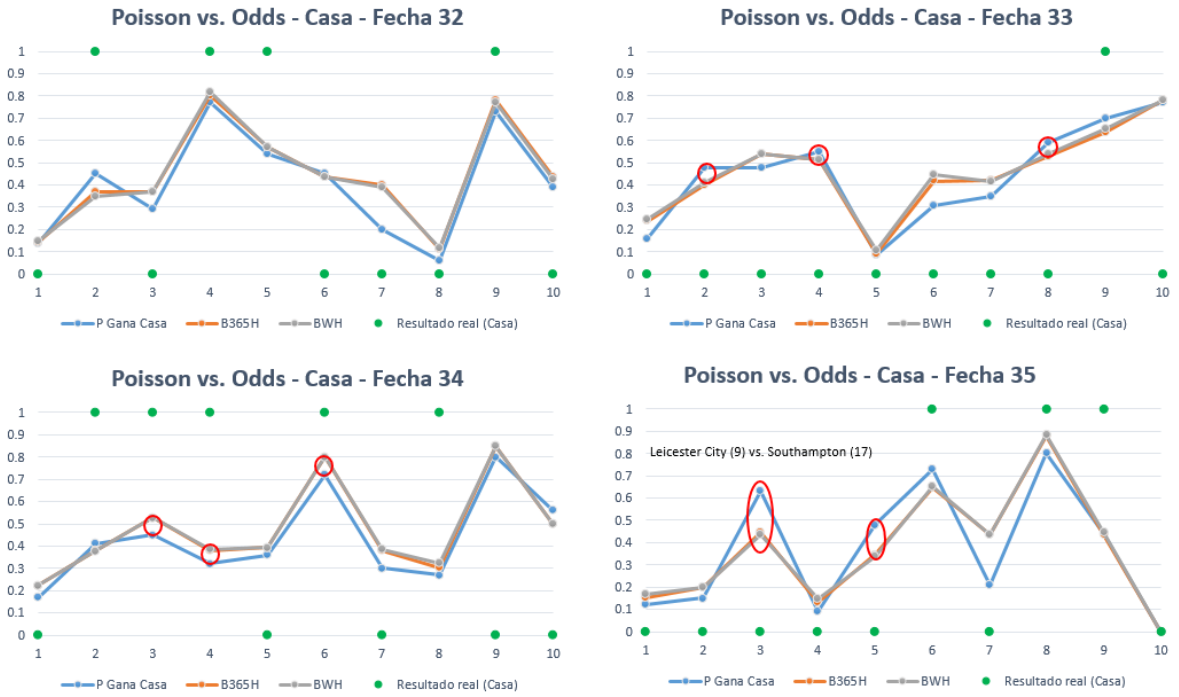


Figura 112 Exactitud del modelo para la categoría “Gana Casa” vs. Posibilidades de las casas de apuestas para las fechas de competición 32, 33, 34 y 35.

Como se puede observar en la figura anterior, la exactitud del modelo, en la mayoría de resultados para la categoría “Gana Casa”, arroja resultados más exactos en cada uno de los partidos. De los 40 partidos analizados, para la categoría “Gana Casa” el modelo basado en resultados anteriores ofrece una exactitud igual o mejor en 32 de los partidos. De los 8 partidos donde la exactitud ha sido menor que la obtenida con las posibilidades de las casas de apuestas, 6 han sido con una diferencia menor a 0.08, y el peor de los casos, correspondiente al partido 3 de la fecha 35, es un caso especial. Para este partido, se enfrentaba el Leicester City en casa y el Southampton de visita. El equipo de casa, para ese momento de la competición, ocupaba el puesto 9 en la tabla de competición, mientras que el equipo visitante, ocupaba el antepenúltimo puesto. Los resultados arrojaban que el equipo de casa tenía la ventaja, pero inesperadamente, el equipo visitante ganó el encuentro.

Siguiendo con el análisis comparativo, a continuación, se muestra la exactitud del modelo para la categoría “Gana Visita” versus las posibilidades de las casas de apuesta.

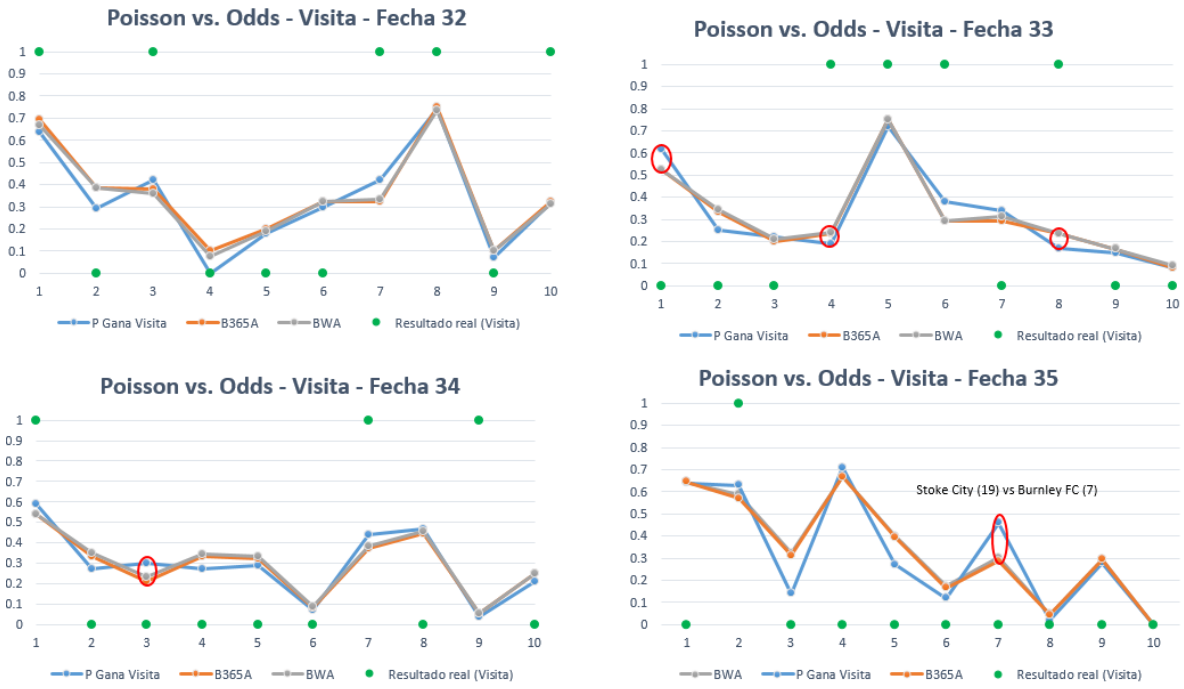


Figura 113 Exactitud del modelo para la categoría “Gana Visita” vs. Posibilidades de las casas de apuestas para las fechas de competición 32, 33, 34 y 35.

Como se puede observar en la figura anterior, la exactitud del modelo, en la mayoría de resultados para la categoría “Gana Visita”, arroja resultados más exactos en cada uno de los partidos. De los 40 partidos analizados, para la categoría “Gana Visita” el modelo basado en resultados anteriores ofrece una exactitud igual o mejor en 35 de los partidos. De los 5 partidos donde la exactitud ha sido menor que la obtenida con las posibilidades de las casas de apuestas, 4 han sido con una diferencia menor a 0.08, y el peor de los casos, correspondiente al partido 7 de la fecha 35, es un caso especial. Para este partido, se enfrentaba el Stoke City en casa y el Burnley FC de visita. El equipo de casa, para ese momento de la competición, ocupaba el puesto 19 en la tabla de competición, mientras que el equipo visitante, ocupaba el puesto 7. Los resultados arrojaban que el equipo visitante tenía la ventaja, pero inesperadamente, los dos equipos empataron el encuentro.

Para finalizar el análisis comparativo, a continuación, se muestra la exactitud del modelo para la categoría “Empate” versus las posibilidades de las casas de apuesta.

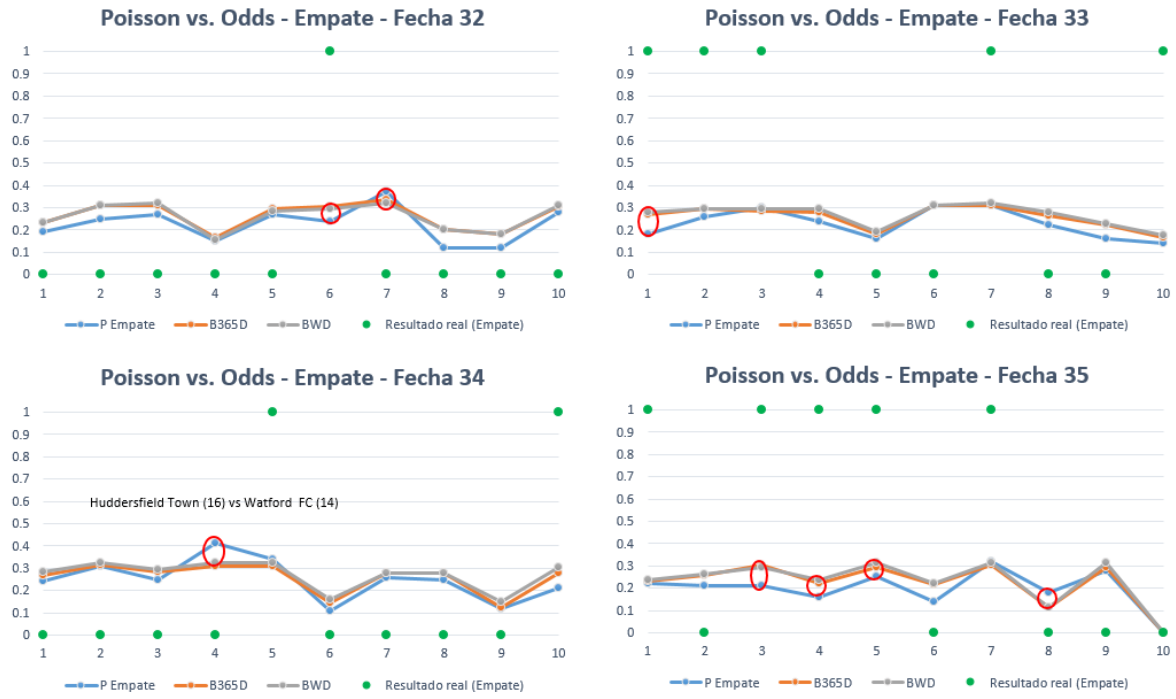


Figura 114 Exactitud del modelo para la categoría “Empate” vs. Posibilidades de las casas de apuestas para las fechas de competición 32, 33, 34 y 35.

Como se puede observar en la figura anterior, la exactitud del modelo, en la mayoría de resultados para la categoría “Empate”, arroja resultados más exactos en cada uno de los partidos. De los 40 partidos analizados, para la categoría “Empate” el modelo basado en resultados anteriores ofrece una exactitud igual o mejor en 33 de los partidos. De los 7 partidos donde la exactitud ha sido menor que la obtenida con las posibilidades de las casas de apuestas, 6 han sido con una diferencia menor a 0.08, y el peor de los casos, correspondiente al partido 4 de la fecha 34, es un caso especial. Para este partido, se enfrentaba el Huddersfield Town en casa y el Watford FC de visita. El equipo de casa, para ese momento de la competición, ocupaba el puesto 16 en la tabla de competición, mientras que el equipo visitante, ocupaba el puesto 14. Los resultados arrojaban que ambos equipos empatarían el encuentro, pero inesperadamente, el equipo de casa ganó el encuentro.

Finalmente, se puede observar que el modelo basado en resultados anteriores puede ser sensibles a resultados especiales inesperados. En el siguiente objetivo

de este trabajo final de máster, se analiza una de las posibles causas de la ocurrencia de este tipo de casos.

4.4.2 Tomar la exactitud del modelo basado en opiniones en twitter y compáralo con el modelo basado en resultados anteriores y los datos de las casas de apuestas.

Para comprar los resultados del modelo basado en opiniones en Twitter con el modelo basado en resultados anteriores y las posibilidades de las casas de apuestas, se cogió el modelo ya entrenado de 10 semanas acumuladas y se realizaron pruebas partido a partido con los Tweets para las fechas de competición 32, 33, 34 y 35. Para visualizar los resultados, se cogieron las gráficas realizadas en el apartado anterior como punto de partida y se incluyeron los resultados del modelo basado en opiniones de Twitter. A continuación, se muestran los resultados para las fechas de competición 32, 33, 34 y 35 para la categoría donde el equipo de casa sale ganador.

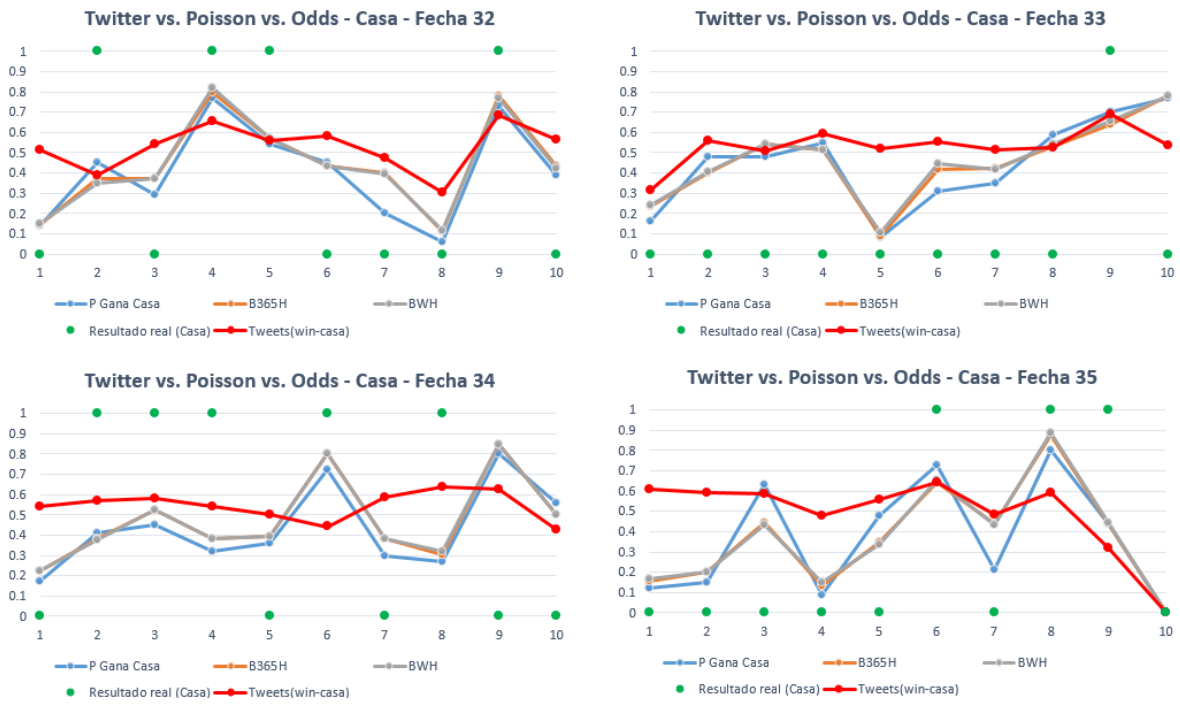


Figura 115 Exactitud del modelo basado Tweets para la categoría “Win-Casa” vs. Exactitud del modelo para la categoría “Gana Casa” vs. Posibilidades de las casas de apuestas - Fechas de competición 32, 33, 34 y 35.

Como se puede observar en la figura anterior, los resultados del modelo basado en opiniones en Twitter se asemejan mucho los resultados obtenidos con el modelo basado en resultados anteriores y a las posibilidades de las casas de apuestas. Al igual que el modelo basado en resultados anteriores, para el modelo basado en opiniones en Twitter, se creó un umbral de confianza. En este caso, el umbral de confianza del modelo fue considerado en 60%. Tomando en consideración este umbral de aceptación, el modelo sugirió considerar que el equipo de casa podía ganar en los siguientes partidos:

Fecha 32:

- Partido 4 - Manchester United vs. Swansea City
- Partido 9 – Arsenal vs. Stoke City

Fecha 33:

- Partido 9 – Arsenal vs. Southampton

Fecha 34:

- Partido 8 – Newcastle United vs. Arsenal
- Partido 9 – Manchester United vs. West Bromwich Albion

Fecha 35:

- Partido 1 – Brighton & Hove Albion vs. Tottenham Hotspur
- Partido 9 – Arsenal vs. West Ham United

De los 7 partidos mencionado anteriormente, el modelo pudo acertar en 5 oportunidades, dejando así un porcentaje de exactitud del 71%.

A continuación, se muestran los resultados para las fechas de competición 32, 33, 34 y 35 para la categoría donde el equipo de visita sale ganador.

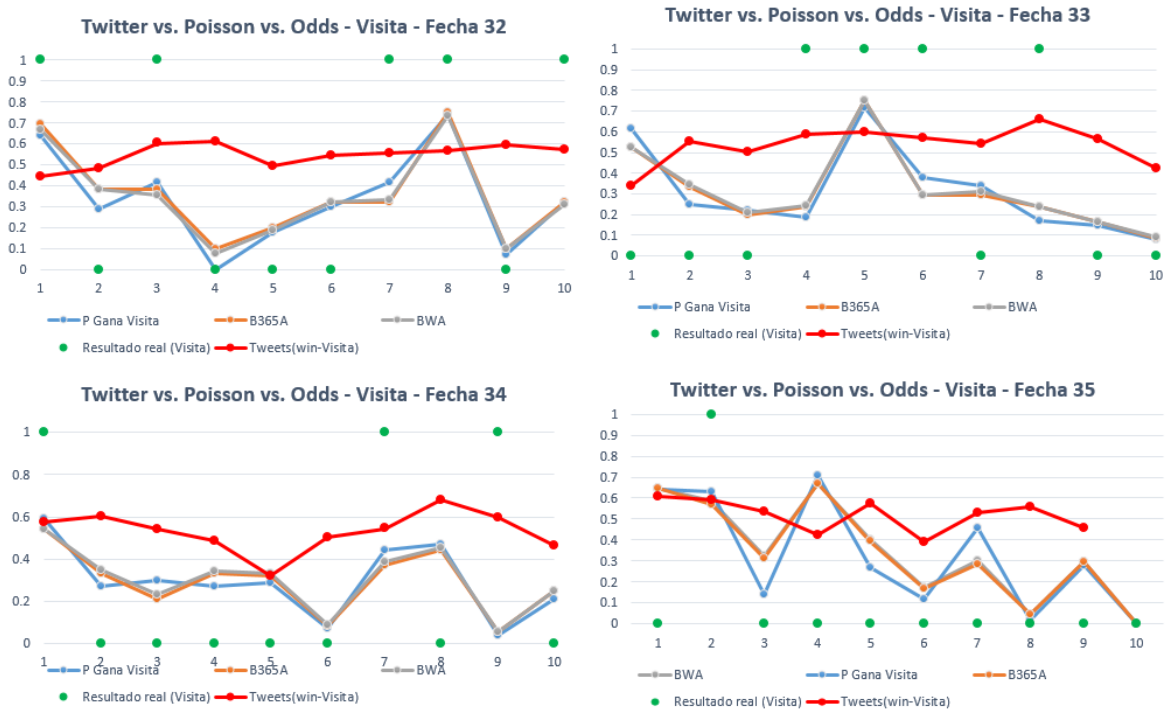


Figura 116 Exactitud del modelo basado Tweets para la categoría “Win-Visita” vs. Exactitud del modelo para la categoría “Gana Visita” vs. Posibilidades de las casas de apuestas - Fechas de competición 32, 33, 34 y 35.

Como se puede observar en la figura anterior, los resultados del modelo basado en opiniones en Twitter no se asemejan tanto con los resultados obtenidos con el modelo basado en resultados anteriores y con las posibilidades de las casas de apuestas en comparación con la categoría “Gana Casa”. Sin embargo, se estableció el mismo umbral de confianza de 60%. Tomando en consideración este umbral de aceptación, el modelo sugirió considerar que el equipo de visita podía ganar en los siguientes partidos:

Fecha 32:

- Partido 3 - Brighton & Hove Albion vs. Leicester City

Fecha 33:

- Partido 8 – Manchester City vs. Manchester United

Fecha 34:

- Partido 2 – Burnley FC vs. Leicester City

De los 3 partidos mencionado anteriormente, el modelo pudo acertar en 2 oportunidades, dejando así un porcentaje de exactitud del 66%.

A continuación, se muestran los resultados para las fechas de competición 32, 33, 34 y 35 para la categoría donde los equipos empatan el partido.

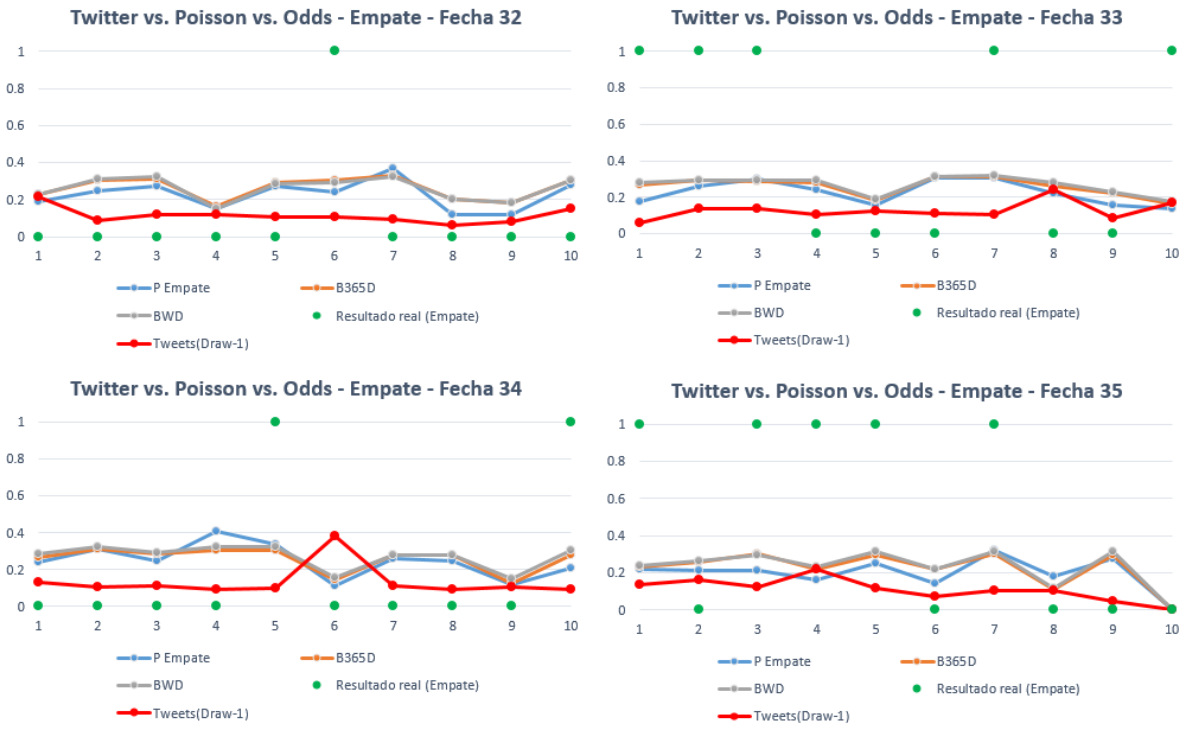


Figura 117 Exactitud del modelo basado Tweets para la categoría “Draw” vs. Exactitud del modelo para la categoría “Empate” vs. Posibilidades de las casas de apuestas - Fechas de competición 32, 33, 34 y 35.

Como se puede ver en la figura anterior, los resultados del modelo basado en opiniones en Twitter son casi constantes y se asemejan al modelo basado en resultados anteriores y a las posibilidades de las casas de apuestas.

Debido a su poca variabilidad, resulta un poco difícil establecer un umbral de confianza para esta categoría. Por esta razón, para esta categoría, no se estableció

un umbral de confianza. Esto resulta en la imposibilidad de poder predecir un resultado para esta categoría.

4.5 Objetivo No.5: Análisis de competitividad de la Premier League.

4.5.1 Crear método de medición de la competitividad

Antes de crear el método de medición de la competitividad de cualquier competición, hay que entender cómo funciona la competición y cuáles son sus características. No todas las competiciones de futbol son iguales en cuanto a su formato de competición. Existen competiciones que funcionan con eliminación directa, otras competiciones que poseen un formato de todos contra todos en donde cada encuentro es jugado dos veces (uno en cada casa de cada equipo), y otros donde ambos formatos se combinan.

Para el caso de la Premier League (y la mayoría de las ligas futbol de Europa), El formato de competición que se utiliza es el formato todos contra todos, donde los 20 equipos se enfrentan entre ellos dos veces, una vez por cada equipo estando en casa y otra vez estando de visita. Además, la competición adopta un sistema de puntos para poder clasificar a cada equipo y al final, poder elegir a un ganador. Este sistema de puntos es bastante básico; tres puntos para el ganador, cero puntos para el perdedor y un punto para los equipos que empatan.

Con base en esta información, el método de medición de la competitividad consta de un grafo de 20 nodos, uno por cada equipo, donde el tamaño del nodo representa la cantidad de puntos que cada equipo obtuvo en la competición. De esta manera se puede diferenciar los equipos grandes y poderosos de aquellos que no lo son.

Adicionalmente, el método de medición de la competitividad consta de conexiones entre nodos. Dos equipos están conectados si cada equipo al menos gana un punto en los dos partidos que disputan entre ellos. En otras palabras, dos equipos no se encuentran conectados si un equipo gana ambos encuentros. De esta manera, se puede medir la capacidad que tienen los equipos pequeños (o no tan pequeños) que quitarles puntos a los equipos grandes.

En general, para una liga con altos niveles de competitividad, se podría observar un grafo con nodos no muy grandes con respecto a los demás y con muchas conexiones entre ellos, mientras que, para una liga con bajos niveles de competitividad, se podría observar un grafo con algunos nodos muy grandes y con pocas conexiones entre ellos.

4.5.2 Aplicar el método de medición en la Premier League

Una vez creado el método para la medición de la competitividad, se tomarán los resultados que se han obtenido en la Premier League y se aplicarán los criterios expuestos en el punto anterior para crear un grafo donde se pueda visualizar la competitividad de la Premier League. A continuación, se muestra el grafo de competitividad para la temporada 2017/2018 de la Premier League.

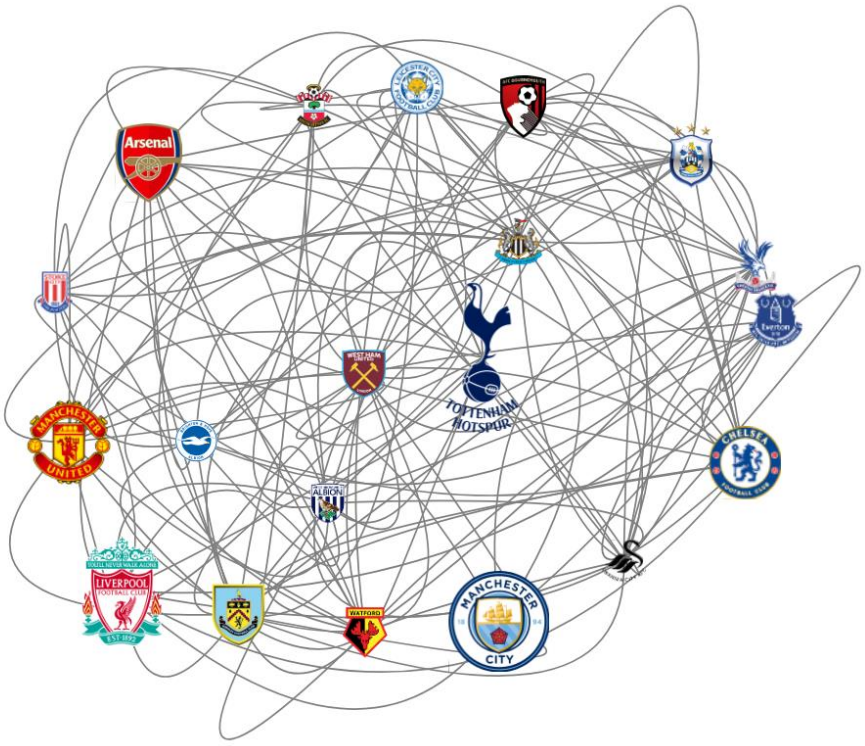


Figura 118 Grafo de competitividad de la temporada 2017/2018 de la Premier League

Como se puede observar en la figura anterior, no existe una diferencia muy grande de tamaño entre los nodos que componen el grafo. Sin embargo, se puede observar los equipos principales que lideran la competición, entre los cuales se encuentran el Manchester City, el Manchester United, el Liverpool, el Chelsea, el Arsenal y el

Totttenham. Adicionalmente, se puede observar que el grafo tiene una densidad de conexiones importantes entre los nodos. Se puede ver que equipos que están últimos en la clasificación como el Albion, el Stoke y el Swansea tienen una cantidad de conexiones importantes con el resto de los equipos. Esto quiere decir que, a pesar de que los equipos sean pequeños y estén en los últimos puestos de la competición, tienen la suficiente calidad y competitividad para quitarle puntos a los demás equipos.

4.5.3 Comparar la competitividad de la Premier Legue con otra liga de futbol europea

Una vez aplicado el método de medición de la competitividad a la Premier League, se aplicó el mismo método a La Liga de España para comparar los resultados. A continuación, se muestra el grafo de competitividad para la temporada 2017/2018 de La Liga de España.

Figura 119 Grafo de competitividad de la temporada 2017/2018 de La Liga de España

Como se puede observar en la figura anterior, existen cuatro equipos que se diferencian claramente del resto. Estos cuatro equipos son el Barcelona, el Real Madrid, el Atlético de Madrid y el Valencia. Por otro lado, se puede observar que este grafo posee menos conexiones de las que posee el grafo de la Premier League. Equipos pequeños como el Deportivo La Coruña, el Leganés o el Athletic de Bilbao, entre otros.

CAPITULO IV: CONCLUSIONES Y RECOMENDACIONES

A continuación, se listan las conclusiones que se obtuvieron en el presente trabajo final de máter.

Objetivo No.1: Descargar, clasificar, guardar y depurar información referente a Tweets, resultados de partidos anteriores y posibilidades de las casas de apuestas de equipos de futbol de la Premier League

- Una de las tareas más complejas y de las que más toma tiempo es la recolección y clasificación de los datos, sobre todo cuando se trata de la recolección de Tweets. Normalmente este proceso es muy lento y deben tomarse medidas de control para poder clasificar los datos de forma correcta para que los datos estén bien clasificados por semana y por equipo.
- El proceso de depuración de la información en vital para el posterior análisis de la información. Durante este proceso, se debe decidir qué parte de la información se pretende utilizar para el posterior análisis y, además, se deben aplicar técnicas de depuración especiales cuando se trata de información en texto y palabras. Técnicas como eliminar mayúsculas, eliminar valores numéricos, eliminar palabras vacías, eliminar signos de puntuación, eliminar los espacios en blanco, entre otras, son de vital importancia para realizar un análisis de sentimiento.

Objetivo No.2: Crear un modelo basado en las opiniones emitidas en la red social Twitter.

- Existe una relación inversamente proporcional entre el porcentaje de distribución de datos máximo entre las tres categorías y la exactitud del modelo. Lo ideal es que el porcentaje de datos máximo nunca supere el 50% para asegurar que el modelo posea una exactitud considerablemente aceptable.
- Mientras menos uniforme sea la distribución de los datos entre las categorías, menor será el porcentaje de exactitud del modelo.

- Existe una relación lineal entre el porcentaje de datos de una categoría y la exactitud del modelo para esa categoría.
- La exactitud del modelo se ve afectada por la exactitud del modelo para cada una de las categorías.
- Es recomendable que exista una distribución de datos equitativa entre todas las categorías del modelo para que su exactitud no esté afectada.
- El método Word2Vec puede ser utilizado para encontrar palabras que tengan más contexto con el conjunto de datos. En el caso de este trabajo de investigación, aun cuando el método calcula palabras según el contexto de cada categoría, las palabras que encuentra no tienen una relación única con cada categoría. Es el caso de las categorías “Draw” y “Lose”, las palabras que encuentra el método no poseen suficiente relación con la categoría para diferenciar una de otra. Sin embargo, para la categoría “Win”, el método encuentra palabras que están relacionadas fuertemente con la categoría y que ayudan a esta categoría a ser diferenciada de las otras dos.

Objetivo No.3: Crear un modelo estadístico basados en los resultados de partidos anteriores que tome en consideración una regresión lineal generalizada.

- En promedio, se anotan más goles cuando el equipo se encuentra en casa que de visita. Esto se debe a la bien conocida “ventaja de casa”.
- La cantidad de goles en un partido se puede modelar con una distribución de Poisson.
- Aun cuando existen diversas variables que pueden influir en el resultado de un partido de futbol, considerar los goles que han sido anotados por los equipos, pueden ayudar a predecir los resultados. En el caso de este estudio, con respecto a las fechas de competición 32, 33, 34 y 35, el modelo pudo acertar los resultados en un 51%.
- Si se desea aumentar la exactitud del modelo de Poisson, se pueden considerar umbrales de confianza con rangos de probabilidades mayores.

Esto aumentará considerablemente la exactitud del modelo, pero se reducirán la cantidad de partidos para predecir.

Objetivo No.4: Comparar los resultados del modelo estadístico basado en resultados anteriores con los datos de casas apuestas y los resultados del modelo basado en opiniones en Twitter.

- El modelo de predicción basado en resultados de partidos anteriores ofrece más información para predecir un partido de fútbol de la Premier League que solo tomar en consideración las posibilidades de las casas de apuestas.
- El modelo de predicción basado en resultados de partidos anteriores es sensible a resultados de partidos inesperados, por ejemplo, que el equipo que ocupa la posición 9 en la tabla de la competición pierda frente al antepenúltimo de la tabla.
- El modelo basado en opiniones en Twitter se puede utilizar para predecir partido de futbol, siempre que se considere un umbral de confianza del 60% o más.
- El modelo basado en opiniones en Twitter tiene inconvenientes para predecir resultados en donde ambos equipos empaten en el partido. Para este caso, la predicción se puede apoyar en el modelo de resultados anteriores o en las posibilidades de las casas de apuestas.

Objetivo No.5: Análisis de competitividad de la Premier League.

- La Premier League posee altos niveles de competitividad en los equipos que la conforman. Aun cuando existen equipos pequeños, éstos tienen la capacidad para ser suficientemente competitivos ante equipos más poderosos.
- En comparación con La Liga de España, la Premier League posee un nivel de competitividad mayor que hace que predecir resultados no sea una tarea fácil.

- Puede ser interesante aplicar los modelos de predicción en otras ligas menos competitivas como La Liga de España para explorar el impacto que pueden tener sobre la exactitud de los modelos.

REFERENCIAS BIBLIOGRAFICAS

Academy Khan. (2010). *Squared error of regression line*. Obtenido de https://www.youtube.com/watch?v=6OvhLPS7rj4&index=52&list=PL1328115D3D8A2566

Aggarwal, C. C. (2015). *Data Classification: Algorithms and Applications.* Obtenido de http://www.charuaggarwal.net/classbook.pdf

Aslam, S. (2018). Obtenido de https://www.omnicoreagency.com/twitter-statistics/

Asur, Huberman. (2010). *Predicting the Future With Social Media.* California.

AZARplus. (17 de Noviembre de 2017). Obtenido de http://www.azarplus.com/2017-11-17/arjel-registra-un-incremento-del-23-de-las-160apuestas-deportivas/14432/noticia/politica-de-privacidad.php

bet365. (2018). Obtenido de https://www.bet365.es/es/

Blue Square. (2018). Obtenido de https://www.bluesq.com/

Facebook, Inc. (31 de Enero de 2018). *Facebook Reports Fourth Quarter and Full Year 2017.* Obtenido de https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-Fourth-Quarter-and-Full-Year-2017-Results/default.aspx

Federico Pozzi, Elisabetta Fersini, Enza Messina, Bing Liu. (2017). *Sentiment Analisys in Social Networks.* Massachusetts: Elsevier Inc.

football-data.co.uk. (2018). Obtenido de http://www.football-data.co.uk/englandm.php

Gallagher, J. (12 de Mayo de 2017). *Home advantages and wanderlust*. Obtenido de https://jogall.github.io/2017-05-12-home-away-pref/

Jurafsky, Manning. (4 de Abril de 2012). *Formalizing the Naive Bayes Classifier* . Obtenido de https://www.youtube.com/watch?v=TpjPzKODuXo

Kampakis, S. (2014). *Utilizando Twitter para predecir resultados de Fútbol.* Obtenido de https://arxiv.org/ftp/arxiv/papers/1411/1411.1243.pdf

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Toronto: Morgan & Claypool.

Mónica Martínez; Manuel Marí. (s.f.). *La distribución Poisson.* Obtenido de https://riunet.upv.es/bitstream/handle/10251/7937/Distribucion%20Poisson.pdf

Paul, M; Dredze, M. (2011). *You Are What You Tweet: Analyzing Twitter for Public Health.* Obtenido de http://www.cs.jhu.edu/~mpaul/files/2011.icwsm.twitter_health.pdf

Sheehan, D. (Diciembre de 2016). Obtenido de https://dashee87.github.io/data%20science/football/Europes-Top-Football-Leagues-Getting-Less-Competitive/

Sinha, S; Dyer, C; Gimpel, K; Smith, N. (2013). *Predicting the NFL Using Twitter.* Obtenido de http://www.cs.cmu.edu/afs/cs/usr/kdeng/www/thesis/logistic.pdf

Sinha, S; Dyer, C; Gimpel, K; Smith, N. (2013). *Predicting the NFL Using Twitter.* Obtenido de https://www.cs.cmu.edu/~nasmith/papers/sinha+dyer+gimpel+smith.mlsa13.pdf

Statista. (2018). Obtenido de https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Trifacta Wrangler. (2018). Obtenido de https://www.trifacta.com/products/wrangler/

Tumasjan, A; Sprenger, T; Sandner, P; Welpe, I. (2010). *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* Obtenido de https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/download/1441/1852

Venkata, S; Kamal, N; Ganapati, P;. (2016). *Sentiment Analysis of Twitter Data for the Prediction of Stock Market Movement.* Obtenido de https://arxiv.org/pdf/1610.09225.pdf

Wikipedia. (s.f.). Obtenido de https://es.wikipedia.org/wiki/Premier_League

ANEXOS

Anexo 1

	Equipos 2018	Palabras Clave
1	Manchester United	@manutd
2	Swansea City	@swansofficial
3	West Bromwich Albion	@WBA
4	Watford	@WatfordFC
5	Brighton & Hove	@OfficialBHAFC
6	West Ham United	@WestHamUtd
7	Leicester City	@LCFC
8	Tottenham Hotspur	@SpursOfficial
9	Crystal Palace	@cpfc
10	Southampton	@southamptonfc
11	Stoke City	@stokecity
12	Newcastle United	@nufc
13	Arsenal	@arsenal
14	Burnley FC	@BurnleyOfficial
15	Huddersfield Town	@htafcdotcom
16	Chelsea	@chelseafc
17	Bournemouth	@afcbournemouth
18	Liverpool	@lfc
19	Everton	@Everton
20	Manchester City	ManCity

Anexo 2

```python
# -*- coding: utf-8 -*-
import sys,getopt,datetime,codecs
if sys.version_info[0] < 3:
    import got
else:
    import got3 as got

def main(argv):

    if len(argv) == 0:
        print('You must pass some parameters. Use \"-h\" to help.')
        return

    if len(argv) == 1 and argv[0] == '-h':
        f = open('exporter_help_text.txt', 'r')
        print f.read()
        f.close()

        return

    try:
        opts, args = getopt.getopt(argv, "", ("username=", "near=", "within=", "since=", "until=", "querysearch=", "toptweets", "maxtweets=", "output="))

        tweetCriteria = got.manager.TweetCriteria()
        outputFileName = "output_got.csv"

        for opt,arg in opts:
            if opt == '--username':
                tweetCriteria.username = arg

            elif opt == '--since':
                tweetCriteria.since = arg

            elif opt == '--until':
                tweetCriteria.until = arg

            elif opt == '--querysearch':
                tweetCriteria.querySearch = arg

            elif opt == '--toptweets':
                tweetCriteria.topTweets = True

            elif opt == '--maxtweets':
                tweetCriteria.maxTweets = int(arg)

            elif opt == '--near':
                tweetCriteria.near = '"' + arg + '"'

            elif opt == '--within':
                tweetCriteria.within = '"' + arg + '"'

            elif opt == '--within':
                tweetCriteria.within = '"' + arg + '"'

            elif opt == '--output':
                outputFileName = arg

        outputFile = codecs.open(outputFileName, "w+", "utf-8")

        outputFile.write('username;date;retweets;favorites;text;geo;mentions;hashtags;id;permalink')

        print('Searching...\n')

        def receiveBuffer(tweets):
            for t in tweets:
                outputFile.write(('\n%s;%s;%d;%d;"%s";%s;%s;%s;"%s";%s' % (t.username, t.date.strftime("%Y-%m-%d %H:%M"), t.retweets, t.favorites, t.text, t.geo,
            outputFile.flush()
            print('More %d saved on file...\n' % len(tweets))

        got.manager.TweetManager.getTweets(tweetCriteria, receiveBuffer)

    except arg:
        print('Arguments parser error, try -h' + arg)
    finally:
        outputFile.close()
        print('Done. Output file generated "%s".' % outputFileName)

if __name__ == '__main__':
    main(sys.argv[1:])
```