UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

EFECTO DEL BALANCEO DE CLASES AL EVALUAR EL F-SCORE PARA UN CLASIFICADOR DE TEXTO EN ANÁLISIS DE SENTIMIENTO

Trabajo Final de Investigación Aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Computación e Informática para optar al grado y título de Maestría Profesional en Computación e Informática

CARLOS FRANCISCO SOLÍS FONSECA

Ciudad Universitaria Rodrigo Facio, Costa Rica

2020

Dedicatoria

Este trabajo tan importante se lo dedico en primer lugar a Dios, definitivamente su palabra me llenó de fe para creer en mi persona y motivarme a atreverme a entrar a este tan importante y difícil programa. En segundo lugar, a mis padres porque a pesar de que vivimos situaciones difíciles siempre me consideraron capaz de seguir adelante. Por último, quiero dedicarle este trabajo a mi novia Malory Fallas Cerna y a nuestro primer hijo porque a pesar de que tuve que invertir mucho de mi tiempo en este trabajo, se mantuvieron a la par mía.

Agradecimientos

En primer lugar, quiero agradecer a los profesores Gilberth Brenes Camacho, Fernando Ramírez Hernández y Edwin Chaves Esquivel porque su enseñanza en la estadística me motivó a querer terminar la carrera y buscar superarme como profesional, además de que fueron las personas que, muy amablemente, accedieron a recomendarme al Programa de Posgrado en Computación e Informática.

En segundo lugar, quiero agradecer a todos los profesores del Posgrado en Computación e Informática porque en todos los cursos me motivaron a siempre dar lo mejor y despertaron en mí el interés por la investigación científica. Especialmente agradezco al Dr. Édgar Casasola Murillo que accedió a ayudarme a desarrollar el tema del cual se trata este documento, sin duda alguna no hubiese sido posible sin su apoyo.

También quiero agradecer a la Dra. Marcela Alfaro Córdoba y al Msc. Aurelio Sanabria Rodríguez que con su guía me ayudaron a mejorar y revisar el trabajo realizado en este documento. Adicionalmente tengo un gran agradecimiento a Mayid Sauma Ruiz quien fuera mi jefe en el departamento de Inteligencia de Negocios de la gerencia de Crédito y Cobro en BAC Credomatic ya que siempre me motivó a seguir estudiando y me permitió tomar las clases del posgrado sacrificando horas laborales.

También un agradecimiento muy fuerte a Valero Ramírez Talbott, el que fue en su momento supervisor del equipo de analistas de inteligencia de negocios, ya que también me permitió seguir las clases aún en momentos de mucho trabajo en la oficina. Por último, pero no menos importante, agradezco muchísimo a todas las personas que más que amigos los considero familia y que me ayudaron en muchas de las diferentes etapas del posgrado. A todos mi mayor y sincero agradecimiento.

"Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Computación e Informática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Computación e Informática."

Firmado digitalmente por JORGE ANTONIO LEONI DE LEON (FIRMA)
Fecha: 2021.01.31 08:09:36 -06'00'

Dr. Jorge Antonio Leoni de León
Representante del Decano
Sistema de Estudios de Posgrado

EDGAR ENRIQUE CASASOLA MURILLO (FIRMA)

Firmado digitalmente por EDGAR ENRIQUE CASASOLA MURILLO (FIRMA)
Fecha: 2021.02.01 20:36:55 -06'00'

Dr. Édgar Casasola Murillo
Profesor Guía

Marcela Alfaro Córdoba

Digitally signed by Marcela Alfaro Córdoba
Date: 2021.01.31 17:18:35 -06'00'

Dra. Marcela Alfaro Córdoba
Lectora

FIRMADO DIGITALMENTE POR AURELIO SANABRIA

TEC | Tecnológico de Costa Rica
Ingeniería en computación
Centro Académico Alajuela

MSc. Aurelio Sanabria Rodríguez
Lector

GABRIELA MARIN RAVENTOS (FIRMA)

Firmado digitalmente por GABRIELA MARIN RAVENTOS (FIRMA)
Fecha: 2021.02.02 09:57:44 -06'00'

Dra. Gabriela Marín Raventós
Directora
Posgrado en Computación e Informática

CARLOS FRANCISCO SOLIS FONSECA (FIRMA)

Firmado digitalmente por CARLOS FRANCISCO SOLIS FONSECA (FIRMA)
Fecha: 2021.01.30 19:31:16 -06'00'

Carlos Francisco Solís Fonseca
Sustentante

Índice

Resumen

En los últimos años los métodos de aprendizaje de máquina han sido incluidos en muchas áreas de investigación para dar soporte al análisis de datos. Los modelos de clasificación, los cuales corresponden a métodos de aprendizaje no supervisado, se han convertido en un instrumento muy importante para el análisis de texto y el análisis de sentimiento no es la excepción. Por esta razón es importante tener en consideración las características de los datos ya que, dependiendo de su naturaleza, pueden afectar la calidad del clasificador entrenado.

Este trabajo se centra sobre el problema en las clases desbalanceadas. Mediante dos métodos de balanceo, submuestreo y sobremuestreo, se lleva a cabo un experimento estadístico para determinar si balancear un conjunto de datos con clases desbalanceadas mejora o no la calidad de un clasificador usando máquinas de soporte vectorial.

Usando dos diferentes modelos para vectorizar palabras, TF-IDF y Word2Vec, se evalúa mediante un análisis de varianza el F-Score del clasificador de texto obteniendo como resultado un F-Score mayor cuando se usa sobremuestreo para balancear clases en comparación al F-Score obtenido del clasificador usando los datos con las clases desbalanceadas.

Palabras Clave: Balanceo de clases, análisis de sentimiento, aprendizaje de máquina, clasificación de texto, aprendizaje no supervisado, representación vectorial de palabras.

Lista de cuadros

Lista de figuras

Lista de gráficos

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Carlos Francisco Solís Fonseca, con cédula de identidad 1-1398-0914, en mi condición de autor del TFG titulado EFECTO DEL BALANCEO DE CLASES AL EVALUAR EL F-SCORE PARA UN CLASIFICADOR DE TEXTO EN ANÁLISIS DE SENTIMIENTO

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Carlos Francisco Solís Fonseca

Número de Carné: A76326 Número de cédula: 1-1398-0914

Correo Electrónico: solfonca@gmail.com

Fecha: 26 de enero de 2021 Número de teléfono: (506)8788-4190

Nombre del Director (a) de Tesis o Tutor (a): Édgar Casasola Murillo

CARLOS FRANCISCO SOLIS FONSECA (FIRMA)

Firmado digitalmente por CARLOS FRANCISCO SOLIS FONSECA (FIRMA)
Fecha: 2021.01.29 22:44:25 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no solo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

1. Introducción

Este documento consiste en una propuesta para trabajo final de investigación aplicada el cual explica y detalla un diseño de experimentos para determinar el efecto de balancear las clases de un conjunto de datos con clases desbalanceadas para análisis de sentimiento sobre el F1-Score resultante de un clasificador de texto.

Para lo cual, este documento se estructura mediante la presentación del problema de investigación del cual se deriva la pregunta de investigación. Seguidamente, se plantea el objetivo general y objetivos específicos con los que se busca responder a la pregunta del problema. Se finaliza la sección introductoria del documento con la justificación del aporte, beneficios y para qué sirve la investigación propuesta.

La segunda sección del documento presenta los resultados obtenidos mediante una revisión de literatura sobre trabajos de clasificación de texto. La tercera sección introduce una serie de conceptos clave para explicar y entender el contexto de la metodología a utilizar.

Por último, la cuarta sección establece la metodología de trabajo para alcanzar los objetivos trazados en la segunda sección del presente documento.

1.1 Problema

Actualmente las técnicas de aprendizaje de máquina son una herramienta que muchos profesionales y académicos utilizan para solucionar problemas (Japkowicz, 2000). La calidad de los resultados depende en muchas ocasiones no tanto de los algoritmos de clasificación como de la calidad de los datos. El tratamiento de los datos utilizados es fundamental y debe llevarse a cabo antes de entrenar un clasificador (Japkowicz, 2000). Realizar un análisis descriptivo previo de los datos, puede evidenciar patrones o información valiosa que puede ser utilizada para construir un clasificador. De esta forma, se puede lograr que el clasificador se vea menos afectado por el ruido o sesgo de los datos.

El resultado de algunas investigaciones, en el área de análisis de texto, evidencia mejora en los resultados obtenidos aplicando algoritmos para tratar el desbalanceo en los datos (Xu, y otros, 2015) y (Pramokchon & Piamsa-nga, 2014). La existencia de un desbalance en las clases de la variable o característica por estudiar agrega incertidumbre sobre los resultados obtenidos mediante un clasificador de texto (Hassanzadeh, Groza, Nguyen, & Hunter, 2014).

A pesar de haber documentos de análisis de texto que hablan sobre el desbalanceo en los datos y autores que obtienen una mejora en resultados, ninguna investigación construye una herramienta científica que permita hacer una comparación de resultados donde sea posible concluir si la mejora en los resultados se debe al balanceo de clases o no.

Lo anterior da como resultado el planteamiento de la siguiente pregunta la cual se desea responder mediante un diseño experimental. ¿Balancear las clases de un conjunto de datos mejora el resultado obtenido en el *F-Score* mediante el entrenamiento de un clasificador cuando se aplica a un problema de análisis de sentimiento?

1.2 Objetivos

1.2.1 Objetivo general

Evaluar mediante un diseño experimental si la aplicación de submuestreo y sobremuestreo para balancear las clases de un conjunto de datos mejora el *F-Score,* en comparación con un análisis que no balancee las clases al entrenar un clasificador para análisis de sentimiento.

1.2.2 Objetivos específicos

1. Seleccionar mediante revisión de literatura un método de sobremuestreo y un método de submuestreo para balancear las clases en un corpus de comentarios.
2. Construir un modelo de clasificación a partir de vectores de palabras usando un corpus de comentarios.
3. Establecer los factores fijos o aleatorios que se usarán en el diseño experimental.

4. Identificar diferencias, en el resultado del F-Score calculado, entre los métodos de balanceo de clases, sobremuestreo y submuestreo, y las clases sin balancear mediante un análisis de varianza.

1.3 Justificación

En análisis de texto o de sentimiento la fuente de datos está compuesta por documentos que a su vez están conformados por texto, y en algunos casos, etiquetados asignándole a cada texto una clase. Esta fuente de datos se conoce como corpus etiquetado. El corpus debe estar seleccionado equilibradamente para tener una mejor representatividad. Esto quiere decir, que, si el corpus es etiquetado por clases, su material debe de ser relativamente proporcional, evitando ser tendencioso a una parte únicamente tal y como lo expone Sierra (2015).

Por lo tanto, con la ayuda de un diseño experimental, la investigación a llevar a cabo pretende servir como evidencia científica para determinar si equilibrar las clases en un corpus mejora estadísticamente los resultados obtenidos mediante un clasificador de texto. De esta forma, esta investigación pretende determinar en qué medida las técnicas de reequilibrio de las clases puede ayudar a minimizar el sesgo. Ya que, existen otros factores que pueden afectar el resultado de un clasificador y conviene determinar cuál es el efecto del balanceo de clases sobre la calidad de los resultados. Y se recalca, que no ha sido posible encontrar algún trabajo específico orientado a llevar a cabo esta evaluación en este dominio del análisis de sentimiento en español.

Con los resultados del experimento a realizar se pretende generar información estadísticamente válida que permita beneficiar a los investigadores que trabajan con clasificadores de texto y permitirles no solo obtener mejores resultados, sino ayudarles con la construcción de corpus de texto para la creación de modelos de clasificación confiables.

Los resultados del experimento propuesto tienen un gran aporte científico ya que sirve como evidencia para justificar que el balanceo de clases permite obtener mejores resultados mediante el entrenamiento de un clasificador de texto. También, permite realizar mayor investigación para determinar nuevos algoritmos para balanceo de clases con el fin de obtener mejores resultados.

1.4 Alcance y limitaciones

La presente investigación se limitará a utilizar un único corpus, el cual es el utilizado para la competencia de InterTASS 2018 (SEPLN, 2020). Además, los resultados obtenidos en este experimento solamente pueden generalizarse sobre el corpus de texto usado en esta investigación. Se aclara que no se puede generalizar fuera de este dominio ya que para hacer una inferencia es necesario obtener una muestra aleatoria a partir de uno o varios corpus.

Debido a que el objetivo de la investigación es meramente comparativo, se usarán solamente factores fijos en el diseño semi-experimental. Además, todos los métodos seleccionados serán usados con la configuración de parámetros por defecto. El clasificador de texto a usar serán las máquinas de soporte vectorial, ya que, según los resultados del TASS 2018, son las que mejores resultados obtuvieron, sistema INGEOTEC_run 1 (Martínez-Cámara, y otros, 2018).

2. Estado del arte

Considerando que, según Sierra (2015), el equilibrio es una de las características que debe de cumplir un corpus lingüístico, el objetivo de realizar la revisión de literatura es determinar cómo los autores buscan satisfacer esta característica.

Al momento de realizar la revisión de literatura y usando la hilera de búsqueda: ("Data Balance") AND ("Text Classification" OR "Sentiment Analysis") se encontraron un total de 34 documentos. Como se explica en el gráfico 1, inicialmente solo se encontraron 8 documentos, los cuales tratan sobre el problema de investigación. Para iniciar la revisión de literatura se seleccionan otros 17 documentos los cuales son los más relevantes sobre clasificación de texto y/o análisis de sentimiento por lo cual se revisaron un total de 25 artículos de fuentes como Wiley, Springer y Redalyc. De los documentos obtenidos, en ningún trabajo se orientó la investigación hacia la evaluación del efecto del rebalanceo sobre la calidad de los resultados de clasificación. A continuación, se presentan los resultados de la revisión de literatura.

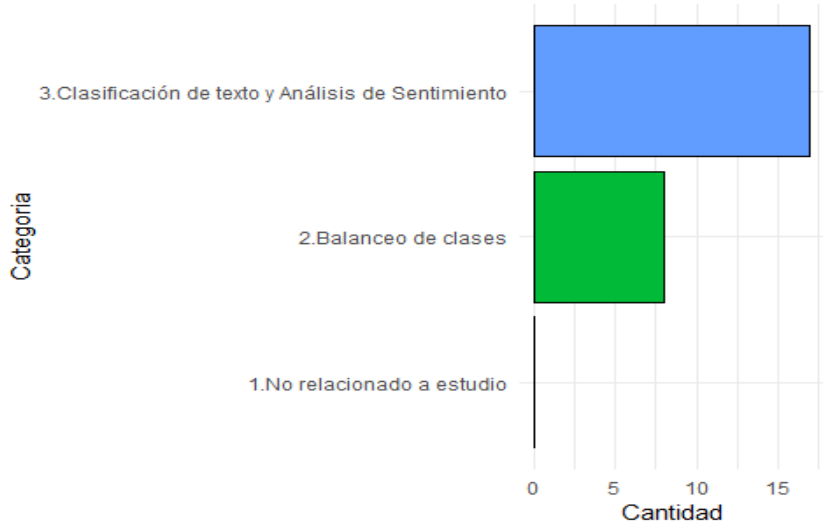

Gráfico 1. *Cantidad de documentos usados según la categoría*

De los 17 documentos sobre clasificación de texto y/o análisis de sentimiento. Uno de estos documentos consideró tratar el desbalance de observaciones en las clases a predecir mediante el entrenamiento de clasificador de texto (Chen, Mckeever, & Delany, 2016). De los 17

documentos sobre clasificación de texto y análisis de sentimiento, solo se encontró un documento sobre análisis de sentimiento (ver gráfico 2).

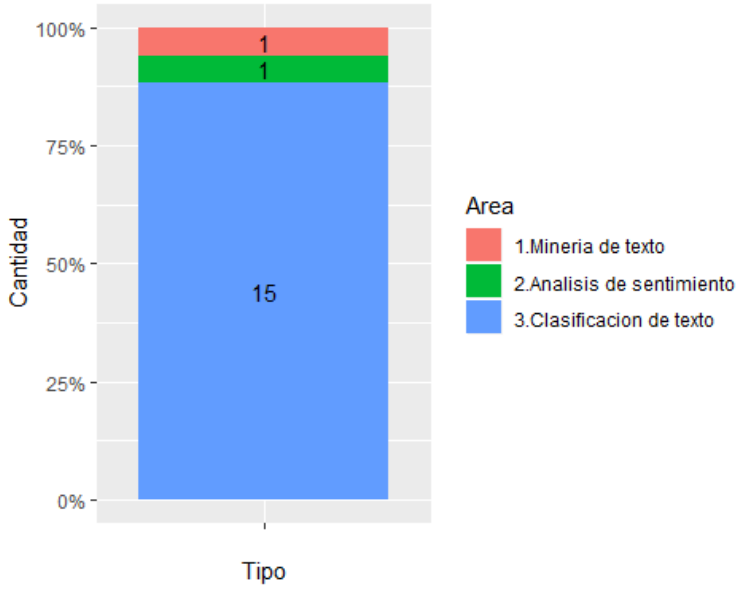

Gráfico 2. *Distribución porcentual según área de aplicación*

Mientras que el gráfico 3 muestra que, de los 8 documentos sobre balance de clases, la cantidad de documentos es la misma para análisis de sentimiento y clasificación de texto.

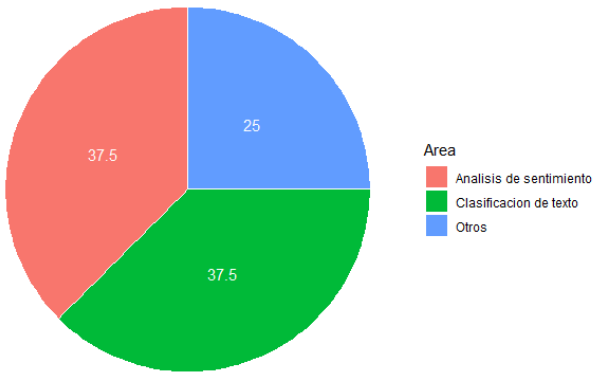

Gráfico 3. *Distribución de documentos según área de estudio*

La mayor parte de los documentos analizados permiten concluir que son pocos los autores que toman en consideración balancear las clases para disminuir el sesgo provocado por el desbalanceo en las predicciones. En los artículos revisados solamente se encontró un par de autores que consideran evaluar el efecto sobre los resultados obtenidos mediante el balanceo de clases (Xu, y otros, 2015) y (Pramokchon & Piamsa-nga, 2014).

Sin embargo, el autor no lleva a cabo un diseño experimental, sino que emplea elementos de estadística descriptiva (especialmente comparaciones entre los métodos estudiados) para evaluar la existencia o no de diferencias significativas.

La revisión realizada refleja que existen pocos trabajos que logren sustentar la hipótesis de que gracias al balanceo de clases se obtienen mejores resultados, ya que muchos de ellos trabajan sobre nuevos algoritmos de balanceo (Liu, y otros, 2014). Otros autores comparan los métodos de balanceo para determinar cuál es mejor (Filiberto, Caballero, Bello, & Frías, 2011) y (Li, Wang, Wang, Liang, & Li, 2019). Por lo tanto, existe una gran oportunidad de investigación para analizar mediante un diseño experimental los efectos del balanceo de las categorías empleando un análisis de sentimiento para el idioma español.

3. Marco teórico

En esta sección se definen algunos conceptos necesarios e importantes para lograr comprender la metodología a utilizar en el desarrollo del experimento. Estos conceptos ayudarán a esclarecer la razón por la que se escogen las metodologías escogidas y que aplicación y resultado tienen sobre el trabajo realizado.

Para lograr tener un mejor entendimiento de los conceptos se decide dividirlos en los temas centrales que abarca esta investigación. El primer de estos temas es el de representación de palabras en vectores sobre espacios numéricos o matemáticos. El segundo de los temas es sobre la definición de lo que son las clases de datos desbalanceadas y por siguiente los métodos existentes que pueden ser usados como medidas remediales a este problema.

El tercer tema abarca la definición de un clasificador y los elementos necesarios para poder entrenarlo para obtener predicciones mediante un modelo. El cuarto tema consiste en definir los conceptos clave para entender el objetivo y configuración de un diseño de experimentos. El quinto tema abarca los conceptos de los indicadores y métodos que se usarán para evaluar los resultados del clasificador entrenado.

Por último, el sexto tema comprende los conceptos necesarios para entender cómo se evalúan los resultados de un diseño experimental a nivel estadístico.

3.1 Representaciones vectoriales de palabras

La representación vectorial de palabras o *word embeddings* son representaciones numéricas de palabras según su contexto en un corpus. Las representaciones vectoriales de palabras capturan la información semántica y sintáctica de las palabras la cual puede ser usada para medir la similitud de palabras dentro de un espacio vectorial (Liu, Liu, Chua, & Sun, 2015).

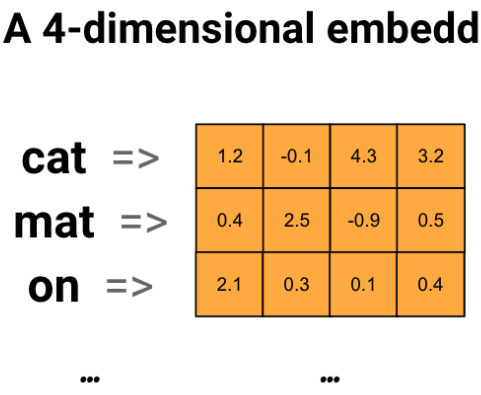

Figura 1. *Representación de palabras en vectores de cuatro dimensiones*
Tomado de: https://www.tensorflow.org/tutorials/text/word_embeddings

La figura 1 muestra un ejemplo de cómo se representa una palabra, en el ejemplo anterior tres palabras, en un vector de cuatro dimensiones. Gracias a esta metodología se pueden realizar operaciones algebraicas para determinar, dados vectores de palabras, cuál es la palabra más cercana. Un ejemplo básico de la funcionalidad de esta metodología es por ejemplo cuál es la palabra más cercana para resolver la pregunta ¿Cuál es la palabra para resolver la analogía Roma es la capital de Italia como París es la capital de?

Para contestar la pregunta anterior se realiza la siguiente operación matemática:

$$\text{vector X} = \text{vector}(\text{“Italia”}) - \text{vector}(\text{“Roma”}) + \text{vector}(\text{“París”})$$

Bajo un contexto semántico y sintáctico se esperaría que la respuesta o resultado de esta operación sea “Francia”, siempre y cuando la palabra exista en el corpus usado (Mikolov, Chen, Corrado, & Dean, 2013).

3.1.1 Construcción de representaciones vectoriales de palabras

Según (Angulo, 2019), en la construcción de representaciones vectoriales de palabras hay que considerar tres conceptos importantes: el corpus o textos de donde se obtienen las entradas de datos, los algoritmos utilizados para la creación de *word embeddings* y el contexto utilizado por ese algoritmo.

Con respecto a los algoritmos, (Baroni, Dinu, & Kruszewski, 2014), citados por (Angulo, 2019), menciona dos categorías de algoritmos para crear *word embeddings.* El primer método remarcado son los que se basan en métodos de conteo mediante cálculos estadísticos. El segundo método son los que se basan mediante una predicción a partir de las palabras vecinas en un espacio vectorial.

Existen varios algoritmos para construir *Word embeddings*: Tf-idf, GloVe y Word2Vec (Jurafsky y Martin, 2018), citado por (Angulo, 2019). Sin embargo, para esta investigación se explicará el algoritmo Word2Vec. Según lo antes mencionado, este algoritmo se basa en un algoritmo predictivo según la cercanía de otras palabras dentro de un espacio vectorial.

3.1.2 Word2Vec

Es un método predictivo para representar las palabras como vectores cortos y densos (donde la mayoría de los valores son distintos a cero). El método predictivo utiliza una técnica denominada ventana de contexto la cual recorre palabra por palabra y captura el contexto de una palabra objetivo, que será computado para crear los *word embeddings* (Angulo, 2019).

La idea principal de este método es que en lugar de contar la frecuencia con la que una palabra *w* aparece cerca de otras, se entrena un clasificador sobre una predicción binaria, basada en una palabra objetivo *w* y una palabra *c,* la cual es: ¿La palabra *c* aparece en el contexto de *w*? (Angulo, 2019).

Este método es propuesto por Tomas Mikolov y compañía en 2013 en el artículo llamado *"Efficient estimation of Word representations in vector space".* En este artículo los autores proponen dos arquitecturas para el método. La primera de ellas es *"Continuos Bag-of-Words"* (CBOW), el cual consiste en tratar de predecir la palabra objetivo basándose en las palabras que están en el contexto. La segunda arquitectura es el Skip-grama, el cual predice el contexto basándose en la palabra objetivo. Esta se utiliza como entrada para un clasificador lineal-logarítmico y se predicen las palabras dentro de un rango a la izquierda y a la derecha de la palabra (Angulo, 2019).

3.1.3 Corpus lingüístico

Un corpus lingüístico se puede definir como un conjunto de textos de materiales escritos y/o hablados, debidamente recopilados para realizar ciertos análisis lingüísticos. Para una definición de esta naturaleza es necesario detallar ciertas características que deben de cumplir este conjunto de textos para poder ser tratados como corpus lingüísticos (Sierra, 2015).

La primera característica de un corpus es que deben de ser representativos. La segunda característica es que deben de ser debidamente recopilados. La tercera característica es está relacionada directamente con la primera y la segunda, y es que el objetivo de un corpus lingüístico debe de ser construido para un análisis lingüístico (Sierra, 2015).

3.2 Tratamiento a las clases desbalanceadas

Dado un conjunto de datos cuando se trata de resolver un problema de clasificación es necesario tomar en cuenta la cantidad de observaciones en cada clase. Tener una buena representación de las diferentes instancias o clases es un elemento importante para una tarea de clasificación. Sin embargo, la mayor cantidad de veces nos encontramos con que las clases están sesgadas debido a que una de las clases tiene una proporción más alta de observaciones (Hassanzadeh, Groza, Nguyen, & Hunter, 2014)

La naturaleza de los datos o clases desbalanceadas reduce el rendimiento de los métodos de clasificación debido al sesgo que provoca la clase con mayor cantidad de observaciones. Lo anterior provoca que los errores de clasificación en la clase o clases con menor cantidad de observaciones sean mayores en comparación a las clases con más información, es decir falsos positivos (Hassanzadeh, Groza, Nguyen, & Hunter, 2014).

Cuadro 1. *Distribución de comentarios por categorías en el corpus InterTASS_CR*

Categoría	Total
N	912
NEU	297
NONE	447
P	677
Total	**2.333**

Con el Cuadro 1 se puede ejemplificar el desbalance en las clases. En este caso la clase con más observaciones es la N, la cual representa comentarios negativos, con 912 (39% del total). La clase con menor cantidad de observaciones es NEU, la cual representa comentarios neutros, con 297 (13% del total). Si se entrena un clasificador con estos datos posiblemente habría más probabilidad de obtener falsos positivos a la hora de clasificar comentarios que por su naturaleza son neutros.

Este tipo de problema ha provocado que haya investigaciones cuyo objetivo es crear algoritmos que funcionen como medida remedial. Volviendo con el ejemplo del cuadro 1 en donde tenemos cuatro clases, lo ideal sería que para cada una de ellas haya 583 observaciones aproximadamente, de esta forma nos aseguramos de que la información que provee cada clase sea la misma.

En este trabajo se van a usar dos técnicas, las cuales son las que más se usan según la revisión de literatura. La primera de ella se basa en algoritmos de submuestreo, usando el algoritmo *NearMiss*. La segunda técnica se basa en el sobre-muestreo usando el algoritmo SMOTE. Estas dos técnicas y algoritmos respectivos se explicarán más adelante.

3.2.1 Algoritmos de submuestreo

Los algoritmos de submuestreo se basan en disminuir *"artificialmente"* el número de observaciones que pertenecen a la clase con mayor número de elementos (Romero, Iglesias, & Borrajo, 2012). Este proceso se lleva a cabo hasta balancear las clases. Los algoritmos de esta clase trabajan bajo el supuesto que en la clase con mayor número de elementos existe cierta redundancia, por lo que una muestra con menor número de elementos es representativa para el resto (Dal Pozzolo, Caelen, Johnson, & Bontempi, 2015).

Algunos de los algoritmos para submuestreo son los siguientes: La regla del vecino cercano condensado, método del vecino cercano editado, método vecino cercano repetido, submuestreo aleatorio, Clúster generados por centroides, AllKNN, submuestreo por umbrales, regla de limpieza de vecinos, enlaces removidos de Tomek, *NearMiss* (Lemaitre, Nogueira, & Aridas,

2017). Debido a que en esta investigación se utilizará el método NearMiss, se procede a describirlo con mayor detalle a continuación.

3.2.2 Método *NearMiss*

Este algoritmo se basa la teoría del vecino más cercano o KNN por sus siglas en inglés (*K-Nearest Neighbor),* de esta forma obtiene una observación de entre los vecinos más cercanos (Zhang & Mani, 2003).

Específicamente, este método busca obtener, de la clase con más observaciones, los elementos cuyas distancias con los elementos de la clase o clases con menor cantidad de observaciones se la menor posible (Zhang & Mani, 2003).

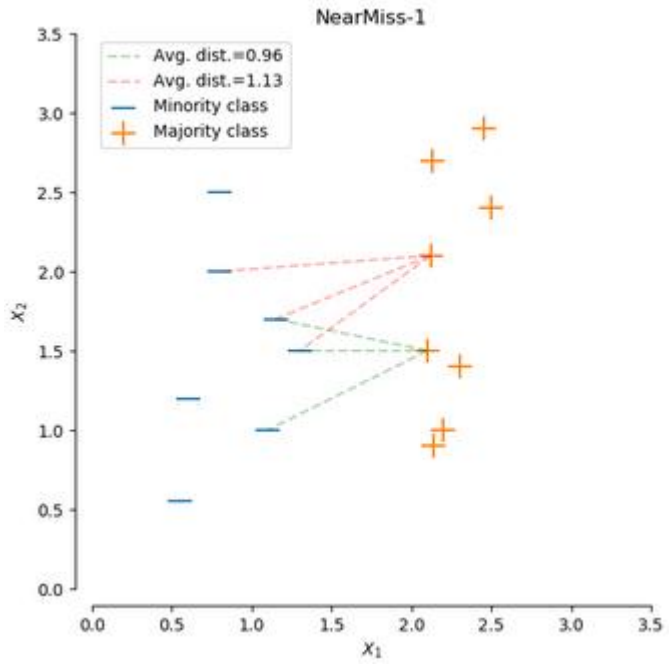

Figura 2. *Método NearMiss*
Tomado de: https://imbalanced-learn.org/stable/under_sampling.html#controlled-under-sampling

3.2.3 Algoritmos de sobremuestreo

Los algoritmos de sobremuestreo se basan en completar el número de observaciones de las clases con menor cantidad de ocurrencia, "duplicando" los casos mediante un muestreo. Lo que buscan estos algoritmos es que cada una de las clases tengan igual cantidad de observaciones. La desventaja de este método es que al duplicar la clase con menor cantidad de observaciones si agrega un sesgo (Dal Pozzolo, Caelen, Johnson, & Bontempi, 2015).

En el caso de sobremuestreo también existen varios algoritmos, algunos de ellos son: algoritmo adaptativo sintético (ADASYN), SMOTE, frontera con SMOTE, KNN con SMOTE, SVM-SMOTE y sobremuestreo aleatorio (Lemaitre, Nogueira, & Aridas, 2017). Debido a que en este trabajo se utilizará el método SMOTE se procede a describir el mismo con mayor detalle a continuación.

3.2.4 Método SMOTE

A continuación, se explica el método SMOTE que significa *Synthetic Minority Over-sampling Technique*. Esta técnica consiste en sobre muestrear las clases con menor cantidad de observaciones usando los mismos elementos de la clase minoritaria creando datos sintéticos usando vecinos cercanos (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

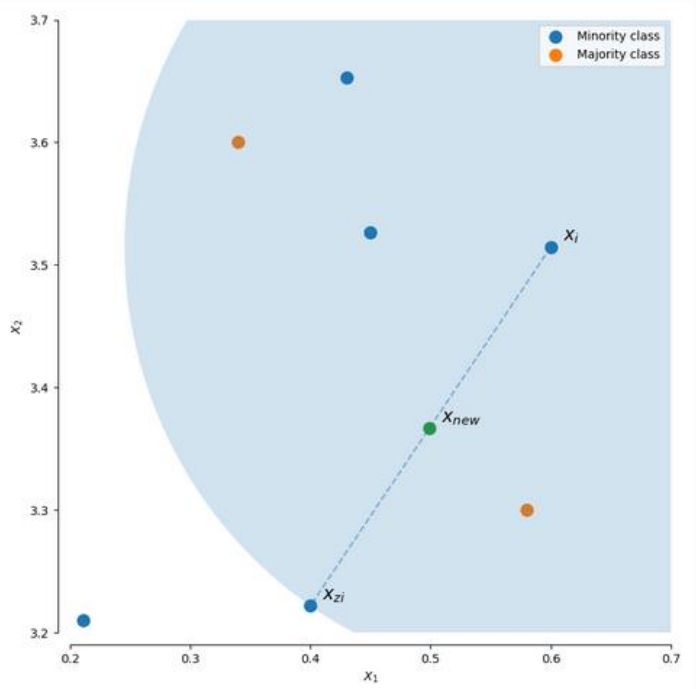

Figura 3. *Método SMOTE tomado de: https://imbalanced-learn.org/stable/over_sampling.html#smote-adasyn*

En la figura 3 se aprecia mejor el concepto de datos sintético, en la figura x_{new}. Este dato se calcula de la siguiente manera:

$$x_{new} = x_i + \lambda\, x\, (x_{zi} - x_i) \quad (1)$$

Donde λ es número aleatorio en el rango [0,1].

3.3 Clasificación de la polaridad

La clasificación de polaridad tiene la tarea de asignar, etiquetar o categorizar un texto o documento en específico. En el procesamiento de lenguaje natural los clasificadores de texto pueden analizar y asignar un conjunto de etiquetas o categorías predefinidas basadas en el contenido de los textos de forma automática (Angulo, 2019).

En el trabajo de Angulo, 2019, divide el trabajo de clasificación de texto en dos fases. La primera de ellas es la extracción de características mediante un método como TF-IDF, TF, Word2Vec, entre otras. La segunda fase consiste en utilizar una técnica de clasificación, por ejemplo: naive bayes, máquinas de soporte vectorial, redes neuronales entre otras.

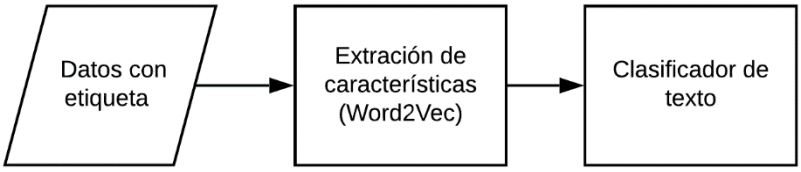

Figura 4. *Proceso clasificación de polaridad de texto*

En la figura 4 se detalla el proceso para clasificar documentos de texto. En la entrada de datos el documento consiste en los documentos de texto y si estos están clasificados en positivo, neutro o negativo. El siguiente paso es extraer las características de los documentos de texto. El tercer y último paso es usar las características extraídas para entrenar el clasificador de texto.

3.3.1 Conjunto de entrenamiento

Un conjunto de entrenamiento es una serie de ejemplos utilizados para aprender (Rashcka & Mirjalili, 2017) citado por (Angulo, 2019). Con este conjunto de datos el clasificador aprende los patrones que existen para clasificar los datos nuevos. En el caso de esta investigación el conjunto de entrenamiento consistirá en un porcentaje del corpus anotado.

3.3.2 Modelo de clasificación

Para el ejercicio de este experimento se hará uso de las máquinas de soporte vectorial como modelo de clasificación. Estas corresponden a máquinas de aprendizaje que toman distintas características de los elementos que se quieren clasificar y los llevan a un espacio vectorial multidimensional. Es en este espacio, donde el algoritmo identifica, de forma óptima, un hiperplano que separa a los vectores de una clase del resto (ver figura 5). Es en ese concepto de "separación óptima" donde reside la característica fundamental de estos algoritmos (Cárdenas, Olivares, & Alfaro, 2014).

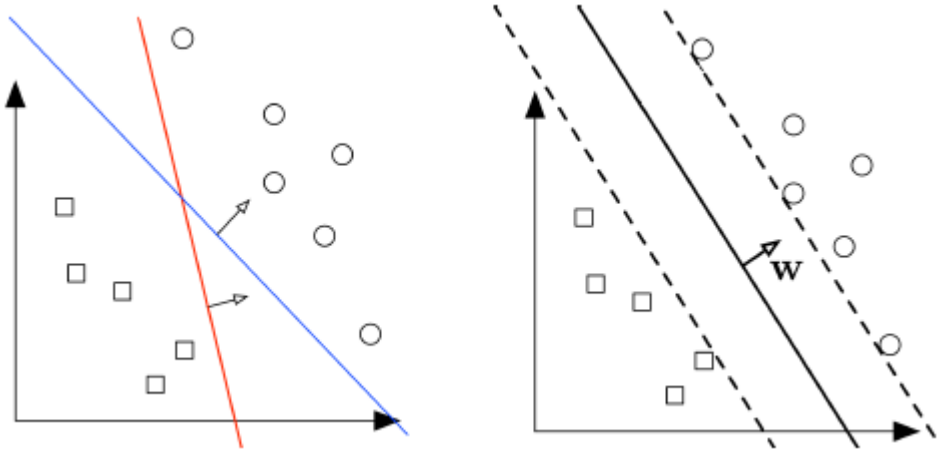

Figura 5. *Espacio bi-dimensional e hiperplano separando dos clases. Tomado de (Angulo, 2019)*

3.3.3 Análisis de sentimiento

El análisis de sentimiento se define como el estudio computacional de opiniones, sentimientos y emociones expresadas en textos. El objetivo principal del análisis computacional de sentimientos consiste en determinar la actitud de un escritor ante determinados productos, situaciones, personas u organizaciones, identificar los aspectos que generan opinión, quién las posee, y cuál es el tipo de emoción (me gusta, me encanta, lo valoro, lo odio) o su orientación semántica (positiva, negativa, neutra). (Dubiau & Ale, 2013).

El tipo de información que puede obtenerse utilizando sistemas de análisis de sentimientos incluye: polaridad de sentimientos en críticas sobre arte, productos o servicios; nivel de fidelización de clientes: opinión pública sobre representantes políticos o situaciones de interés social; predicciones sobre resultados de elecciones; tendencias de mercado, etc. (Dubiau & Ale, 2013).

3.4 Diseño de experimentos

El diseño de experimentos es un método estadístico que se aplica en una amplia gama de disciplinas tales como la medicina, agronomía, entre otras. La experimentación corresponde a una de las etapas del método científico cuyo propósito es construir una herramienta que permita conocer mejor cómo funciona un proceso o sistema mediante la construcción de conjeturas las

cuales se prueban mediante los resultados obtenidos los cuales, a su vez, puedan ser utilizados para crear nuevas conjeturas y ponerlas a prueba nuevamente (Montgomery, 2012).

En el trabajo propuesto se busca entender científicamente el efecto del balanceo de datos en la clasificación de sentimientos mediante el análisis de texto.

3.4.1 Factores

Para explicar los factores dentro de un experimento se usará un ejemplo, para de esta forma, explicar el efecto de este sobre la variable de estudio. Suponga que se desea saber qué efecto existe en la aplicación de la cantidad de watts en la vida útil de componente electrónico. En el caso del ejemplo anterior se tiene que la cantidad de watts es el factor ya que la aplicación de diferentes niveles de watts podría tener un efecto sobre la vida útil del componente electrónico.

Existen dos formas diferentes para determinar el efecto de los niveles del factor sobre la variable de estudio, los efectos fijos y los efectos aleatorios, ambos explicados más adelante. Para efectos de este trabajo se aplicarán efectos fijos.

Los efectos aleatorios son utilizados cuando dentro de los posibles niveles del factor existe un número grande de posibilidades. Un ejemplo de esto es cuando se elige usar un algoritmo de clasificación cuyo parámetro se encuentra en un rango [0,1]. A pesar de que sean números pequeños, en ese intervalo existe una cantidad de números infinita, por lo que se podría obtener una muestra aleatoria dentro de ese rango y aplicarlo como niveles en un factor, a lo anterior se le llaman efectos aleatorios (Montgomery, 2012).

Por otra parte, los efectos fijos corresponden a cuando el factor se puede configurar sobre un número finito de niveles. Para ejemplificar los efectos fijos se usará la investigación propuesta en este documento. En este caso el factor corresponde al tipo de técnica para balancear. Los niveles son finitos ya que sólo se cuenta con tres: sobremuestreo, submuestreo y un grupo control (Montgomery, 2012).

3.4.2 Tamaño de muestra

En diseño de experimentos el tamaño de la muestra es muy importante ya que como resultado de su tamaño se verá afectado el error de tipo 1 y el error de tipo 2. En diseño de experimentos el tamaño de la muestra viene dado por lo que se conocen como réplicas (Montgomery, 2012).

Por réplica se entiende la ejecución de la combinación de los niveles de un factor independiente a alguna otra ejecución realizada. El tamaño de la muestra está relacionado con error tipo 1, el cual consiste en la probabilidad de refutar estadísticamente nuestra hipótesis cuando es verdadera. El error de tipo 2 es la probabilidad de no refutar estadísticamente la hipótesis de investigación cuando ésta es falsa (Montgomery, 2012).

Cuando se desea minimizar esta probabilidad es necesario trabajar con una muestra mayor (Montgomery, 2012). Para obtener el tamaño de muestra se usa una función programada en el lenguaje de programación R la cual se llama *power.anova.test.* Esta función se enfoca en el análisis de varianza y utiliza como parámetros los siguientes valores:

- Varianza Entre Grupos: Este valor consiste en la varianza existente entre los promedios de los grupos.
- Varianza Intra Grupos: Este valor consiste en la varianza de todas las observaciones.
- Nivel de Significancia: Probabilidad de cometer error tipo 1 (5%)
- Tamaño de Muestra por Grupo: Este valor consiste en la cantidad de observaciones para cada grupo. En el experimento actual es el valor buscado.
- Poder de la Prueba: El poder de la prueba se calcula restándole a uno la probabilidad de cometer el error tipo 2. Para este experimento se busca que exista un 25% de probabilidad de cometer el error tipo 2, por lo tanto, el poder de la prueba buscado es del 75%.

3.4.3 Análisis de varianza de una vía

Montgomery, 2012, define el análisis de varianza como el análisis de una vía o análisis de varianza de un solo factor ya que se está investigando un único factor. Para ejemplificar esta definición en términos del problema en cuestión, el factor a estudiar es el balanceo de las categorías de los

datos. Existen dos modelos que son usados para realizar este análisis. El primero de ellos se llama el modelo de promedios y se define de la siguiente forma:

$$y_{ij} = \mu_i + \varepsilon_{ij}\{i = 1,2,\dots,a \;\; j = 1,2,\dots,n\} \quad (2)$$

Donde, y_{ij} es la ij-ésima observación, μ_i es el promedio del i-ésimo nivel del factor o tratamiento, y ε_{ij} es el componente de error aleatorio que es incorporado por todas las fuentes de variabilidad en el experimento. El segundo modelo se llama el modelo de efectos y se define de la siguiente forma:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}\{i = 1,2,\dots,a \;\; j = 1,2,\dots,n\} \quad (3)$$

En esta forma del modelo, μ es un parámetro común entre todos los tratamientos y se llama el promedio global, y τ_i es un parámetro único asociado al i-ésimo tratamiento llamado i-ésimo efecto de tratamiento. Para probar de forma correcta la hipótesis, se establecen los siguientes supuestos: los errores del modelo deben de distribuirse de acuerdo con una distribución normal y ser independientes con promedio cero y varianza σ^2 asumiendo que es constante en todos los tratamientos.

3.4.4 Contraste de hipótesis

Una hipótesis estadística es una afirmación sobre el valor de unos parámetros o distribución la cual refleja una conjetura sobre algún problema en específico (Montgomery, 2012). Mediante el uso de técnicas estadísticas tales como análisis de varianza, comparación de medias entre diferentes grupos, entre otros, se busca obtener evidencia estadística para, por lo general, refutar la conjetura planteada.

La conjetura planteada es tratada como la hipótesis nula, generalmente denotada de la forma H_0, la cual da como resultado de forma automática la existencia de una hipótesis alternativa, generalmente denotada de la forma H_1. La decisión que se tome respecto a la hipótesis nula debe de basarse en un cierto número de observaciones.

La prueba de hipótesis es un problema típico de decisión ante la incertidumbre. Con ayuda de la probabilidad y la teoría estadística se obtiene y recopilan los datos suficientes para tomar una decisión respecto a la hipótesis nula y además a disminuir el grado de incertidumbre.

3.5 Evaluación de los resultados de clasificación

En la evaluación de los resultados de clasificación se tomarán las medidas del conjunto de prueba usado en el clasificador. Para esta medición se tabulan los resultados para comparar el valor real de cada clase con el valor asignado por el clasificador. Una vez tabulados los resultados se pueden obtener tres métricas que más adelante son explicadas con detalle. Estas métricas son: la precisión, la exactitud y la cobertura o exhaustividad. En conjunto corresponden un primer método de medición.

Un segundo método de medición es el F-Score. Este método es el usado para evaluar los clasificadores a usar en el diseño experimental. Consiste en un indicador compuesto por las tres métricas mencionadas anteriormente. Se describe con mayor detalle en esta sección.

3.5.1 Conjunto de pruebas

En el caso de la clasificación de texto, cada elemento de un conjunto de pruebas está constituido por un texto, y su etiqueta o polaridad. A diferencia del conjunto de entrenamiento que es utilizado para entrenar los clasificadores, el conjunto de pruebas es utilizado para evaluar los modelos de clasificación (Angulo, 2019).

El conjunto de pruebas tiene que ser independiente al conjunto de entrenamiento, esto es, que no tiene que existir intersección entre los elementos, es decir, una observación que se encuentre en el conjunto de datos de entrenamiento no puede existir en el conjunto de datos de prueba y viceversa. Ambos conjuntos deben de seguir la misma distribución probabilística (Angulo, 2019).

3.5.2 Métricas

Para determinar las métricas necesarias para el cálculo del indicador F-Score es necesario definir en primer lugar lo que se conoce como la matriz de confusión que consiste en una tabla de contingencia.

Cuadro 2. *Matriz de confusión para tres clases*

Real / Predicho	Ap	Bp	Cp
Ar	ArAp	ArBp	ArCp
Br	Br Ap	Br Bp	Br Cp
Cr	Cr Ap	Cr Bp	Cr Cp

En el Cuadro 2 las filas representan las clases reales, mientras que las columnas representan las clases predichas por el clasificador. De este cuadro es que salen los indicadores necesarios para el cálculo final del F-Score. Es importante mencionar que los valores en la diagonal de la matriz mostrada en el Cuadro 2 corresponde a los casos verdaderos mientras cualquier otro valor corresponderá a una clasificación falsa por parte del clasificador.

Las métricas siguientes se calculan a partir de la matriz obtenida del Cuadro 2.

Exactitud o *Accuracy*

Esta métrica corresponde a la suma de los elementos de la diagonal de la matriz dividido entre la suma de casos total. Esta métrica se puede definir de la siguiente manera:

$$\frac{\sum V}{\sum V + \sum F} \tag{4}$$

Donde V corresponde a los elementos de la matriz en el cuadro 1 que pertenecen a la diagonal y F corresponde a los elementos de matriz en el cuadro 1 que no pertenecen a la diagonal.

Precisión o *Precision*

Esta métrica corresponde al elemento verdadero de la fila que contiene a las clases reales dividido entre la suma de los elementos de la clase real. Esta métrica se puede definir de la siguiente manera:

$$\frac{\mathrm{ArAp, Br\ Bp, Cr\ Cp}}{\sum A_r, \sum B_r, \sum C_r} \tag{5}$$

Cobertura o exhaustividad o *Recall*

Esta métrica corresponde al elemento verdadero de la columna que contiene a las clases predichas dividido entre la suma de los elementos de la clase predicha. Esta métrica se puede definir de la siguiente manera:

$$\frac{\text{ArAp, Br Bp, Cr Cp}}{\sum A_p \, , \sum B_p \, , \sum C_p} \tag{6}$$

Puntaje F

El puntaje F o macro *F-Score* es una medida estadística para medir la precisión predictiva de un clasificador entrenado mediante alguna técnica de aprendizaje de máquina. Este indicador se calcula utilizando la información del indicador de precisión y la información del *recall*.

$$F_\beta = (1 + \beta^2) x \frac{Precisión \; x \; Recall}{\beta^2 Precisión \; + \; Recall} \tag{7}$$

En la fórmula 7, el valor del parámetro β = 1.

3.6 Evaluación de los resultados del experimento

Para evaluar los resultados del experimento es necesario llevar a cabo un análisis estadístico que permita establecer conclusiones sobre la existencia de un efecto o no del balanceo de clases sobre el F-Score. El análisis escogido para llevar a cabo el experimento es el análisis de varianza de una vía, el cual permitirá saber si hay un grupo diferente si los datos cumplen con los supuestos del modelo.

Si se determina la existencia de una diferencia es necesario evaluar dónde existe esa diferencia usando una técnica de comparación.

3.6.1 Análisis estadístico

Para el análisis estadístico se asume que los errores o incertidumbre que afecta el análisis se distribuye de manera normal y es independiente, los mismos supuestos aplican para el F-Score. El estadístico por utilizar para poner a prueba la hipótesis nula se llama el estadístico F y este se obtiene del cuadrado medio de los tratamientos y el cuadrado medio del error (Montgomery, 2012).

Cuadrado medio de los tratamientos

Esta medida de la suma de cuadrados de los tratamientos. La suma de cuadrados de los tratamientos consiste en la multiplicación del tamaño de muestra o replicas por la suma de la diferencia al cuadrado del promedio del tratamiento i con el promedio global. Esta se define de la siguiente manera:

$$SC_{Tratamientos} = n\sum_{i=1}^{a}(\overline{y_{i.}} - \overline{y_{..}})^2 \tag{8}$$

La suma de cuadrados de tratamientos se divide entre la cantidad de tratamiento menos uno y da como resultado el cuadrado medio de los tratamientos. Es decir,

$$CM_{Tratamientos} = \frac{SC_{Tratamientos}}{a-1} \tag{9}$$

Donde a es la cantidad de tratamientos.

Cuadrado medio de los errores

Esta medida se compone de la suma de cuadrados del error, la cual se calcula mediante la diferencia de la suma de cuadrados total y la suma de cuadrados de tratamientos. En primer lugar, se define de la siguiente manera la suma de cuadrados total:

$$SC_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 \tag{10}$$

Por lo tanto:

$$SC_E = SC_T - SC_{Tratamientos} \tag{11}$$

Por último, el cuadrado medio del error se divide por la diferencia de la cantidad total de observaciones y cantidad de tratamientos.

$$CM_E = \frac{SC_E}{N - a} \tag{12}$$

Estadístico F

Este estadístico sigue una distribución de probabilidad F y se calcula dividiendo el cuadrado medio de los tratamientos entre el cuadrado medio de los errores. Este estadístico se define de la siguiente manera:

$$F_0 = \frac{CM_{Tratamientos}}{CM_{Error}} \tag{13}$$

Este estadístico se interpreta como el estadístico calculado y se compara contra un estadístico teórico que se define mediante la distribución de probabilidad. En este experimento la regla para este estadístico se determinará mediante el valor p, el cual determina la probabilidad de error del experimento y la cual se compara contra una probabilidad teórica la cual es 5%.

3.6.2 Análisis por intervalos de confianza

Este análisis se utiliza para expresar los resultados finales en términos de intervalos de confianza y además desea mostrar que tan amplios pueden ser de acuerdo con la configuración del experimento (Montgomery, 2012). Los intervalos de confianza se calculan de la siguiente manera:

$$\bar{y}_{..} - t_{\frac{\alpha}{2},a(n-1)}\sqrt{\frac{MS_{Tratamientos}}{an}} \leq \mu \leq \bar{y}_{..} + t_{\frac{\alpha}{2},a(n-1)}\sqrt{\frac{MS_{Tratamientos}}{an}} \tag{14}$$

El análisis mediante intervalos de confianza es útil ya que si alguno de los intervalos se interseca entonces podemos concluir que no hay diferencia estadística, mientras si los intervalos no se intersecan entonces si existe una diferencia entre promedios.

3.6.3 Pruebas de comprobación supuestos

Cumplir con los supuestos es muy importante para que la prueba tenga validez estadística, ya que la violación de estos supuestos puede significar que la configuración del experimento no es necesariamente la mejor. Para este experimento hay dos supuestos fundamentales a verificar. El primero es la prueba de normalidad y el segundo es de igualdad de varianzas (Montgomery, 2012).

Prueba de normalidad

Normalmente este supuesto se verifica de manera visual graficando los errores sobre una distribución teórica normal.

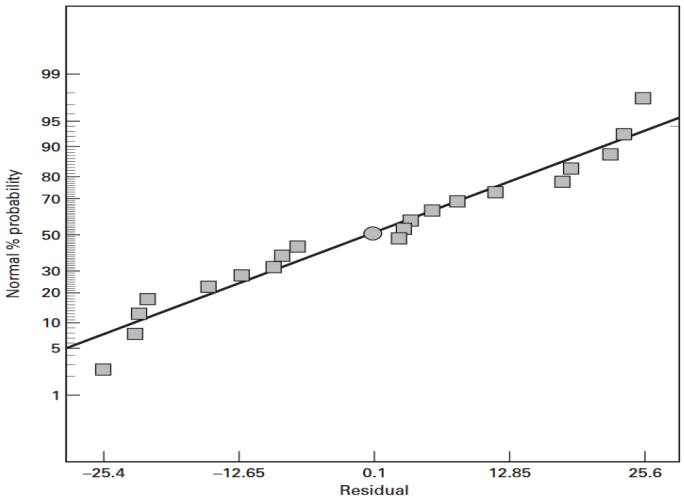

Figura 6. *Gráfico probabilidad normal de residuos*

En la figura 6, se aprecia que los errores mantienen una distancia mínima sobre la línea teórica de probabilidad, por lo que no se viola el supuesto de normalidad. Sin embargo, existen pruebas estadísticas que permiten determinar si la distribución de las observaciones sigue una distribución de probabilidad normal. Una de estas pruebas es la de Shapiro-Wilk, la cual se basa en el cálculo de un estadístico W el cual se compara con una distribución de probabilidad teórica (Shapiro & Wilk, 1965).

Prueba para normalidad Shapiro – Wilk

La prueba de normalidad de Shapiro – Wilk se basa en el cálculo del estadístico W, el cual se define de la siguiente forma:

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

La prueba de hipótesis a contrastar para determinar si un conjunto de datos sigue o no una distribución normal es la siguiente:

$$H_0: Los\ datos\ siguen\ una\ distribución\ normal$$
$$H_1: Los\ \ datos\ no\ siguen\ una\ distribución\ normal$$

Lo que se busca con esta prueba es tener la suficiente evidencia estadística para no rechazar la hipótesis nula.

Prueba igualdad de varianza

Para verificar la igualdad de varianza se usará una prueba formal. Esta prueba formal se llama la prueba de Levene y utiliza la desviación absoluta de las observaciones de cada tratamiento con respecto a la mediana. La hipótesis por evaluar es la siguiente:

$$H_0: \sigma_1^2 = \ \sigma_2^2 = \cdots = \ \sigma_a^2$$
$$H_1: Existe\ al\ menos\ una\ varianza\ distinta$$

Lo que se busca con esta prueba es tener la evidencia estadística suficiente para no rechazar la hipótesis nula de igualdad de varianzas.

3.6.4 Violación de supuestos

Los principales supuestos del análisis de varianza de una vía son el de normalidad y varianza equitativa. Actualmente existen técnicas que permiten llevar a cabo el análisis aun cuando estos son violados. Estas técnicas son conocidas como medidas remediales. A continuación, se detalla el procedimiento a aplicar cuando se viola el supuesto de normalidad y el de igualdad de varianza.

Medida remedial para supuesto de normalidad

Cuando la prueba de normalidad evidencia que los datos no siguen una distribución normal, la medida remedial que se aplica es el uso de mínimos cuadrados ponderados el cual se basa en un método generalizado de regresión (Kutner, Nachtsheim, Neter, & Li, 2005).

Medida remedial para supuesto de varianza

Cuando el supuesto de igualdad de varianza no se cumple se puede usar la misma medida remedial usada cuando no se cumple el supuesto de normalidad. Sin embargo, existe un análisis de varianza que se puede usar aun cuando hay diferencias de varianza. Esté método se llama el Análisis de Varianza de Welch.

El análisis de varianza de Welch es una variante del análisis de varianza convencional el cual no contempla el supuesto de igualdad de varianza (Glen, 2016). El estadístico F se define de la siguiente forma:

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^{k} w_j\left(\bar{x_j} - \bar{x}'\right)^2}{1 + \frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left(\frac{1}{n_j - 1}\right)\left(1 - \frac{w_j}{w}\right)^2} \tag{15}$$

Dónde,

$$w_j = \frac{n_j}{{s_j}^2} \qquad w = \sum_{j=1}^{k} w_j \qquad \bar{x}' = \frac{\sum_{j=1}^{k} w_j \bar{x}_j}{w}$$

La distribución teórica de la ecuación (15) se modela de la siguiente forma: F ~ F(k – 1, gl), dónde:

$$gl = \frac{k^2 - 1}{3\sum_{j=1}^{k}\left(\frac{1}{n_j - 1}\right)\left(1 - \frac{w_j}{w}\right)^2}$$

3.6.5 Pruebas de comparación

Las pruebas de comparación se deben de realizar si el análisis de varianza determina la existencia de una diferencia de promedios. Si esto ocurre es necesario realizar las comparaciones entre tratamientos para determinar cuál de ellas es la diferente.

En este experimento uno de los tratamientos consiste en un grupo control, por lo tanto, se usará la prueba Post-Hoc no paramétrica Games-Howell para determinar si existe una diferencia estadística. Esta prueba se define de la siguiente manera (Schlegel, 2016):

$$\bar{x}_i - \bar{x}_j > q_{\sigma,k,gl} \tag{15}$$

Donde σ es.

$$\sigma = \sqrt{\frac{1}{2}\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)}$$

Los grados de libertad se calculan de la siguiente forma (corrección de Welch):

$$\frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\frac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i - 1} + \frac{\left(\frac{s_j^2}{n_j}\right)^2}{n_j - 1}}$$

El valor t se calcula con la prueba de Welch:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$$

4. Metodología

La metodología de trabajo se desarrolló en su mayoría en el curso de PF-3397 Procesamiento de Lenguaje Natural y además se profundizaron algunos aspectos en el curso PF-3394 Recuperación de Información, ambos del Programa de Posgrado en Computación e Informática. Esta metodología se centrará en **4 etapas,** cada una orientada en cumplir con los objetivos específicos planteados.

La primera fase, asociada al primer objetivo específico, consiste en el trabajo previo que debe realizarse al conjunto de datos que será usado para el experimento. En esta etapa se explora la distribución del corpus con respecto a las categorías para determinar el número de clases a ser usadas. Por ejemplo: en Intertass se cuenta con 4 categorías donde se incluye NEUTRO y NONE, pero vale la pena unificarlas para reducir la clasificación a 3 categorías. En esta etapa también se hace la configuración del ambiente de trabajo.

La segunda fase, asociada al segundo objetivo específico, consiste en hacer uso de librerías especializadas para crear los vectores de palabras y además las librerías para entrenar los modelos de clasificación creadas para Python 3. En esta fase se prepara el diseño experimental, aleatorizando las unidades de estudio a los tratamientos a ser aplicados.

La tercera fase, asociada al tercer objetivo específico, consiste en entrenar el clasificador para cada una de las observaciones aplicando el tratamiento determinado para el tratamiento. En esta fase también se lleva a cabo un análisis de poder del experimento para determinar el tamaño de muestra. Finalmente, **la cuarta fase** asociada al cuarto objetivo específico, consistirá en la verificación de supuestos y a realizar el análisis de varianza y la comparación entre tratamientos para lograr hacer una conclusión.

Adicionalmente se agrega una **quinta fase** la cual consiste en ejecutar el experimento usando TF-IDF como algoritmo para vectorizar palabras. Esta fase consiste en comparar resultados entre el método de *Word2Vec* y TD-IDF.

4.1 Fase de preparación

En esta fase se explorará la distribución de las categorías del corpus a usar para determinar el número de clases a usar para el experimento. Una vez analizada la distribución de las clases se procede con el preprocesamiento del conjunto de datos. Para esta etapa se seguirá la estrategia de (Angulo, 2019). Se tomará en cuenta los diacríticos, emoticones, *hashtag,* mayúsculas, mención, numeral, puntuación, stopwords, url, fechas, elongaciones y se realizarán recodificaciones de caracteres.

Una vez el corpus haya sido normalizado o pre procesado, la última etapa de esta fase consistirá en la creación de los vectores de palabras mediante el uso de Word2Vec. En el cuadro 3 se muestran los parámetros con los que se construye el vector de palabras usando la librería gensim para Python.

Cuadro 3. *Valor de parámetros para modelo de Word2Vec*

Parámetro	Valor
size (Tamaño vectores)	300
window (Ventana de palabras)	2
Sg (Algoritmo entrenamiento)	1 (Skip-gram)
negative (muestreo negativo)	10
Iter (iteraciones)	5

4.2Fase de configuración experimento

En esta fase se utiliza *NearMiss* como el método a usar para submuestreo ya que la metodología se basa en vecinos cercanos, el cual es la base del resto de métodos de submuestreo. La misma decisión se toma para el método de sobremuestreo ya que el resto de los métodos son un algoritmo modificado del SMOTE. Tanto al submuestreo como al sobremuestreo tienen métodos aleatorios, sin embargo, para no incurrir en un sesgo provocado por aleatoriedad se decide usar otro método.

En esta fase se elaborará un pequeño experimento con el clasificador para tener la información suficiente para determinar el tamaño de muestra para el experimento. Después de determinar el tamaño de muestra se obtendrán diferentes conjuntos de entrenamiento y prueba, similar a una validación cruzada, sin embargo, no se trata de una validación cruzada.

Este proceso se debe de llevar a cabo de esta forma para aleatorizar la aplicación de los tratamientos, ya que este es el principio fundamental de un diseño experimental. En total se aplicarán tres tratamientos. El primero es entrenar el clasificador sin balancear las clases. El segundo tratamiento es utilizar la técnica SMOTE para sobremuestreo. El tercer y último tratamiento consiste en aplicar la técnica *NearMiss* para submuestreo.

4.3Fase de entrenamiento clasificador

La asignación de los tratamientos se describe en la figura 7. Una vez que se asigna el tratamiento al conjunto de entrenamiento y prueba se llevará a cabo el entrenamiento del clasificador para cada uno de esos conjuntos. El método de entrenamiento escogido es una máquina de soporte vectorial siguiendo la configuración de (Angulo, 2019).

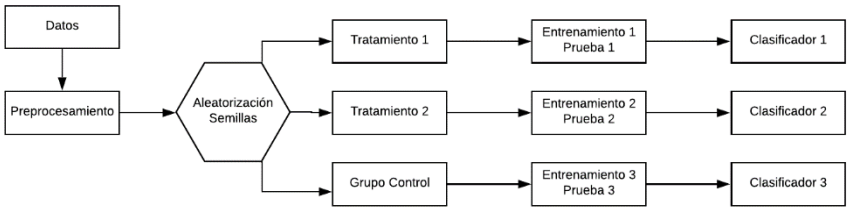

Figura 7. *Proceso de aleatorización para aplicación de tratamientos*

A cada uno de estos conjuntos se le calculara el F-Score para proceder a la última fase del experimento.

4.4Fase de pruebas y conclusiones

En esta fase se recopilan los diferentes F-Score obtenidos de los clasificadores entrenados y se procederá a verificar los supuestos establecidos para un análisis de varianza de una vía. En primer lugar, se evalúan los supuestos de normalidad e igualdad de varianza usando las pruebas de

Shapiro-Wilk y Levene respectivamente. Una vez se verifican los supuestos se evalúan las siguientes hipótesis mediante el análisis de varianza.

H_0: F-Score submuestreo = F-Score grupo control
H_a: F-Score submuestreo ≠ F-Score grupo control

La prueba de hipótesis anterior corresponde a la comparación del tratamiento 1 con el grupo control.

H_0: F-Score sobremuestreo = F-Score grupo control
H_a: F-Score sobremuestreo ≠ F-Score grupo control

La prueba de hipótesis anterior corresponde a la comparación del tratamiento 2 con el grupo control.

Una vez evaluadas las pruebas de hipótesis, se procederá a determinar en cuál de los grupos se presenta un F-Score mayor mediante la diferencia en los F-Score promedio obtenidos. Una vez realizado los análisis se procederá a redactar las conclusiones obtenidas.

4.5Fase de comparación Word2Vec y TF-IDF

La fase 5 consiste en repetir la fase 2, 3 y 4 cambiando como método para vectorizar palabras Word2Vec por TF-IDF. El objetivo principal en esta fase es comparar el valor del F-Score obtenido de un clasificador de texto entre un método de aprendizaje supervisado (Word2Vec) y un método de aprendizaje no supervisado (TF-IDF) para vectorizar palabras.

5. Resultados

En esta sección se presentan los resultados obtenidos siguiendo las fases explicadas en la sección de metodología. Los resultados se presentan con el objetivo de comparar las técnicas de balanceo de clases y el conjunto de datos sin balanceo. El conjunto de datos usado en el Inter TASS 2018 se compone de 4 clases: N, P, NEU y NONE. El cuadro 4 muestra la distribución porcentual de los comentarios usados.

Cuadro 4. *Distribución porcentual comentarios según categoría sentimiento*

Categoría	Cantidad	%
N	1.406	34%
P	1.123	27%
NONE	1.023	25%
NEU	562	14%
Total	**4.114**	**100%**

Inicialmente se consideró unir las categorías NEU y NONE, sin embargo, para efectos del objetivo de la investigación se mantiene la distribución porcentual. Previo a la ejecución del experimento se ejecuta un pequeño ensayo para determinar el tamaño de muestra necesario. En el cuadro 5 se detalla los valores obtenidos del F1-Score al ejecutar tres veces el clasificador para cada uno de los grupos o tratamientos (incluido el grupo control).

Cuadro 5. *Valor parámetros usados en función power.anova.test*

Grupo	F1-Score Promedio	Varianza Entre Grupos	Varianza Intra Grupos	Nivel de Significancia	Tamaño de Muestra por Grupo	Poder de la Prueba
Control	0,37	0,0049	0,0044	5%	**5**	**75%**
Submuestreo	0,26					
Sobremuestreo	0,39					

Para poder determinar diferencias entre los grupos y mantener un poder de prueba al 75% es necesario obtener 5 observaciones para cada uno de los grupos, es decir, una muestra de 15 observaciones en total. Como resultado del ensayo se escogen un total 15 semillas las cuales se aleatorizan para cada uno de los grupos en el experimento. Se escogen tres grupos de semillas: de la 1 a la 5, de la 11 a la 15 y de la 21 a la 25. En el cuadro 6 se detalla la asignación aleatoria de las semillas a cada grupo del experimento.

Cuadro 6. *Distribución semillas según grupo*

Grupo	Submuestreo	Sobremuestreo	Grupo Control
Semillas	14, 25, 24, 5 y 1	22, 21, 12, 11 y 15	13, 3, 2, 4 y 23

La figura 9 muestra el comportamiento del F1-Score promedio para cada una de las semillas usadas en el experimento. En general se puede apreciar que entrenar el clasificador usando el método de sobremuestreo (*SMOTE*) logra un F1-Score mayor en comparación al grupo control y a la técnica de submuestreo. Por otro lado, la técnica de submuestreo es la que evidencia tener un F1-Score menor, por lo que el clasificador es de menor calidad.

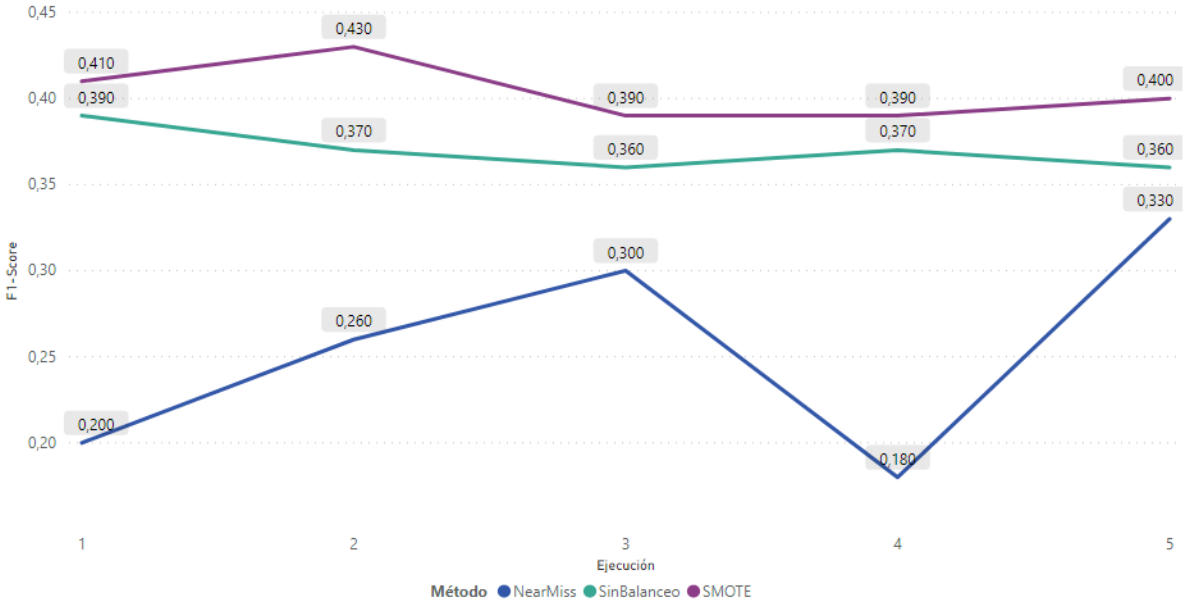

Figura 9. *F1-Score según grupo por ejecución*

En el cuadro 7 se detalla el resultado del análisis de poder del experimento. Los resultados obtenidos se asemejan mucho a los que se obtuvieron con el experimento previo. La diferencia encontrada se debe a que para el tratamiento de submuestreo se encontró una mayor variabilidad

Cuadro 7. *Análisis poder de prueba para análisis de varianza*

Grupo	F1-Score Promedio	Varianza Entre Grupos	Varianza Intra Grupos	Nivel de Significancia	Tamaño de Muestra por Grupo	Poder de la Prueba
Control	0,37	0,0062	0,0057	5%	**5**	**74%**
Submuestreo	0,25					
Sobremuestreo	0,40					

A continuación, se presentan los resultados de los experimentos divididos en dos secciones. La primera compara el grupo control con la técnica *NearMiss* para submuestreo. La segunda sección compara el grupo control con la técnica *SMOTE* para sobremuestreo. En ambas secciones se evidencia las pruebas estadísticas para determinar si se cumple o no con los supuestos para un análisis de varianza de una vía.

5.1 Grupo Control vs Submuestreo

Según el cuadro 8 el valor p evidencia la suficiente información estadística para no rechazar el supuesto de que los errores del modelo siguen una distribución normal, en la figura 10 se puede observar el comportamiento mediante un gráfico de normalidad.

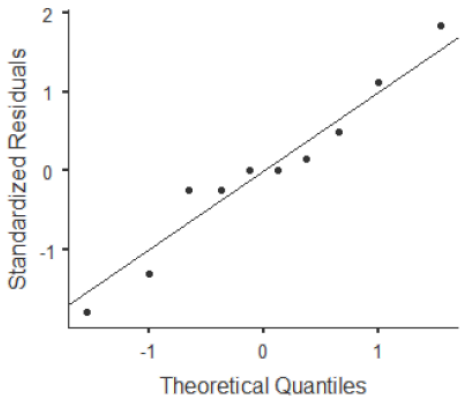

Figura 10. *Gráfico probabilidad normal de residuos grupo control vs submuestreo*

En el caso para el supuesto de igualdad en las varianzas el valor p para la prueba de Levene hay evidencia estadística suficiente para rechazar la hipótesis de igualdad de varianzas.

Cuadro 8. *Resultado pruebas estadísticas para supuestos del análisis de varianza (submuestreo)*

Prueba	Supuesto	Estadístico	Valor p
Shapiro-Wilk	Normalidad errores	W = 0.959	0,769*
Levene	Igualdad varianza	F = 10,7; gl1 = 1; gl2 = 8	0,011*

Nota: *5% de significancia

Debido a que el supuesto de Igualdad de varianzas no se cumple se utiliza el análisis de varianza de Welch para varianzas diferentes. La prueba estadística para el análisis de varianza refleja un valor p de 0,014, por lo que es suficiente evidencia estadística para rechazar la hipótesis nula de que el F1-Score promedio es igual en el grupo control y el grupo al que se le aplicó submuestreo. Finalmente, la diferencia promedio es de 0,116 a favor del grupo control. El cuadro 9 muestra la diferencia y el valor p el cual confirma la existencia de diferencias entre los grupos.

Cuadro 9. *Prueba Post-Hoc Games-Howell para el F-Score*

	Indicador	Grupo Control
NearMiss	Diferencia promedio	-0,116
	Valor-p	0,014

5.2 Grupo Control vs Sobremuestreo

Según el cuadro 10 el valor p evidencia la suficiente información estadística para no rechazar el supuesto de que los errores del modelo siguen una distribución normal, en la figura 11 se puede observar el comportamiento mediante un gráfico de normalidad.

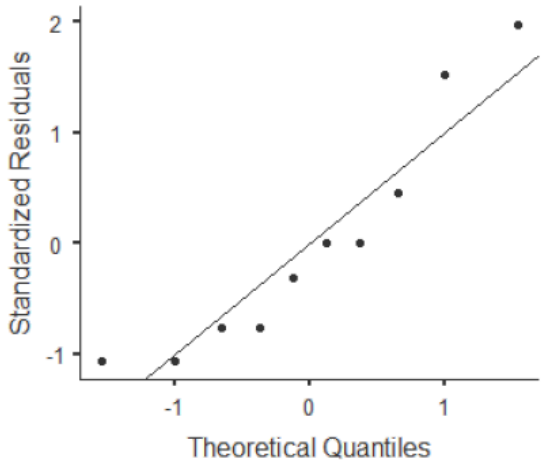

Figura 11. *Gráfico probabilidad normal de residuos grupo control vs sobremuestreo*

En el caso para el supuesto de igualdad en las varianzas el valor p para la prueba de Levene hay evidencia estadística suficiente para no rechazar la hipótesis de igualdad de varianzas.

Cuadro 10. *Resultado pruebas estadísticas para supuestos del análisis de varianza (sobremuestreo)*

Prueba	Supuesto	Estadístico	Valor p
Shapiro-Wilk	Normalidad errores	W = 0.882	0,138*
Levene	Igualdad varianza	F = 0,793; gl1 = 1; gl2 = 8	0,399*

Nota: *5% de significancia

El análisis de varianza realizado a estos grupos da como resultado un valor p del 0,006 que, comparado con el 5% de significancia, es evidencia estadística suficiente para determinar que el promedio de ambos grupos es diferente. La diferencia promedio entre ambos grupos es de 0,0340 a favor del grupo tratado con sobremuestreo. El cuadro 11 muestra la diferencia y el valor p el cual confirma la existencia de diferencias entre los grupos.

Cuadro 11. *Prueba Post-Hoc Games-Howell para el F-Score*

	Indicador	Grupo Control
SMOTE	Diferencia promedio	0,0340
	Valor-p	0,007

5.3Resultados con TF-IDF

En esta sección se presentan los resultados del experimento para el método TF-IDF para vectorizar palabras. En la figura 12 se presentan los gráficos QQ para verificar el supuesto de normalidad en los residuos del modelo de análisis de varianza. En ambos gráficos se determina un patrón normal en los residuos.

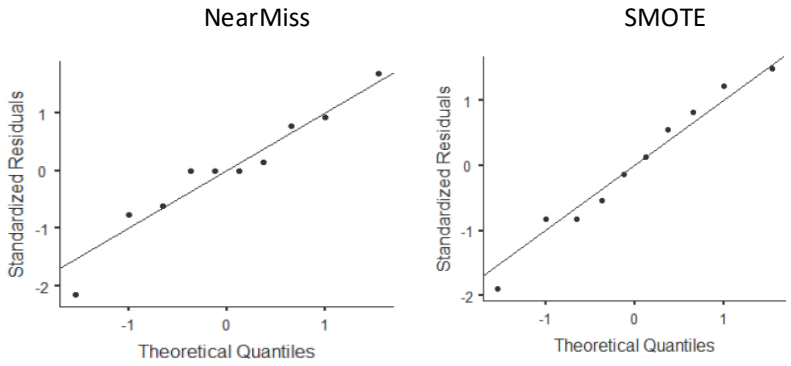

Figura 12. *Gráficos normalidad residuos según tratamiento vs grupo control*

En el cuadro 12 se muestran los resultados de las pruebas formales para verificar el supuesto de normalidad en residuos, así como el supuesto en igualdad de variables.

Cuadro 12. Resultado pruebas para supuestos análisis de varianza según tratamiento vs grupo control

Tratamiento	Prueba	Estadístico	Valor-p
NearMIss	Shapiro-Wilk	W=0,954	0,712
	Levene	F=3,84; gl1=1; gl2=8	<0.086
SMOTE	Shapiro-Wilk	W=0,972	0,910
	Levene	F=0,549; gl1=1; gl2=8	0,480

Con un 5% de significancia para ambos tratamientos se cumple el supuesto de normalidad. En el caso de igualdad de varianza solamente el tratamiento con sobremuestreo cumple el supuesto. Por esta razón, para comparar tratamiento de submuestreo con el grupo control se realiza un análisis de varianza de Welch, el cual es no paramétrico.

Cuadro 13. *Análisis de varianza para el F-Score según tratamiento vs grupo control*

Tratamiento	F	gl1	Gl2	Valor-p
NearMiss	40,0 (Welch)	1	5,06	0,001
SMOTE	192	1	8	<0.001

Al igual que en el caso con Word2Vec, el cuadro 13 muestra que en ambas comparaciones existen diferencias. En el caso del submuestreo con el grupo control existen diferencias estadísticamente justificables. Lo mismo ocurre en el caso del sobremuestreo con el grupo control. En el cuadro 14 se detalla el resultado de las diferencias promedio.

Cuadro 14. *Prueba Post-Hoc Games-Howell para el F1-Score según tratamiento vs grupo control*

Tratamiento	Diferencia promedio	Valor-p
NearMiss	-0,0580	0,001
SMOTE	0,144	<0.001

Según las pruebas post-hoc hay evidencia estadística suficiente para determinar que el F-Score del grupo control es diferente al de los tratamientos. En comparación con el submuestreo el grupo control destaca por entrenar un mejor clasificador. Para el caso de sobremuestreo las diferencias son a favor del tratamiento.

Por último, el cuadro 15 muestra el poder del experimento realizado para el método TF-IDF.

Cuadro 15. *Análisis poder de prueba para análisis de varianza*

Grupo	F1-Score Promedio	Varianza Entre Grupos	Varianza Intra Grupos	Nivel de Significancia	Tamaño de Muestra por Grupo	Poder de la Prueba
Control	0,41	0,0108	0,0079	5%	**5**	**84%**
Submuestreo	0,35					
Sobremuestreo	0,55					

Para el caso del TF-IDF se alcanza un poder en la prueba del 84%, es decir, la probabilidad de cometer el error tipo 2 es del 16% con 5% de significancia. En comparación al método de Word2Vec, estadísticamente se obtienen resultados más concluyentes y confiables. También se puede apreciar mediante el F1-Score promedio que F-Score mejora para cada uno de los grupos.

Esto se debe a que el método Word2Vec se basa en un algoritmo de aprendizaje supervisado el cual también puede verse afectado por las clases desbalanceadas. Mientras que el método TF-IDF, al ser un método de aprendizaje no supervisado puede capturar más información valiosa para el clasificador debido a que se basa en las palabras incluidas en un conjunto de documentos.

Un último resultado de la investigación realizada es, habiendo analizado el análisis de varianza para cada una de las comparaciones, el cambio en la cantidad de observaciones para cada clase aplicando los diferentes métodos de balanceo. Este resultado es generalizable para el corpus usado en el experimento.

Con respecto al resultado obtenido usando el método *NearMiss,* la baja calificación del F-Score puede deberse a la pérdida de información en las clases con mayor número de observaciones, ya que cada clase tendría 562 comentarios. Usando un 70% de los datos para entrenar el clasificador cada clase pasaría a tener 393 comentarios. En el cuadro 16 se detalla la disminución porcentual para cada clase.

Cuadro 16. *Cambio en la distribución de las clases según grupo*

Clase	Original	Submuestreo	Sobremuestreo
N	1.406	562 (-60,03%)	1.406(0,00%)
P	1.123	562(-49,95%)	1.406(25,20%)
NONE	1.023	562(-45,06%)	1.406(37,44%)
NEU	562	562(0%)	1.406(150,18%)
Total	**4.114**	**2.248**	**5.624**

Por otra parte, si se usa el sobremuestreo se tendría el efecto contrario, es decir, se ganaría información para las clases con menor cantidad de observaciones. Cada clase tendría 1.406 observaciones, 984 observaciones por clase para entrenar el clasificador.

6. Conclusiones y trabajo futuro

En primer lugar, se detallan las conclusiones más importantes para cada uno de los objetivos específicos planteados para explicar la conclusión asociada al objetivo general. Por último, se detallan los puntos que se deben de trabajar a futuro de acuerdo con los hallazgos en los resultados.

- La primera conclusión de esta investigación es que debido a la normalización de texto en el preprocesamiento de los datos permite construir, en este caso, dos modelos de vectorización de palabras lo suficientemente eficaces para entrenar un clasificador de texto.
- Con respecto a las técnicas de balanceo de datos, la revisión de literatura permite establecer el método *NearMiss* para balancear las clases mediante submuestreo y el método *SMOTE* para balancear las clases mediante sobremuestreo. Ambos métodos se basan bajo el mismo concepto de vecinos cercanos por lo cual permite aumentar la validez del experimento.
- Utilizar como factores fijos los valores por defecto de los diferentes algoritmos usados en este proyecto permitió que la comparación de los resultados obtenidos se viera menos afectada por el ruido que implica cambiar alguno de estos parámetros.
- Para el último objetivo específico, se logra determinar mediante el análisis de varianza que existen diferencias a nivel estadístico en el F-Score obtenido usando los métodos de submuestreo y sobremuestreo para balancear las clases.
- Las diferencias encontradas evidencian que el F-Score obtenido balanceando las clases con submuestreo es menor al F-Score del grupo al que no se le aplica el balanceo de clases. Por el contrario, el F-Score obtenido balanceando las clases con sobremuestreo es mayor al F-Score del grupo al que no se le aplica el balanceo de clases.
- Un último hallazgo importante es que el F-Score obtenido, tanto en los grupos a los que se les aplicó los tratamientos como el grupo control, es mayor usando el método TF-IDF

para vectorizar palabras en comparación a si se vectorizan las palabras usando Word2Vec.

De acuerdo con los hallazgos obtenidos en esta investigación, se plantean los siguientes puntos como trabajo futuro que ayude a reforzar la validez de los resultados obtenidos.

- En primer lugar, es utilizar un corpus lingüístico más grande ya que el corpus del interTASS es un corpus relativamente pequeño en cantidad de comentarios.
- El segundo punto importante por llevar a cabo es probar con diferentes tamaños de muestra ya que, según los resultados, usar submuestreo disminuye la cantidad total de observaciones, por lo cual es importante medir si el tamaño de muestra puede estar relacionado a que el F-Score sea menor.
- Otro punto importante para evaluar en un experimento futuro es utilizar diferentes algoritmos de submuestreo y sobremuestreo para determinar el efecto sobre el calculo del F-Score obtenido en el clasificador.
- Por último, entrenar el clasificador aplicando validación cruzada para optimizar los parámetros de entrada del modelo usado para asignar las clases al recibir datos nuevos.

Bibliografía

Adams, B., & McKenzie, G. (2018). Crowdsourcing the character of a place: Character-level convolutional networks for multilingual geographic text classification. *Transactions in GIS*, 394-408. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1111/tgis.12317

Ahmed, H., Traore, I., & Saad, S. (2017). Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1:e9. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1002/spy2.9

Albitar, S., Fournier, S., & Espinasse, B. (2014). *An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification.* Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/content/pdf/10.1007%2F978-3-319-11749-2_8.pdf

Angulo, C. (2019). *Repositorio Kérwá Universidad de Costa Rica.* Obtenido de http://repositorio.ucr.ac.cr/bitstream/handle/10669/79814/Desarrollo%20de%20representaciones%20vectoriales%20de%20palabras%20Costa%20Rica.pdf?sequence=1&isAllowed=y

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238-247.

Cárdenas, J. P., Olivares, G., & Alfaro, R. (2014). *Clasificación automática de textos usando redes de palabras.* Obtenido de https://www.redalyc.org/articulo.oa?id=157032730001

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artifical intelligence research*, 321-357.

Chen, H., Mckeever, S., & Delany, S. J. (2016). Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media. *Springer, Cham*, vol 513. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/chapter/10.1007/978-3-319-46562-3_12

Colas, F., & Brazdil, P. (2006). *Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks .* Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/content/pdf/10.1007%2F978-0-387-34747-9_18.pdf

Contreras, M. (2016). *Minería de texto en la clasificación de materialbibliográfico.* Obtenido de https://dialnet.unirioja.es/descarga/articulo/5733173.pdf

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Obtenido de Research Gate: https://www.researchgate.net/profile/Andrea_Dal_Pozzolo/publication/283349138_Calibrating_Probability_with_Undersampling_for_Unbalanced_Classification/links/5636

06c308ae88cf81bcd9f1/Calibrating-Probability-with-Undersampling-for-Unbalanced-Classification.

Dubiau, L., & Ale, J. M. (2013). Análisis de Sentimientos sobre un Corpus en Español: Experimentación con un Caso de Estudio. *Argentine Symposium on Artificial Intelligence*.

Elmarhoumy, M., Fattah, M., Suzuki, M., & Ren, F. (2013). A new modified centroid classifier approach for automatic text classification. *IEEJ Trans Elec Electron Eng*, 8: 364-370. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1002/tee.21867

Filiberto, Y., Caballero, Y., Bello, R., & Frías, M. (2011). *Método para el aprendizaje de reglas de clasificación para conjuntos de datos no balanceados.* Obtenido de https://www.redalyc.org/articulo.oa?id=378343674002

Glen, S. (2016). *StatisticsHowTo.com*. Obtenido de https://www.statisticshowto.com/welchs-anova/

Han, J., Zuo, W., Liu, L., Xu, Y., & Peng, T. (2016). Building text classifiers using positive, unlabeled and 'outdated' examples. *Concurrency Computat.: Pract. Exper.*, 28: 3691–3706. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1002/cpe.3879

Hassanzadeh, H., Groza, T., Nguyen, A., & Hunter, J. (2014). Load Balancing for Imbalanced Data Sets: Classifying Scientific Artefacts for Evidence Based Medicine. *PRICAI 2014: Trends in Artificial Intelligence*, Lecture Notes in Computer Science, vol 8862. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/chapter/10.1007/978-3-319-13560-1_84

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence ICAI.*

Kanaan, G., Al-Shalabi, R., Ghwanmeh, S., & Al-Ma'adeed, H. (2009). A comparison of text-classification techniques applied to Arabic text. *J. Am. Soc. Inf. Sci.*, 60: 1836-1844. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1002/asi.20832

Krawczyk, B., McInnes, B. T., & Cano, A. (2017). Sentiment Classification from Multi-class Imbalanced Twitter Data Using Binarization. *HAIS 2017*, vol 10334. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/chapter/10.1007/978-3-319-59650-1_3

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models.* New York: McGraw-Hill Irwin.

Lemaitre, G., Nogueira, F., & Aridas, C. (2017). *imbalanced-learn*. Obtenido de https://imbalanced-learn.org/stable/about.html

Li, Y., Wang, J., Wang, S., Liang, J., & Li, J. (2019). Local dense mixed region cutting + global rebalancing: a method for imbalanced text sentiment classification. *Int. J. Mach. Learn. & Cyber*, 10: 1805. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/article/10.1007/s13042-018-0858-x

Liu, P., Chen, W., Ou, G., Wang, T., Yang, D., & Lei, K. (2014). Sarcasm Detection in Social Media Based on Imbalanced Classification. *WAIM 2014*, vol 8485. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/chapter/10.1007/978-3-319-08010-9_49

Liu, R.-L. (2009). Context recognition for hierarchical text classification. *J. Am. Soc. Inf. Sci.*, 60: 803-813. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1002/asi.21022

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). *AAAI Conference on Artificial Intelligence.* Obtenido de https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9314/9535

Martínez-Cámara, E., Almeida-Cruz, Y., Días-Galiano, M. C., Estévez-Velarde, S., García-Cumbreras, M. Á., García-Vega, M., . . . Piad-Morffis, A. (2018). Overview of TASS 2018: Opinions, Health and Emotions. *Workshop on Semantic Analysis at SEPLN*, 13-27.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Obtenido de https://arxiv.org/pdf/1301.3781.pdf

Montgomery, D. (2012). *Design and analysis of experiments.* Wiley.

Pérez de Celis, C., Ronquillo, F., Sierra, G., & Salceda, E. (2014). *Creación y uso de una ontología relacionada con genes, síndromes, síntomas y enfermedades para la clasificación de textos biomédicos.* Obtenido de https://www.redalyc.org/articulo.oa?id=157029689005

Pramokchon, P., & Piamsa-nga, P. (2014). Reducing Effects of Class Imbalance Distribution in Multi-class Text Categorization. En S. Boonkrong, H. Unger, & P. Meesad, *Recent Advances in Information and Communication Technology* (págs. 263-272). Springer. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/content/pdf/10.1007%2F978-3-319-06538-0.pdf

Rashcka, S., & Mirjalili, V. (2017). *Python machine learning.* Packt Publishing Ltd.

Romero, Iglesias, & Borrajo. (2012). A Comparative Analysis of Balancing Techniques and Attribute Reduction Algorithms. *6th International Conference on Practical Applications of Computational Biology & Bioinformatics. Advances in Intelligent and Soft Computing*,

vol 154. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/chapter/10.1007/978-3-642-28839-5_10

Santana, P., Rossana, C., & Missio, D. (2014). *Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos.* Obtenido de https://www.redalyc.org/articulo.oa?id=92530455007

Schlegel, A. (2016). *RPubs*. Obtenido de https://rpubs.com/aaronsc32/games-howell-test

SEPLN, T. d. (14 de 06 de 2020). *TASS: Workshop on Semantic Analysis at SEPLN*. Obtenido de http://tass.sepln.org/tass_data/download.php

Shapiro, S. S., & Wilk, M. B. (1965). *JSTOR.* Obtenido de https://www.jstor.org/stable/2333709

Sierra, G. E. (2015). *Introducción a los corpus lingüísticos.* México: Instituto de Ingeniería Universidad Nacional Autónoma de México.

Soto, C., & Jiménez, C. (2011). *Aprendizaje supervisado para la discriminación y clasificación difusa.* Obtenido de https://www.redalyc.org/articulo.oa?id=49622390003

Venegas, R. (2007). *Clasificación de textos académicos en función de su contenido léxico-semántico.* Obtenido de https://www.redalyc.org/articulo.oa?id=157013772012

Wang, Y., & Zhu, L. (2019). Research on improved text classification method based on combined weighted model. *Concurrency Computat Pract Exper*, e5140. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1002/cpe.5140

Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., & Wang, X. (2015). Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification. *Cogn Comput*, 7: 226. Obtenido de https://link-springer-com.ezproxy.sibdi.ucr.ac.cr/article/10.1007/s12559-015-9319-y

Zhang, J., & Mani, I. (2003). Obtenido de http://www.site.uottawa.ca/~nat/Workshop2003/jzhang.pdf?attredirects=0

Zhuo, Y., Tong, Y., Gu, R., & Gall, H. (2016). Combining text mining and data mining for bug report classification. *J. Softw. Evol. and Proc.*, 28: 150– 176. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/10.1002/smr.1770

Zu, G., Ohyama, W., Wakabayashi, T., & Kimura, F. (2005). Automatic text classification of english newshire articles based on statistical classification techniques. *Electrical Engineering in Japan*, Vol. 124-C, No. 3, 852–860. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/epdf/10.1002/eej.20108