UNIVERSIDAD PRIVADA DE TACNA

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

"ANÁLISIS DE COMPORTAMIENTO DE TWEETS RELACIONADOS A LA PANDEMIA DE COVID-19 EN EL PERÚ EMPLEANDO MINERÍA DE TEXTOS, 2022"

PARA OPTAR:

TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

PRESENTAD POR:

Bach. ROSALIA INES MAMANI LLACA

TACNA – PERÚ

2022

UNIVERSIDAD PRIVADA DE TACNA

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

"ANÁLISIS DE COMPORTAMIENTO DE TWEETS RELACIONADOS A LA PANDEMIA DE COVID-19 EN EL PERÚ EMPLEANDO MINERÍA DE TEXTOS, 2022"

Tesis sustentada y aprobada el 25 de junio de 2022, estando el jurado calificador integrado por:

PRESIDENTE : Dra. MARTHA JUDITH PAREDES VIGNOLA

SECRETARIO : Mag. PATRICK JOSÉ CUADROS QUIROGA

VOCAL : Mag. RICARDO EDUARDO VARCÁRCEL ALVARADO

ASESOR : MSc. HUGO MANUEL BARRAZA VIZCARRA

DECLARACIÓN DE ORIGINALIDAD

Yo, Rosalia Ines Mamani Llaca identificada con documento de identidad 45829671 en calidad de: Bachiller en Ingeniería de Sistemas de la Escuela Profesional de Ingeniería de Sistemas.

Declaro bajo juramento:

Soy autor de la tesis titulada: *"Análisis de Comportamiento de tweets relacionados a la pandemia de COVID-19 en el Perú empleando Minería de Textos, 2022"* la cual presento para optar el Título profesional de Ingeniero de Sistemas.

1. La tesis no ha sido plagiada ni total ni parcialmente, para lo cual se han respetado las normas internacionales de citas y referencias para las fuentes consultadas.
2. La tesis presentada no atenta contra derechos de terceros.
3. La tesis no ha sido publicada ni presentada anteriormente para obtener algún grado académico previo o título profesional.
4. Los datos presentados en los resultados son reales, no han sido falsificados, ni duplicados, ni copiados.

Por lo expuesto, mediante la presente asumo frente a la universidad cualquier responsabilidad que pudiera derivarse por la autoría, originalidad y veracidad presentada. En consecuencia, me hago responsable frente a la universidad y a terceros, de cualquier daño que pudiera ocasionar, por el incumplimiento de lo declarado o que pudiera encontrar como causa del trabajo presentado, asumiendo todas las cargas pecuniarias que pudieran derivarse de ello en favor de terceros con motivo de acciones, reclamaciones o conflictos derivados del incumplimiento de lo declarado o las que encontrasen causa en el contenido de la tesis, libro y/o invento.

De identificarse fraude, piratería, falsificación o que el trabajo de investigación haya sido publicado anteriormente; asumo las consecuencias y sanciones que de mi acción se deriven, sometiéndome a la normatividad vigente de la Universidad Privada de Tacna.

Tacna, 15 de mayo de 2022.

Bach. Rosalia Ines Mamani Llaca

DNI: 45829671

DEDICATORIA

A todas las personas que no se rinden y cumplen con sus sueños, metas y objetivos que uno se plantea en la vida con el fin de obtener el éxito.

Bach. Rosalia Ines Mamani Llaca

AGRADECIMIENTO

A Dios por cada día y por cada oportunidad.
A mi asesor a cargo MSc. Hugo Manuel Barraza Vizcarra ya que sin él no hubiera realizado mi tesis.
A la Escuela de Ingeniería de Sistemas de la Universidad Privada de Tacna por darnos la oportunidad de lograr nuestros objetivos.

Bach. Rosalia Ines Mamani Llaca

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

Un evento trascendental como es la COVID-19 genera bastante expectativa por parte de actores nacionales e internacionales lo que genera la producción de bastante contenido en internet y en redes sociales; las redes sociales son en la actualidad grandes repositorios de las opiniones de los usuarios a nivel mundial. A raíz de esta abundante información es que se planteó realizar la investigación del comportamiento de *tweets* relacionados a la pandemia COVID-19 en el Perú empleando minería de textos, 2022.Con respecto al entrenamiento de los *tweets* de la pandemia COVID-19, el algoritmo Máquinas de Vectores de Soporte consiguió una exactitud del 93,3 %, clasificando un 70 % como *tweets* positivos y 30 % como negativos. Luego de haber aplicado el modelo entrenado a 5000 *tweets* se consiguió el 87 % *tweets* positivos y 13 % negativos.

Palabra clave: Análisis de sentimientos, minería de textos, Maquina de vectores de soporte, pandemia Perú, dosis, vacuna, COVID-19.

ABSTRACT

A transcendental event such as COVID-19 generates a lot of expectation on the part of national and international actors, therefore, regarding the abundant information on the internet and on social networks, social networks are currently great repositories of user opinions. a world level. As a result of this abundant information, it was proposed to carry out the investigation of the behavior of tweets related to the COVID-19 pandemic in Peru using text mining, 2022.Regarding the training of the tweets of the COVID-19 pandemic, the Support Vector Machines algorithm achieved an Accuracy of 92,8 %, achieving 70 % positive tweets and 30 % negative ones. After having altered the model and the 5,000 tweets, the Accuracy of 36,875 % was obtained, achieving 87 % positive tweets and 13 % negative ones.

Keywords: Sentiment analysis, text mining, Support Vector Machine, Peru pandemic, dose, vaccine, COVID-19.

INTRODUCCIÓN

El avance de las tecnologías de la información que existe hoy en día exige a las personas actualizarse y afrontar los desafíos el cual nos permite descubrir y analizar.

En este trabajo de investigación se ha desarrollado el análisis de comportamiento de tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos. El presente trabajo está estructurado en cinco (5) capítulos a continuación se presentan.

Capítulo I Planteamiento del problema, se describen el porqué de la elaboración, y los objetivos planteados.

Capítulo II Marco teórico, se narra la base teórica que está relacionada con la investigación.

Capítulo III Marco metodológico, se narra el tipo y diseño de investigación y el análisis de datos realizado.

Capítulo IV Resultados, se narra los resultados conseguidos a través de figuras en donde se describe la gráfica del comportamiento de tweets.

Capítulo V Discusión, se describe la discusión obtenida.

Por último, se hace la presentación de las conclusiones y recomendaciones y referencias bibliográficas y anexos de matriz de consistencia del trabajo de investigación

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción del problema

En el Perú, en marzo del 2020, llegó la COVID-19, encontrando un sistema de salud deficiente, poco integrado y con varias limitaciones para el diagnóstico, la vigilancia e incluso el manejo de los fallecidos. Esto generó una crisis sanitaria, social y hasta política. La percepción del manejo de esta crisis se vio expresada en los medios de comunicación, proclamas de distintas organizaciones civiles y también en distintas redes sociales, como por ejemplo Twitter, en donde cualquier ciudadano puede expresar sus opiniones mediante “tweets”. Las redes sociales son muy usadas a fin de medir las reacciones del público hacia un evento en específico.

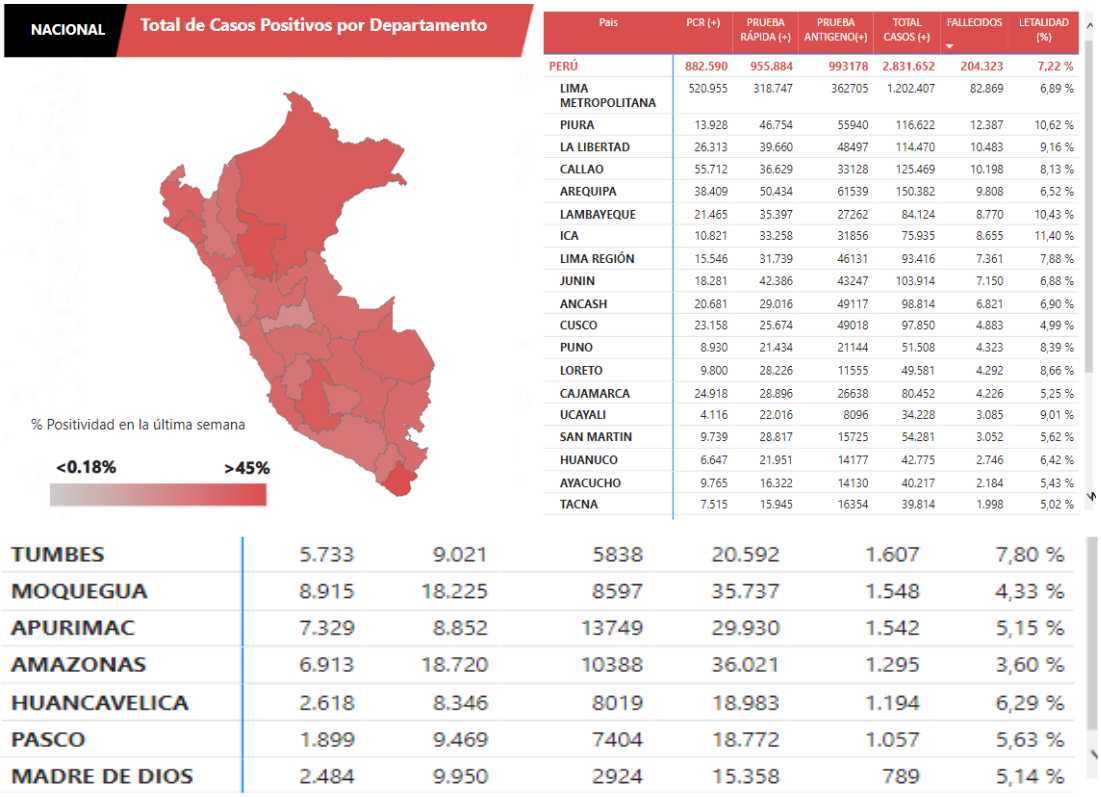
La minería de textos es una herramienta que se puede usar para el análisis de sentimientos en redes sociales. Sabiendo que en las redes sociales se encuentra abundante cantidad de información disponible para poder analizar, el uso de técnicas de análisis de textos puede producir nuevo conocimiento provechoso para la toma de decisiones.

Diariamente los medios de comunicación, como por ejemplo los sistemas de noticias tradicionales, y los usuarios de internet publican una gran cantidad de información relacionados a la pandemia de COVID-19 en sus medios sociales, es por esto que es muy importante el uso de herramientas de análisis de información para que los datos almacenados en bases de datos y en internet sean correctamente aprovechados, de tal manera que puedan servir para encontrar patrones de conducta, intenciones o tendencias, y esto contribuya a mejorar la toma de decisiones, tal como se puede apreciar en la Figura 1.

A nivel nacional, según el portal web del Ministerio de Salud (MINSA), en los últimos reportes del 23 de enero del 2022 se muestran el total de casos positivos por departamento; en el Perú hay un total de 2,831,652 de casos positivos (+); 204,323 fallecidos y una letalidad (%) del 7,22 %.

Figura 1.

Total, de casos positivos por departamento



País	PCR (+)	PRUEBA RÁPIDA (+)	PRUEBA ANTIGENO(+)	TOTAL CASOS (+)	FALLECIDOS	LETALIDAD (%)
PERÚ	882.590	955.884	993178	2.831.652	204.323	7,22 %
LIMA METROPOLITANA	520.955	318.747	362705	1.202.407	82.869	6,89 %
PIURA	13.928	46.754	55940	116.622	12.387	10,62 %
LA LIBERTAD	26.313	39.660	48497	114.470	10.483	9,16 %
CALLAO	55.712	36.629	33128	125.469	10.198	8,13 %
AREQUIPA	38.409	50.434	61539	150.382	9.808	6,52 %
LAMBAYEQUE	21.465	35.397	27262	84.124	8.770	10,43 %
ICA	10.821	33.258	31856	75.935	8.655	11,40 %
LIMA REGIÓN	15.546	31.739	46131	93.416	7.361	7,88 %
JUNIN	18.281	42.386	43247	103.914	7.150	6,88 %
ANCASH	20.681	29.016	49117	98.814	6.821	6,90 %
CUSCO	23.158	25.674	49018	97.850	4.883	4,99 %
PUNO	8.930	21.434	21144	51.508	4.323	8,39 %
LORETO	9.800	28.226	11555	49.581	4.292	8,66 %
CAJAMARCA	24.918	28.896	26638	80.452	4.226	5,25 %
UCAYALI	4.116	22.016	8096	34.228	3.085	9,01 %
SAN MARTIN	9.739	28.817	15725	54.281	3.052	5,62 %
HUANUCO	6.647	21.951	14177	42.775	2.746	6,42 %
AYACUCHO	9.765	16.322	14130	40.217	2.184	5,43 %
TACNA	7.515	15.945	16354	39.814	1.998	5,02 %
TUMBES	5.733	9.021	5838	20.592	1.607	7,80 %
MOQUEGUA	8.915	18.225	8597	35.737	1.548	4,33 %
APURIMAC	7.329	8.852	13749	29.930	1.542	5,15 %
AMAZONAS	6.913	18.720	10388	36.021	1.295	3,60 %
HUANCAVELICA	2.618	8.346	8019	18.983	1.194	6,29 %
PASCO	1.899	9.469	7404	18.772	1.057	5,63 %
MADRE DE DIOS	2.484	9.950	2924	15.358	789	5,14 %

Nota. El total de casos positivos corresponde a los 26 departamentos del país Perú. Fuente: (MINSA, 2022).

1.2 Formulación del Problema

1.2.1 Problema General

¿Cómo es el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos,2022?

1.2.2 Problemas Específicos

a. ¿Cómo clasificar el comportamiento de los tweets relacionados a la pandemia de COVID-19?
b. ¿Cuál es el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos?

1.3 Justificación e importancia

Actualmente se evidencia la falta de concientización que hay en el Perú respecto a la importancia de encontrar información verídica respecto al tema de COVID-19 lo cual termina causando confusión en los ciudadanos y estos pueden adoptar medidas incorrectas debido a la mala calidad de la información que reciben.

La presente investigación demuestra que la importancia de los comentarios de las redes sociales en la actualidad realizando el análisis del sentimiento de los tweets en el Perú relacionados a la pandemia COVID-19.

En efecto, se justifica la ejecución del proyecto puesto que facilitara la labor a Establecimientos de Salud y Centro de Salud, al Ministerio de Salud (MINSA) en la toma de decisiones.

1.3.1 Desde el punto de vista científico

Desde el punto de vista científico, en el presente trabajo usa técnicas computacionales para el análisis de sentimientos de textos y se validará la información generada por la técnica. Esto permitirá generar nuevo conocimiento científico que podrá ser utilizado en las futuras investigaciones.

1.3.2 Desde el punto de vista social

Desde el punto de vista social, los resultados obtenidos en el presente trabajo de investigación favorecerán a la sociedad, al Ministerio de Salud (MINSA), a los establecimientos de Salud o Centros de Salud, hogar, comunidad y otros facilitando la toma de decisiones.

El análisis de comportamiento de tweets beneficia al Ministerio de Salud, a los establecimientos de salud y centros de salud, en la que participan médicos y enfermeras. Al diario el comercio y las noticias del Perú.

Favorece a la sociedad a nivel local, nacional e internacional.

Las enfermeras también se ven beneficiados ya que ayuda a tomar decisiones.

1.3.3 Desde el punto de vista económico

Desde el punto de vista económico, el presente trabajo de investigación permite recoger la opinión de los ciudadanos relacionada a la pandemia COVID-19 que es abundante en las redes sociales, lo cual permitirá poder armar estrategias de concientización a la población. Esto permitirá que ahorrar en los altos costos que representa la contratación de un servicio de estudio de mercado, aprovechando la información que ya se encuentra disponible en internet.

1.4 Objetivos

1.4.1 Objetivo general

Analizar el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos, 2022.

1.4.2 Objetivos específicos

a. Elaborar un clasificador de sentimiento de los tweets relacionados a la pandemia de COVID-19.

b. Calcular el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos.

1.5 Hipótesis

1.5.1 Hipótesis General

Se podrá analizar el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos,2022.

1.5.2 Hipótesis Específicos

a. Se podrá clasificar el sentimiento de los tweets relacionados a la pandemia de COVID-19.
b. Se podrá calcular el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos.

CAPÍTULO II: MARCO TEÓRICO

2.1 Antecedentes del estudio

2.1.1 Antecedentes Internacionales:

En el estudio de (Condor Tinoco, Loa Navarro, Huarcaya Ccoicca y Castro Buleje, 2020)", en su Actas del III Congreso Internacional de Ingenieria de Sistemas "Minería de datos en Twitter: análisis del sentimiento del desempleo en la población hispanohablante en tiempos del COVID-19". En su trabajo de investigación tiene el propósito de conocer la opinión de la población y desarrollar un modelo que permita analizar el sentimiento relacionado al desempleo, los autores hicieron uso del método SEMMA (muestrea, explora, modela y evalúa) para el análisis del sentimiento. Y para el preprocesamiento se empleó la herramienta Orange Canvas. Como consecuencia, de los seis mil tweets relacionados al desempleo, la palabra "desempleo" se asocia con las palabras pandemia, trabajo, pobreza y país. Asimismo, se distinguió que los sentimientos de los usuarios son negativo y neutro.

En conclusión, en el trabajo de investigación señalado anteriormente, de los usuarios que hicieron los comentarios, la palabra con mayor frecuencia es "desempleo", de la cual se repiten 5998 veces, "pobreza" aparecen 704 veces, "pandemia" aparecen 637 veces, "país" aparecen 521 veces, "economía" aparece 468 veces y "trabajo" aparecen 408 veces, quiere decir que estas palabras aparecen en los comentarios de tweets. Para la búsqueda de los tweets se empleó la palabra "desempleo".

Los tweets se recolectaron los días 25, 26 y 27 de agosto del 2020, el resultado obtenido es del 79,21 % posee sentimiento neutral, seguido del 19,42 % posee sentimiento negativo y finalmente el 1,37 % posee sentimiento positivo.

En el estudio de (Flores González, Contreras M. y Andrade del Cid, 2020), en su artículo "Comportamiento de las comunidades digitales en Twitter durante las elecciones México 2018", se estudió el comportamiento de las comunidades digitales en las elecciones en la ciudad de México 2018. La recolección de los tweets se realizó utilizando el lenguaje de Programación "R" y a través de un filtro de un libro de códigos.

Como resultado se obtuvo para el candidato José Antonio Meade el 16

% de tweets positivos, el 20 % de tweets negativos y el 11 % de tweets neutrales sumando el 47 % en total. Y para el candidato Andrés Manuel López Obrador se emitieron el 10 % tweets de positivos, el 12 % tweets de negativos y el 7 % tweets de neutrales sumando el 29 % en total.

Por último, emitieron el 4 % tweets de positivos, el 13 % tweets de negativos, el 7 % tweets de neutrales sumando el 24 % en total para el candidato Ricardo Anaya.

Como conclusión para este artículo se detectó que mayormente los tweets relacionados al candidato José Antonio Meade tienen un sentimiento clasificado como negativo e igualmente para, Andrés Manuel López Obrador y Ricardo Anaya.

En el estudio de (Pineda Briseño y Chire Saire, 2020), en su artículo "Minería de texto para identificar las principales preocupaciones de los usuarios de Twitter durante COVID-19 en la Ciudad de México", explica el inicio del brote epidemiológico que aconteció en China. Esta enfermedad se ha esparcido velozmente a nivel mundial. Aprovechar las tecnologías de información es uno de los retos de la comunidad científica internacional con el propósito de monitorear el comportamiento de la población en tiempo real. Las personas comparten información de su salud, sus datos personales desde cualquier lugar del mundo en las redes sociales. En este estudio se hizo uso de la técnica de minería de texto. Se realizó la extracción, visualización y análisis de los tweets, concluyendo que las preocupaciones de la población están relacionados a la cantidad de casos confirmados y a las medidas de seguridad sanitaria que fueron implementadas por el Gobierno de México.

2.1.2 Antecedentes Nacionales:

En el estudio de (Paez Guarnizo y Monroy, 2020), en su tesis "Implementación de un modelo de análisis de sentimientos con respecto a la Jurisdicción Especial para la Paz basado en minería de datos en Twitter", proponen una metodología de seis fases. Lo primero es armar un conjunto de datos, lo que sigue es el preprocesamiento del Lenguaje Natural, para luego realizar la extracción de características, luego la Máquinas de Vectores de Soporte a través de cross-validation para el muestreo de datos. Por otro lado, para los algoritmos Random Forest y Naive Bayes, la segmentación de los

datos se realiza en porcentajes de 70,75 y 80 % y la prueba de datos se muestra en porcentajes de 30,25 % y 20 %, luego se realiza el entrenamiento y clasificación para cada uno de los algoritmos, por ultimo las técnicas implementadas de desempeño fueron evaluadas. En la última fase se muestra que el mejor método de clasificación de texto es el algoritmo Random Forest que obtiene como unas métricas de 74,56 % de precisión, el 70,15 % de recall y 68,10 % de F1-Score.

En el estudio de (Mamani Coaquira, J. Ibarra, Mamani Vilca, Ordoñez Ramos y Aquino Cruz, 2021) en su artículo denominado “Identificar Sentimientos en Cuarentena por la COVID-19 mediante Clasificador Léxico y Aprendizaje Supervisado”, se trabaja con un conjunto de tweets recolectados durante la pandemia para clasificar los sentimientos en negativos y positivos. Los resultados fueron las palabras más frecuentes: COVID-19, Cuarentena y Casa. Los términos “Ganar” y “Bueno” poseen un sentimiento positivo, el término “extraño” posee un sentimiento negativo. A través del algoritmo Máquinas de Vectores de Soporte y Clasificador Léxico de palabras se obtuvo el modelo de clasificación de sentimientos de precisión es del 91,5 %.

En el estudio de (Tapia Perales, Ruiz Montalvo y Chirinos Mundaca, 2014) en su proyecto denominado “Modelo de clasificación de opiniones subjetivas en redes sociales” utiliza la técnica de minería de datos para analizar la opinión favorable o desfavorable de los usuarios de las redes sociales que realizan publicaciones de un producto, servicio. Utiliza el algoritmo de clasificación redes bayesianas, emplean la metodología de dos etapas Crisp-DM para la minería de datos y XP para el desarrollo del prototipo de la red social. El modelo obtiene una eficiencia de 73,66 % y valida los comentarios reales con el cálculo del modelo.

2.1.3 Antecedentes locales:

En el estudio de (Poma Pancaya, 2020), en su proyecto “Análisis del comportamiento de los tweets, de la participación electoral de los principales candidatos a la presidencia del Perú utilizando minería de textos en el año 2016”, se realiza un análisis de comportamiento de los tweets relacionados a los dos candidatos presidenciales con mayor votación en el proceso electoral del año 2016 empleando el clasificador Naive Bayes.

Poma concluye que en la primera vuelta electoral para, Pedro Pablo

Kuczynski obtuvo el 86,8 % de exactitud y los tweets positivos son el 58,2 % y los tweets negativos son 41,8 %. Asimismo, el 47,3 % de tweets positivos y 52,7 % de tweets negativos para la segunda vuelta.

De la misma manera en primer lugar el 88,5 % de exactitud para, Keiko Fujimori, logrando tweets positivos del 52,4 % y negativos del 47,6 %. Del mismo modo el 47,5 % tweets positivos y el 52,5 % tweets negativos en la segunda vuelta.

2.2 Bases Teóricas

2.2.1 Minería de texto

Se explica la minería de texto reune diferentes técnicas propuestas en el campo de la recuperación de textos (Senso y Eíto Brun, 2004, p. 1).

Asimismo, la minería de textos indica a manera de una aplicación complementaria a la minería de datos que analiza y aspira identificar patrones en los datos almacenados en bases de datos relacionales y data warehouse (Senso y Eíto Brun, 2004, p. 2).

Esta segunda definición señala que ésta tiene como fin hallar información y conocimiento que antes no se conocia, y que no se mostraba en ninguno de los documentos analizados. Conforme con esta interpretación, la minería de textos es un proceso que aspira descubrir conocimiento (Senso y Eíto Brun, 2004, p. 2).

Inicialmente, la minería textual ayuda el análisis de abundantes conjuntos de textos que en efecto nos sería de muy difícil manejo a causa de su magnitud. De este modo, un científico lograra estudiar estos datos, conocer relaciones entre documentos y extraer conclusiones. sino que es una composicion de análisis humano y automatizado que sea capaz de dar excelentes resultados. En efecto la minería textual es el hallazgo de lo desconocido semi automatizado de modelos y tendencias en inmensos conjuntos de datos (Senso y Eíto Brun, 2004, p. 2).

Por lo tanto, el objetivo de la minería de textos es descubrir nuevos conocimientos, procesar y presentar la información disponible en una gran cantidad de formatos de documentos que facilitan la comprensión y el análisis. Con esta definición damos una definición más práctica de ciencia (Senso y Eíto Brun, 2004, p. 2).

De manera similar, Sullivan afirma que la minería de textos es el proceso de recopilar, organizar y analizar grandes volúmenes de documentos para apoyar la difusión de información a analistas y tomadores de decisiones y para descubrir relaciones entre hechos relevantes repartidos en diferentes campos de estudio (Senso y Eíto Brun, 2004, p. 3).

Sin embargo, no queremos terminar esta sección sin resumir las definiciones encontradas en varios documentos oficiales de IBM, uno de los creadores de la herramienta comercial de minería de textos de la que hablaremos más adelante. En estos documentos, la minería de texto se

define como el proceso de extraer automáticamente información significativa del texto, descubrir automáticamente temas importantes en un conjunto de documentos y buscar texto relevante mediante una consulta flexible y de alto rendimiento (IBM, 1998, p. 50) (Senso y Eíto Brun, 2004, p. 3).

Esta definición amplía el pragmatismo que mencionamos anteriormente. Evita mencionar la posibilidad de identificar nuevos conocimientos a partir de documentos existentes y alejarse así de la propuesta original de Hirst. En esta definición, la recuperación de información está incluida en la operación de minería de texto. Esta definición nos parece correcta porque incluye la funcionalidad real implementada por las aplicaciones de minería de texto actuales de las que podemos extraer información (Senso y Eíto Brun, 2004, p. 3).

Otra definición de Business Papers, en este caso del fabricante de aplicaciones de análisis SAS, establece que la minería de texto es el proceso de examinar una gran cantidad de documentos de texto libre para descubrir y explotar el conocimiento disponible en la colección. Esta definición resume la funcionalidad que podemos encontrar en las aplicaciones comerciales: ayudar a comprender e interpretar la información recopilada en muchos documentos (Senso y Eíto Brun, 2004, p. 3).

La minería de texto tiene como objetivo encontrar información relevante para un propósito, especialmente encontrar información relevante del texto en lenguaje natural. Ian h Witten, Frank Abe y A. Marks lo definen como un método que tiene como objetivo encontrar patrones en el texto. Para lograrlo, analiza el texto para extraer información útil para un propósito específico (Godoy Viera, 2017, p. 6).

Ronen Sign y James Singer lo describieron como un proceso de conocimiento cercano. Entre ellos, los usuarios utilizaron herramientas de análisis para interactuar con la documentación con un conjunto de kits de tiempo; Intentó obtenerlo de la fuente de datos identificando y explorando patrones interesantes para obtener una utilidad de la fuente de datos. Información. Estos incluyen una colección de documentos de origen de datos. Un modelo interesante está en el texto no estructurado habitual (Godoy Viera, 2017, p. 7).

La minería de texto se deriva en gran medida de la investigación de minería de datos y, por lo tanto, tiene similitudes en la arquitectura de alto nivel; por ejemplo, ambos sistemas se basan en rutinas de preprocesamiento, algoritmos para encontrar patrones y capas de elementos de presentación

que incluyen herramientas de visualización para mejorar la navegación en clústeres receptivos (Godoy Viera, 2017, p. 7).

Dado que la minería de textos se centra en el análisis de textos en lenguaje natural, se basa en otras disciplinas informáticas que se ocupan del procesamiento del lenguaje natural. También utiliza técnicas y métodos de los campos de la recuperación de información, la minería de información y la lingüística de datos, principalmente (Godoy Viera, 2017, p. 7).

Podemos distinguir la minería de datos de la minería de texto, como lo señala Sholom M. Weiss et al., quienes afirman que la primera se organiza en hojas de cálculo y la segunda en formato de documento, para aprender comienza con el esquema utilizado en el mundo de los documentos, que es una variante de formato XML (Godoy Viera, 2017, p. 7).

Para Feldman y Sanger, también señalan las diferencias entre minería de datos y minería de texto. Para ellos, en el primero, los datos se almacenan en formatos estructurados, y gran parte de su preprocesamiento se centra en limpiar y normalizar los datos, así como en crear un gran número de uniones de tablas. Por el contrario, el preprocesamiento de minería de texto se centra en identificar y extraer características representativas de documentos en lenguaje natural. Estas funciones pueden incluir la identificación de palabras clave relacionadas, nombres personales, organizaciones y similares. El propósito del preprocesamiento es transformar los datos no estructurados que se encuentran en una colección de documentos en un formato estructurado medio más claro (Godoy Viera, 2017, p. 7).

Para Witten et al. Explique la diferencia entre minería de datos y minería de texto y muestre que la primera intenta extraer información indirecta, previamente desconocida y potencialmente útil de grandes cantidades de datos. En segundo lugar, la información extraíble del texto está escrita de forma clara y sin ambigüedades. Sin embargo, un problema importante para los usuarios es la dificultad de acceder y leer la gran cantidad de documentos de texto en formato digital que actualmente están disponibles con fines informativos, estratégicos o de entretenimiento (Godoy Viera, 2017, pág. 8).

Por lo tanto, la minería de texto estudia diferentes métodos de representación de documentos de texto para que puedan ser utilizados por los sistemas informáticos y las personas que no tienen tiempo para leer la gran cantidad de contenido disponible en los medios digitales (Godoy Viera, 2017, p. 7).

Desde el punto de vista de los sistemas automatizados de procesamiento de texto, el problema es que los documentos de texto en su mayoría no están estructurados, por el contrario, son amorfos y difíciles de procesar, y la mayoría de los documentos no están representados adecuadamente en una forma que pueda ser utilizada directamente por el sistema de mineria de texto (Godoy Viera, 2017, p. 8).

Para Michael W. Berry y Jacob Kogan, sin embargo, confirman que los temas de investigación clave en la minería de textos son la extracción de palabras clave, la clasificación, la agrupación, la extracción de nombres y entidades, la detección de anomalías y tendencias y el flujo de texto. Cada uno de estos temas es parte del subcampo de minería de texto (Godoy Viera, 2017, p. 8).

En la subsección de resumen de texto, el resultado del sistema de minería de texto es un resumen de las características más importantes del gran cuerpo de texto (los pasajes relevantes del documento). Otro subcampo de la minería de texto es la clasificación, donde cada instancia representa un documento y las categorías son temas. Por ello, se utilizan diversas técnicas de minería de textos para clasificar los documentos en función de las palabras que aparecen en ellos (Godoy Viera, 2017, p. 8).

La agrupación de documentos es otro subcampo de la minería de texto que agrupa las palabras encontradas en función de sus criterios de similitud. La característica principal de esta técnica es que no existen categorías predefinidas, sino que permite crear varios grupos para el conjunto de documentos a procesar (Godoy Viera, 2017, p. 8).

Sobre Weiss et al. Mencionaron algunas áreas de minería de texto de aplicación, como la clasificación de documentos, recuperación de información, agrupación de documentos y organizaciones. Esto también se aplica al análisis sensorial, el análisis de Internet, los textos de las redes sociales, etc (Godoy Viera, 2017, p. 8).

Se puede decir que la minería de textos es una herramienta que cubre una amplia gama de campos que van desde la búsqueda y extracción de información, presentación, agregación de múltiples documentos, minería de datos para texto, etc. Es un término cuyo uso se limita o amplía según el autor que lo utilice (Botta Ferret y Cabrera Gato, 2007, p. 7).

La minería de texto se define como la búsqueda de patrones o patrones en un texto basado en técnicas de aprendizaje automático; por lo tanto, se considera una de las muchas ramas de la lingüística computacional. Como

proceso, se trata del descubrimiento de conocimientos que no están en el texto, sino correlacionando el contenido de varios textos, y se divide en varias etapas (Botta Ferret y Cabrera Gato, 2007, p. 7).

Herramientas de gestión del conocimiento minero según texto definido. Es un proceso que se ocupa del descubrimiento de conocimientos que no están contenidos en los textos, se crea correlacionando el contenido de varios textos, basándose en las técnicas de aprendizaje automático de regularidades o patrones encontrados en ellos, ya que son textos especialmente desestructurados, para la información almacenada como texto no estructurado: informes, correos electrónicos, actas de reuniones, etc (Botta Ferret y Cabrera Gato, 2007, pp. 9-10).

La minería de textos es la adquisición de conocimiento en forma textual a partir de diversas fuentes de información, lo que nos permite tomar decisiones estratégicas (Gavilanes, Rio, Cilleruelo y Garechana, 2011, p. 2).

El análisis basado en la minería de texto se divide en dos partes: 1) el refinamiento o preprocesamiento del texto, que produce una forma intermedia de información que se puede controlar mediante métodos computacionales. 2) Extraer conocimiento de esta interfaz usando herramientas comunes de minería de datos para extraer patrones, relaciones o conocimiento (Gavilanes et al., 2011, p. 2).

En particular, una forma intermedia basado en documentos puede transformarse en una forma basada en conceptos si de él se extrae información relevante de acuerdo con los intereses del análisis., Tan (1999) (Gavilanes, Rio, Cilleruelo y Garechana, 2011, p. 2).

Otros autores han señalado que parte de la riqueza semántica contenida en el texto fuente se pierde cuando se utilizan métodos inductivos clásicos de minería de textos, y han propuesto diferentes aproximaciones a esto, proponiendo procedimientos de minería de textos para apoyar hipótesis de desarrollo experto o como búsqueda. herramientas conflictivas, etc. Sanchez et al. (2008) (Gavilanes et al., 2011, p. 2).

De manera similar, Porter (2007a) reconoce las limitaciones de las fuentes empíricas y recomienda combinarlas con fuentes basadas en conocimiento experto. Un enfoque interesante es crear métricas de innovación basadas en datos e invitar a expertos comerciales y técnicos a revisarlos en busca de errores, sugerir formas de ampliar la cobertura y explicar (Gavilanes et al., 2011, p. 2).

La definición de minería de texto o minería de opinión se centra en el procesamiento automático de información de opinión, que permite extraer del texto polaridades positivas, negativas y neutras o mixtas (Pang y Lee, 2008, p. 12).

Definición de minería de texto (MT) La minería web, especialmente la minería de contenido, impregnada como un campo de procesamiento automático de texto, la minería de texto es el proceso de recopilación, organización y análisis de grandes volúmenes de documentos para apoyar la difusión de información a analistas y tomadores de decisiones. descubrir relaciones entre hechos relevantes que se extienden a través de varios campos de estudio (Angulo Cuentas et al., 2009, p. 7).

La minería de textos no debe confundirse con las funciones avanzadas de los sistemas de gestión de bases de datos o buscadores de Internet, que tienen como objetivo encontrar documentos en una gran cantidad de documentos que satisfagan las necesidades descritas en las solicitudes de los usuarios, mientras que la minería de textos se basa en los mismos autores. que quieren que el Conocimiento se encuentre en una gran cantidad de texto, literalmente no registrado en ningún documento. Paralelamente a la minería de datos, que extrae información útil de grandes cantidades de datos, la minería de textos es un proceso que se aplica a grandes cantidades de texto libre no estructurado (Angulo Cuentas et al., 2009, p. 7).

La minería de texto tiene tres componentes importantes:

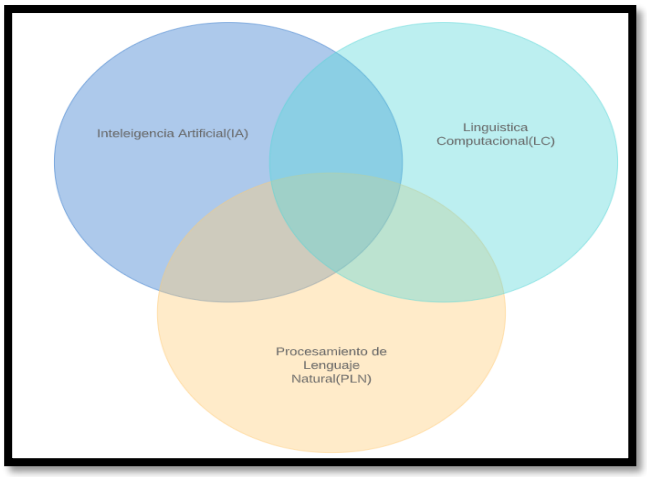
- Recuperación, es la extracción de textos para que luego se realice su transformación.
- Tratamiento de la información, la extracción de patrones de los datos recuperados conseguidos en la fase de recuperación. En este apartado es acerca de ordenar, clasificar y cuantificar el material primero no estructurado.
- La Integración de la información, es la combinación de los procesos cognitivos humanos del procesamiento de la información en la salida de la computadora.

La minería de textos, denominada minería de datos de texto, es, desde un punto de vista práctico, el proceso de extraer información cualitativa de un texto dado. Así el análisis de texto logra encontrar patrones en un conjunto de textos para facilitar una mejor toma de decisiones. Sin embargo, dice que el texto es una de las fuentes más comunes y más grandes de big data, por lo que los datos de texto se encuentran en correos electrónicos, mensajes de

texto, tweets, entradas de redes sociales, etc., como blogs, wikis, WhatsApp, Viber y Line. mensajes instantáneos, Gmail, WhatsApp, Facebook Messenger Live Chat, fuentes de datos especiales, así como libros, informes, estudios, artículos de noticias, contenido de sitios web. El análisis de texto o el análisis de texto en forma de procesamiento de lenguaje natural y otras disciplinas como la inteligencia artificial y la lingüística computacional son disciplinas enfocadas que fueron ampliamente reconocidas hace muchos años que muestran en la figura 2 (Joyanes Aguilar, 2013, p. 55). tal como se puede apreciar en la Figura 2.

Figura 2.

Disciplinas de la minería de texto



Nota. La figura muestra 3 disciplinas de la mineria de texto. Fuente: Elaboración propia.

La minería de datos implica encontrar datos. Del mismo modo, la minería de texto está tratando de encontrar el estado en el texto: es un proceso de análisis de texto para obtener información. Comparado con los datos cubiertos en este libro, el texto es desestructurado, amorfo y voluminoso. Sin embargo, en la cultura occidental moderna, el texto es el medio más común para transmitir información formalmente. Motivación para tratar de obtener información. Fue tranquilizador para él, incluso si el éxito fue solo parcial (Witten et al., 2011, p. 421).

Las similitudes superficiales entre la minería de texto y de datos ocultan diferencias reales. Describimos la minería de datos como la extracción de información indirecta, previamente desconocida y potencialmente útil de los datos. Pero con la minería de texto, la información que se va a extraer se

especifica de forma clara y sin ambigüedades en el texto. No está oculto en absoluto: la mayoría de los escritores hacen todo lo posible para asegurarse de expresarse de manera clara y sin ambigüedades. Desde una perspectiva humana, el único significado de la palabra "anteriormente desconocido" es que las limitaciones de tiempo impiden que las personas lean el texto por sí mismas. El problema, por supuesto, es que la información no se edita para que sea adecuada para el procesamiento automático. La extracción de texto trata de presentarlo en un formato adecuado para el uso de computadoras o personas que no tienen tiempo para leer el texto completo (Witten et al., 2011, p. 421).

Un requisito común en la minería de datos y textos es que la información extraída debe ser potencialmente útil. En cierto sentido, significa orientado a la acción, capaz de proporcionar la base para que las acciones se realicen automáticamente. En el caso de la minería de datos, este concepto se puede expresar de una manera relativamente independiente del dominio: los patrones de acción permiten predicciones no triviales sobre nuevos datos de la misma fuente. El rendimiento se puede medir por el éxito y el fracaso computacional, las técnicas estadísticas se pueden usar para comparar diferentes métodos de extracción de datos en el mismo problema, etc. Sin embargo, en muchas situaciones de minería de texto, es difícil describir qué significa "operativo", independientemente del dominio específico. Esto hace que sea difícil encontrar una medida justa y objetiva del éxito (Witten et al., 2011, p. 421).

Como hemos enfatizado a lo largo de este libro, en la minería de datos práctica, el término "potencialmente útil" a menudo tiene una interpretación diferente: la clave del éxito es que la información extraída debe ser comprensible porque ayuda explicar los datos. Esto es necesario si el resultado está destinado al consumo y no (o como y también) como base para la acción automática (Witten et al., 2011, p. 422).

Este estándar no se aplica a la minería de texto porque es diferente de la minería de datos, es comprensible ingresar a sí mismo. La extracción de texto directo con una salida comprensible es equivalente a recopilar las características más excelentes de un gran número. Este es un campus adjunto: revisión de texto (Witten et al., 2011, p. 422).

Ya nos hemos encontrado con un problema importante de minería de texto: la clasificación de documentos, donde cada instancia representa un documento y la clase de instancia es el sujeto del documento. Los

documentos se caracterizan por las palabras que aparecen en ellos. La presencia o ausencia de cada palabra se puede tratar como un atributo booleano, o el documento se puede tratar como una bolsa de palabras en lugar de una colección basada en la frecuencia de palabras (Witten et al., 2011, p. 422).

La minería de texto es el análisis de texto que permite descubrir diferentes tipos de conocimiento que son útiles para muchas aplicaciones, especialmente el conocimiento sobre las opiniones y preferencias de las personas, que a menudo se expresan directamente en datos textuales. Por ejemplo, ahora es común usar datos de texto de opinión, como reseñas de productos, debates en foros y textos de redes sociales para generar opiniones sobre temas de su interés, lo que simplifica varias tareas de toma de decisiones. Seleccione un servicio. Debido a la gran cantidad de información, las personas necesitan herramientas de software inteligentes que les ayuden a descubrir información relevante para optimizar las decisiones o realizar tareas de manera más eficiente (Zhai y Massung, 2016, p. 24).

Mineria de datos de texto (TDM) es la aplicación de minería de datos se basa en descubrir abundante cantidad de información que no se encuentra estructurado. Información no estructurada significa que no se almacenan en un formato de base de datos estructurado, de modo ordenado y posteriormente encontrar y usar para distinto finalidad. Esta información son textos tales son mensajes de WhatsApp, correos electrónicos, twitter, presentación de PowerPoint y documentos en Word, etc.; o estos estan de manera no textual tales como imágenes de formato JPEG, archivos de audio MP3, etc (Torres Samboni, 2015, p. 4).

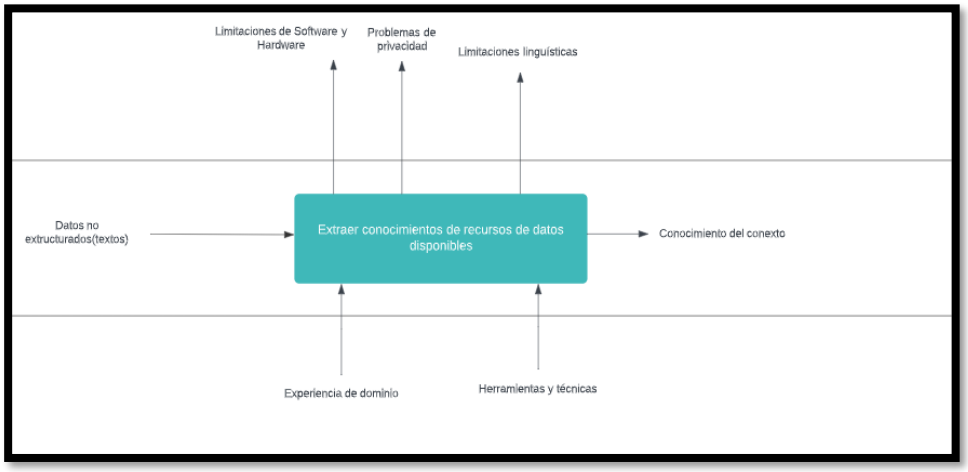
La característica principal de la minería de textos trabaja en lenguaje natural. A diario hablamos los humanos este lenguaje. De esta forma se representa el lenguaje natural en los textos descritos, que es objeto de extracción de conocimiento (Torres Samboni, 2015, p. 5).

Está orientado para descubrir patrones importantes o eventos frecuentes, la finalidad es descubrir desviaciones, tendencias y asociaciones en abundante cantidad de información textual. Las aplicaciones de sistemas de minería de texto tales son la identificación redirección de contenido de correo electrónico; sistemas de vigilancia, análisis de información de los artículos y libros, búsqueda importante en los artículos (Torres Samboni, 2015, p. 5).

El proceso de la minería de textos tiene los siguientes elementos podemos ver en el diagrama de contexto (Torres Samboni, 2015, p. 5). tal como se puede apreciar en la Figura 3.

Figura 3.

Los elementos del proceso de la minería de textos



Nota.La figura muestra los elementos del proceso de la mineria de textos. Fuente: Estudiante universidad San Carlos.

Datos no estructurados es el insumo importante para realizar minería de textos, podemos ver en un ejemplo de tweet.

“Militares se alistan en las aulas para posconflicto. Capacitaciones en #sena Formulación de proyectos. Vía @eltiempo http://app.eltiempo.com/colombia/cali/posconflicto-en-el-valle/16400807…” (Torres Samboni, 2015, p. 5).

El texto no posee el estándar pues tiene muchos símbolos, sitios web que no aporta significado analizamos otro tweet de manera distinto se representa. No tiene normas donde detalle la estructura, por ejemplo, al inicio de tweets debe ir el nombre de la noticia, seguido etiquetas de usuarios y la URL del sitio web, de tal forma que el tweet este organizado (Torres Samboni, 2015, p. 5).

Una vez concluido el estudio se logran los objetivos definidos al inicio de la minería de textos, en la salida se logra el conocimiento extraído del análisis de tendencias, agrupaciones de datos o clasificaciones (Torres Samboni, 2015, p. 5).

El proceso tiene una serie de limitaciones del software tales como la descarga de registros de los navegadores de internet Google, Internet Explorer y redes sociales estos son Facebook y Twitter. Limitaciones de hardware y privacidad para conseguir permisos de usuarios para acceder enormes cantidades de información (Torres Samboni, 2015, p. 5).

La minería de textos tiene una metodología para el proceso de extracción de información (Torres Samboni, 2015, p. 6).

- Establecer el corpus.
- Crear matriz de términos.
- Extraer el conocimiento.

En la primera fase es acerca de la linguistica de corpus es la rama de la linguistica investiga en base a datos obtenidos del corpus, o sea muestras verdaderas del empleo de la lengua. Se emplea en dos contextos, con respecto a la minería de textos recopilan material linguístico, tales como oraciones o textos. La córpora inmenso en formato digital compuesto de la fuente de gran tamaño de información acerca del empleo de la lengua, estos son información gramatical, léxica, semántica, etc. Se emplea la fuente de información para redactar diccionarios y para el proceso automático del lenguaje natural (Torres Samboni, 2015, p. 6).

La finalidad del corpus es recopilar los documentos que se desea analizar. La recopilación está constituida de documentos de texto, XML, correos electrónicos, páginas web, comentarios de redes sociales (Torres Samboni, 2015, p. 6).

Después de concluido el procedimiento de recopilación organizan y transforman los documentos de modo que todos posean igual formato, seguidamente este sea procesado por la máquina Una organización puede ser simple y constar de una colección almacenada en un directorio y constar de varios registros en una base de datos (Torres Samboni, 2015, p. 6).

En la segunda fase se emplea el corpus para la creación de matriz de términos del documento (TDM). En esta matriz, las filas representan los documentos, a la vez que las columnas representan los términos, o sea, las palabras del documento. relacionados a través de términos y documentos se manifiesta por medio de índices, o sea, las medidas relacionales son simples como el número de ocurrencias de cada término en cada documento o el número de ocurrencias del término. El propósito es representar el corpus mediante una matriz de términos (Torres Samboni, 2015, p. 6).

En cambio, no todos los términos incluidos en el documento describen sus características, como artículos, verbos auxiliares, puntuación, pronombres, etc.; ejemplos en las siguientes oraciones: He de hacerlo cuanto antes; Hoy he comido a las dos; El verbo "he" es un auxiliar y no hay diferencia, por lo que no debe formar parte del término matriz. Esta lista de términos que no hay diferencia se denomina palabras sin significado o stopwords. La construcción inicial de la matriz de términos debe incluir todos los términos definidos en el corpus (columnas) excepto los incluidos en la lista de palabras sin significado o stopwords, todos los documentos en el corpus (filas) y las cantidades de apariciones de cada término en cada documento (Torres Samboni, 2015, p. 7).

Asimismo, el método de palabras sin significado o stopwords, también se realiza el llamado stemming es un método consiste en cambiar las palabras por su raíz. Reducir un término a su raíz o lema a través un algoritmo de distintos maneras gramaticales o verbales se identifiquen con el mismo término; por ejemplo, los resultados originales de las siguientes palabras; perro, perra, cachorro, ese perro; distintas maneras que pueden tomar estas palabras se reducen a la forma común (Torres Samboni, 2015, p. 7).

Seguidamente se calcula la frecuencia de los términos y filtrar aquellos que no es importantes. La frecuencia es las veces que aparecen un término en un documento esto significa lo importante que tiene en este documento, por ejemplo, si un término aparece una vez en el documento A y tres veces en el documento B. A continuación, se muestra procesos de estandarización (Torres Samboni, 2015, p. 7).

Frecuencias logarítmicas: la transformación reduce el impacto de las frecuencias originales y afectan resultados de análisis (Torres Samboni, 2015, p. 7).

Frecuencias binarias: Matriz de términos resultante incluirá 1 y 0 el cual indica la presencia o ausencia de las palabras. Esta transformación reduce el impacto de las frecuencias originales en el análisis y cómputos (Torres Samboni, 2015, p. 7).

Frecuencias de documentos inversas: incluye una transformación que es útil porque permite medir la relevancia de los términos, ya que refleja el número de ocurrencias de un término en un documento y tiene varias versiones, generalmente basadas en función inversa del número de documentos que aparece el término (Torres Samboni, 2015, p. 7).

En la tercera fase es extraer el conocimiento, luego de haber

construidos la matriz de términos, los patrones se extraen en el contexto del problema específico bajo estudio. Por ejemplo, en el entorno comercial, es interesante encontrar patrones ocultos de consumo de los clientes para investigar nuevos horizontes. Por ejemplo, un vehículo deportivo corre riesgo de accidente igual que un vehículo normal. Inventar novedoas estrategias (Torres Samboni, 2015, p. 7).

Existen cuatro métodos para extraer la información esto es: clasificación, clustering, asociación y análisis de tendencias (Torres Samboni, 2015, p. 8).

- Clasificación: La clasificación, se basa en agrupar las instancias en categorias. A la clasificación de textos se asigna a diversos documentos una etiqueta relacionada al contenido, a esto se denomina aprendizaje supervisado pues la función es en base a datos de entrenamiento. Por ejemplo, un grupo de asuntos y una serie de documentos de textos en el cual hay que hallar el asunto de los documentos mediante modelos de un grupo de datos de prueba. La clasificación de textos automática es aplicada a spam, clasificación de páginas web, generar automáticamente metadatos, en el caso se toma un documento para publicar en una página web y el clasificador de textos localiza el documento en una categoría, tales son deportes, política, cultura, etc (Torres Samboni, 2015, p. 8).
- Clustering: asimismo se denomina agrupamiento y está constituida los datos en grupos de objetos parecidos métodos de clustering usan algoritmos matemáticos encargadas de agrupar objetos. Se usa la información de las variables que pertenecen a cada objeto, luego se clasifica en clase (Torres Samboni, 2015, p. 8).

  Se aplica este tipo de análisis desconocido etiquetas asociados a datos. Cuando se realiza clústeres, se identifican espacios de características, asimismo descubrir division de patron y Correlación de atributos. A diferencia el clustering es aprendizaje de observación, en tanto la clasificación es aprendizaje de ejemplos (Torres Samboni, 2015, p. 8).

  El clustering se aplica para detectar relaciones de muchos textos, de distribuir en agrupaciones o descubrir temas mas relevantes de sus contenidos y expresar en sus términos. En especifico, el clustering se aplica en descubrimiento de temas (Torres Samboni, 2015, p. 8).

  Asimismo, el clustering es muy importante en aplicaciones tales son la exploración de datos científicos, agrupar documentos, aplicación de base de

datos tales como GIS, aplicación web, marketing (Torres Samboni, 2015, p. 8).

- Asociación: La asociación es la generación de reglas para identificar relaciones entre terminos. Las reglas de asociación son usadas para analizar la literatura que se publicó tales como noticias y artículos académicos publicados en la web, etc. La finalidad es identificar automáticamente las asociaciones de distintos conjuntos (Torres Samboni, 2015, p. 9).

Métodos de minería de datos Un aspecto que debe quedar muy claro sobre el proceso KDD es la diferencia entre tareas y métodos de minería de datos. Las tareas de minería de datos pueden ser tanto predictivas como descriptivas. Las tareas de predicción son preguntas y tareas en las que uno o más ejemplos deben predecir uno o más valores. Dependiendo de cómo los ejemplos coincidan con los valores de salida y cómo se muestren los ejemplos, podemos definir diferentes tareas de predicción (Osorio Zuluaga, 2009, p. 29).

- Clasificación o discriminación
- Clasificación blanda
- Estimación de probabilidad de clasificación
- Categorización
- Preferencias o priorización
- Regresión
- Las tareas descriptivas tienen como objetivo describir los datos existentes. A continuación, las tareas descriptivas más determinadas:
- Agrupamiento por sus siglas en ingles clustering.
- Correlación o factorización
- Criterios de asociación
- Dependencias funcionales
- Detección de valores e instancias anómalas.

Cada una de las primeras tareas requiere métodos, técnicas o algoritmos para resolver. Se pueden visualizar algoritmos y sus métodos correspondientes. La minería de textos es un uso generalizado de dispositivos informáticos y de comunicación que generan grandes cantidades de información digital, especialmente en la producción de datos textuales, y requiere el desarrollo de métodos, algoritmos y sistemas que puedan procesarlos automáticamente. Pueden ser estructurados, semiestructurados

y no estructurados, dando lugar a campos de investigación de la información como la minería de textos, que puede definirse como el proceso de descubrir patrones importantes y nuevos conocimientos en un conjunto de textos, por lo que se encarga del descubrimiento. proceso de conocimiento No está claramente expresado en ningún texto de la colección, y surge nuevamente al vincular varios de estos.

El objetivo de la minería de textos es extraer información de un texto no estructurado, como entidades (personas, organizaciones, fechas, cantidades) y las relaciones entre ellas. La búsqueda semántica, por otro lado, permite a los usuarios especificar no solo elementos que deberían aparecer en una consulta en un documento, sino también entidades y relaciones que se derivan a través del análisis de texto. Una clasificación de documentos de texto es una aplicación de minería de texto que asigna una o más categorías, etiquetas o clases a los documentos en función de su contenido. Es una parte esencial de muchas tareas organizativas y de gestión de la información. Los métodos tradicionales de clasificación de texto, en los que los expertos en el dominio del texto definen manualmente las reglas de clasificación, han sido reemplazados por métodos basados en técnicas de aprendizaje automático o una combinación de aprendizaje automático y otras técnicas (Pérez Abelleira y Cardoso, 2010, p. 2).

Los clasificadores están diseñados para encontrar patrones o patrones en el comportamiento de una variable en relación con el comportamiento de otra variable en una base de datos de incertidumbres y observaciones no cuestionadas. Se divide principalmente en: generador de árboles de decisión, generador de reglas, modelo de regresión, red neuronal, red de características, clasificador bayesiano. Algoritmos genéticos, métodos de visión, clasificaciones difusas (Arribas López, 2008, p. 11).

La minería de datos es el proceso de usar datos, extrayendo información importante previamente desconocida de grandes bases de datos para usarla como un elemento importante en la toma de decisiones (Angulo Cuentas et al., 2009, p. 4).

La minería de datos es un término general que incluye los resultados de la investigación, los métodos y las herramientas utilizadas para extraer información útil de grandes bases de datos. Además, la minería de datos es parte de todo el proceso KDD, por otro lado, los conceptos de minería de datos y KDD se consideran similares entre sí en la literatura (Molina López y García Herrero, 2006, p. 13).

Específicamente, la comunidad de estadísticos, analistas de datos y administradores de sistemas de información a menudo usan el término minería de datos para el proceso general de descubrimiento, mientras que el término KDD es más comúnmente utilizado por expertos en inteligencia artificial.

El propósito de la minería de datos es descubrir hechos, hechos valiosos e información previamente desconocida contenida en una base de datos, desde la academia el concepto de minería de datos es un proceso mayor llamado KDD, en resumen, la minería de datos combina varios campos como la estadística, artificial. inteligencia, infografías, bases de datos y procesamiento masivo que utilizan las bases de datos como materia prima principal.

La causa de la minería de texto es el problema cada vez más apremiante de extraer información automáticamente de grandes cantidades de texto. Se trata de extraer información de datos no estructurados: texto plano. Hay varias formas de representar la información no estructurada: "bolsa de palabras": cada palabra forma una posición en un vector con un valor correspondiente al número de veces que aparece; N-gramas u oraciones: permite considerar el orden de las palabras. Prueba mejores oraciones negativas. " Excepto.", pero no, de lo contrario considerarían que tales palabras están relacionadas. Representación Relacional (primer orden): permite detectar patrones más complejos (cuando la palabra X está a la izquierda de la palabra Y en la misma oración). Categorías de conceptos. Casi todo el mundo se enfrenta a un "problema léxico": tienen problemas con los sinónimos, la polisemia, los lemas, etc. Un ejemplo de una aplicación de minería de texto es la generación automática de índices sobre documentos. Otros, más sofisticados, incluyen escanear el texto completo y mostrar un mapa con las partes o documentos más relevantes colocados uno al lado del otro. En este caso se trataría de analizar las palabras en el contexto en el que se encuentran. En cualquier caso, aunque no se ha avanzado mucho en el campo de la minería de textos, ya existen productos comerciales que utilizan esta técnica para diversos fines (Molina López y García Herrero, 2006, p. 20).

La minería de textos es un área de investigación del procesamiento automático de información. Se define como el proceso de descubrir patrones interesantes y nuevos conocimientos en una colección de textos, es decir, es el proceso encargado de descubrir conocimientos que no se encuentran claramente en ningún texto de la colección, pero que surgen para relacionar

el contenido. de varios de ellos (Contreras Barrera, 2014, p. 4).

La minería de textos explica como el proceso encargado de descubrir patrones interesantes y nuevos conocimientos en un conjunto de textos, es decir, que no están explícitamente presentes en ningún texto de la colección, pero que provienen de asociar varios de ellos con contenido (Hearst, 1999; Kodratoff, 1999).

Este proceso consta de dos etapas principales: el preprocesamiento y la etapa de descubrimiento (Tan, 1999).

El proceso de minería de texto consta de dos etapas principales: preprocesamiento y etapa de descubrimiento. En la primera etapa, los textos se transforman en una forma de representación estructurada o semiestructurada, lo que facilita su posterior análisis, mientras que, en la segunda etapa, se analizan las representaciones para descubrir patrones interesantes o nuevos conocimientos en ellas (Gil Pascual, 2021, pp. 9-11). tal como se puede apreciar en la Figura 4.

Figura 4.

Etapa de preprocesamiento

Etapa de pre-procesamiento	Tipo de representación	Tipo de descubrimientos
Categorización	Vector de temas	Nivel temático
Full-text	Secuencias de palabras	Patrones de lenguaje
Extracción de información.	Tabla de datos	Relaciones entre entidades.

Nota.La figura muestra la etapa de preprocesamiento. Fuente: Montes y Gómez.

- Etapa de Preprocesamiento: En la etapa de preprocesamiento, se transforman los textos en una representación estructurada o semiestructurada de su contenido. Por un lado, estas representaciones intermedias de los textos deben ser faciles para facilitar el análisis de los textos, por otro lado, completas y que puedan permitir el descubrimiento de patrones interesantes, e incluir nuevos conocimientos.

Dos representaciones más utilizadas en la mineria de texto:

- A nivel de documento, cada representación muestra un texto diferente de la colección.
- A nivel conceptual, cada representación representa un objeto, tema o concepto interesante en un dominio de aplicación específico.
- Se utilizan diferentes estrategias para construir representaciones. Por ejemplo, las representaciones a nivel de documento se crean utilizando técnicas de clasificación, texto completo e indexación. Para las representaciones a nivel de concepto, estas se obtienen utilizando técnicas dependientes para el dominio, así como la recuperación de términos importantes y la extracción de información (Montes y Gómez, 2005).
- Etapa de descubrimiento: El descubrimiento de minería de textos divide sus métodos y tareas y se clasifica como: descriptiva y predecible. Por otro lado, se puede clasificar de diferentes maneras. Por ejemplo, una posibilidad es la clasificación de minería de texto, donde el texto es una descripción de una situacion. La representación de texto logrados en la fase de preprocesamiento es una descripción estructurada. Por lo tanto, el descubrimiento de minería de texto se divide en tres enfoques: a) descubrimiento a nivel de presentación, b) descubrimiento a nivel de texto y c) descubrimiento a nivel mundial (Montes y Gómez, 2005).
- Descubrimientos a nivel representación: Método para desarrollar o detectar textos estructurados o semiestructurados. Son encargadas de la clasificación, categorización de textos (Montes y Gómez, 2005).
- Descubrimientos a nivel texto: Hay dos tipos de métodos: métodos que descubren patrones de lenguaje a partir de colecciones de texto y métodos que descubren la organización "oculta" de colecciones de texto.

Destacan los métodos relacionados con el reconocimiento de patrones de lenguaje al tener en cuenta todas las palabras del texto y conservar su orden relativo, es decir, utilizar la representación de texto completo (Full-Text, en inglés). Estos métodos identifican secuencias frecuentes de palabras y, a veces, crean un conjunto de reglas de asociación basadas en estas secuencias para expresar combinaciones de palabras de uso frecuente.

Por otro lado, los métodos relacionados con la agrupación de textos se caracterizan por el uso de diferentes tipos de métodos, desde los

tradicionales métodos euclidiana para medir la distancia entre textos hasta métodos complejos basados en redes neuronales. Tipos de mapas auto organizados.

Estos métodos enfatizan particularmente la visualización e interpretación de los resultados. Por ejemplo, algunos utilizan una interfaz gráfica para analizar agrupaciones, otros identifican etiquetas que describen el contenido de cada grupo y otros identifican documentos representativos de cada clase. Además, la agrupación de texto se utiliza en el análisis exploratorio de colecciones de texto, la generación de resúmenes de varios documentos y otras tareas de descubrimiento, como la detección de asociaciones y desviaciones (Montes y Gómez, 2005).

- Descubrimientos a nivel mundo: Este enfoque considera una variedad de tareas, incluida el descubrimiento de asociaciones, la detección de desviaciones y el análisis de tendencias. Los métodos de este enfoque tienen las siguientes características: a) utilizan la representación de texto tanto a nivel de concepto como de documento; (b) utilizan el conocimiento del dominio, generalmente representado en una jerarquía de conceptos o conjuntos de predicados, y (c) permiten al usuario guiar el proceso de descubrimiento especificando principalmente las regiones y conceptos de mayor interés. Entre los trabajos de descubrimiento de asociaciones, destacan aquellos que proponen descubrir asociaciones temáticas imprecisas de manera similar (A) => B (confianza | soporte) y utilizan elementos estructurados y no estructurados para obtener esta relación.

Por otro lado, los métodos de detección de desviación consideran la detección de textos raros con temas diferentes al promedio del conjunto, así como el descubrimiento de nuevos temas en conjuntos dinámicos como flujos de noticias.

El análisis de tendencias se ocupa de la descripción del desarrollo de una colección de textos. Entre sus métodos, se destacan los siguientes dos enfoques: a) identificar temas de discusión con un comportamiento predeterminado y comparar el tema de dos colecciones de diferentes épocas (Montes y Gómez, 2005).

2.2.2 Análisis de sentimientos

Señala análisis de sentimiento en Twitter. En una primera aproximación, el análisis de sentimientos en Twitter significa que a cada mensaje publicado se le asigna un valor relacionado con la carga emocional que transmite. Se pueden distinguir algunos tipos diferentes de variables en relación con esta carga emocional (Baviera, 2016, pp. 2-3).

- La polaridad: indica si el mensaje es un sentimiento positivo o negativo. En algunos análisis, se introdujo una tercera categoría para clasificar los mensajes neutrales.
- La intensidad: da un valor numérico relativo a la intensidad percibida. Se pueden distinguir intensidades positivas y negativas.
- Las Emociones: clasifica el texto según diferentes tipos de emociones, como feliz, triste o enojado.

El análisis de sentimiento se define como el proceso de determinar opiniones basadas en valoraciones, actitudes y sentimientos sobre un tema en particular. El análisis de sentimientos generalmente tiene dos objetivos: uno es identificar las expresiones emocionales y determinar la dirección en la que los individuos expresan sus sentimientos (Ramon Saura et al., 2018, p. 4).

El análisis de sentimiento intenta clasificar los documentos en función de las opiniones expresadas por los autores. Este nuevo campo, que combina PLN y minería de texto, cubre una amplia gama de tareas que más o menos ya están cubiertas. Existen principalmente dos enfoques diferentes para este problema: la aplicación de aprendizaje automático o métodos semánticos. Hay dos aplicaciones principales: determinar la polaridad de la opinión a nivel de documento, oración o elemento, y determinar si un documento tiene opinión (Martínez Cámara et al., 2011, p. 1).

El análisis de sentimiento puede identificar expresiones positivas, negativas o neutrales de elementos textuales para un tema, producto o servicio específico, entidad, individuo, etc. El análisis de sentimientos puede aplicarse a diferentes métodos y basarse en funciones como etiquetas automáticas en conversaciones, referencias comunes a temas o eventos específicos, emoticonos de uso frecuente o recursos como diccionarios. Tweets positivos, neutrales o negativos (Martínez Cámara et al., 2011, pág. 1).

El Análisis de Sentimiento (AS) es un campo de PLN que tiene como objetivo analizar la percepción que tienen las personas de ciertas entidades tales como productos, servicios, organizaciones, individuos, asuntos, eventos, problemas y sus características. Analizar y determinar polaridad de sus aspectos positivos, negativos o neutral (Henríquez et al., 2017).

El análisis de sentimiento se encarga de clasificar el documento y hace del uso de métodos semánticos (diccionarios de términos), máquinas de aprendizaje (soporte vector machine, Naive bayes, Deep Learning, etc.) según la polaridad de las opiniones expresadas por el autor. En este contexto, determinar la polaridad de una opinión se considera positiva, neutra o negativa respecto de un producto, servicio, organización, individuo o cualquier tipo de entidad que pueda expresar una opinión por escrito (Chanchi G. et al., 2019, pp. 4-5).

El análisis de sentimientos se denomina extracción de opiniones, se define como sentimientos y emociones, es el estudio computacional de opiniones expresadas en documentos, es la actitud del autor ante una situación, producto, empresa o persona. Identificar las cualidades que crean opinión y la clase de emoción que me gusta, lo odio, lo aprecio. y sentimiento, negativo neutral si es positivo (Dubiau y Ale, 2013, p. 1).

El análisis de sentimiento es una ciencia que reconoce el comportamiento humano. El objetivo de la minería es reconocer la polaridad de las opiniones (positivas, negativas o neutrales) por otro lado el análisis de sentimiento está asociado a reconocer emoción. El análisis de sentimiento se define como un proceso, el objetivo es obtener información emocional (Cambria et al., 2012, p. 4).

El objetivo del análisis de sentimiento es obtener opiniones y sentimientos de textos que clasifiquen las emociones expresadas en estos textos utilizando un espectro de polaridad positiva, negativa y neutral, que es la forma en que otros autores clasifican microblogs, publicaciones, comentarios o tweets (Mejova, 2012).

La definición de análisis de sentimientos ofrece un modelo en el que el texto no estructurado se define como datos estructurados. La estructura propuesta es aproximadamente quíntuple, por lo que tenemos un objeto o conjunto de características. La estructura contiene comentarios. Uno puede referirse a la dirección del objeto raíz o una característica específica, al mismo tiempo estas opiniones pueden tener cierta dirección (positiva, negativa, enojado, feliz, etc.) y fuerza, si queremos completar la estructura, también se

puede agregar opiniones no subjetivas como hechos (orientación de opinión neutral). Entonces la estructura tendrá varios hechos y opiniones (Bosch, 2013).

Se refiere a un objeto y sus propiedades.

El análisis de sentimientos, o extracción de opinión, es un campo interdisciplinario que cruza el procesamiento del lenguaje natural, la inteligencia artificial y la minería de textos. surgió como un subcampo de la minería de textos porque analiza opiniones y la mayoría de las opiniones se expresan en formato de texto. Clasificar el texto como subjetivo u objetivo y la orientación del texto como positiva o negativa son las dos tareas principales del análisis de sentimientos. Se utiliza en diversas aplicaciones en compras, marketing, entretenimiento, educación, política y social (Karamibekr y Akbar Ghorbani, 2013).

Los individuos expresan sus opiniones no solo sobre bienes y servicios, sino también sobre diversos temas y problemas, especialmente aquellos que afectan su vida social. Los problemas sociales son problemas relacionados con la vida personal y las interacciones de las personas. La opinión pública sobre temas sociales se utiliza para tomar decisiones que satisfagan los derechos de las personas, así como para juzgar las actitudes y opiniones públicas. Probar estadísticamente que los problemas sociales y de productos difieren desde la perspectiva (Karamibekr y Akbar Ghorbani, 2013).

El análisis de sentimiento depende en gran medida del dominio y del contexto porque utiliza el sentimiento de las palabras, que no solo dependen del dominio sino también del contexto. El contexto se refiere a la combinación de diferentes expresiones, así como a la estructura léxica y sintáctica de la oración. Por ejemplo, mientras que nuestras expectativas para "préstamo" son resultados positivos y negativos para su antónimo "severo", tanto "demasiado fácil" como "demasiado duro" tienen sentimientos negativos sobre el castigo. Una aplicación debe tener en cuenta el contexto. Métodos de análisis de sentimiento. Este artículo se centra en el análisis de los patrones léxico-sintácticos de subjetividad a nivel de oración (Karamibekr y Akbar Ghorbani, 2013).

La clasificación de la polaridad del texto se considera una tarea típica y se identifican dos enfoques:

- Use lexicones con cierta polaridad de sentimientos, pero la desventaja es que estos diccionarios están hechos para un contexto específico.
- Cree un modelo de lenguaje usando datos de entrenamiento y técnicas

> de aprendizaje automático para crear un clasificador que también esté entrenado para un contexto específico pero que capture las características del lenguaje en uso (Mejova, 2012, pp. 24-25).

Una tarea básica y típica del análisis de sentimientos es la clasificación de polaridad de textos en los que las categorías de interés son positivas y negativas, y en ocasiones de clase media mixta. Hay dos formas de abordar esto: LIWC o POMS pueden usar un léxico de sentimientos (una lista de palabras con polaridades de sentimientos conocidas) como se describe anteriormente, o se puede usar datos de entrenamiento para crear un "modelo" de idioma para cada polaridad (Mejova, 2012, pp. 24-25).

A pesar de su simplicidad y facilidad de uso, los métodos basados en diccionarios carecen de flexibilidad debido a la diversidad de temas y estilos de escritura: las palabras del diccionario pueden no aparecer en absoluto en el texto de interés, o pueden usarse de una manera específica. Sin embargo, utilizando técnicas de aprendizaje automático, es posible crear un clasificador entrenado sobre un texto determinado que refleje las características del lenguaje utilizado en él (Mejova, 2012, pp. 24-25).

Se han utilizado varios métodos de visualización de texto para este propósito. La más común es la representación de bolsa de palabras, donde cada palabra se convierte en una característica con un valor binario (1 si aparece en el documento, 0 si no aparece) o algún otro valor (como cuántas veces aparece en el documento). Más complicado. Las representaciones incluyen n-gramas (n palabras consecutivas), oraciones (usando el reconocimiento de palabra a voz) y palabras ricas en negación (distinguiendo entre "malo" y "no está mal"). Estas y muchas otras técnicas para la representación de documentos (que definen el "espacio de características") se han propuesto en la literatura. Sin embargo, debido a la falta de conjuntos de datos y métodos estándar, algunos estudios han producido resultados contradictorios o directamente incomparables. Se puede considerar si una sola representación es suficiente o hay más (Mejova, 2012, pp. 24-25).

2.2.3 Modelos Clasificadores

2.2.3.1 Naive Bayes

El modelo de clasificación utilizado para esta investigación es Naive Bayes (Garcés Chaparro, 2019).

El clasificador Naive Bayes (o Naive Bayes) es un clasificador probabilístico basado en la aplicación del teorema de Bayes y sus características.

Fue estudiado en 1950 y los métodos de clasificación de texto son muy populares, especialmente cuando se combinan con técnicas muy avanzadas como SVM.

En estadística y teoría de probabilidad, representa la probabilidad de un evento basado en el conocimiento de las condiciones relacionados con el evento.

Más precisamente, el teorema de Bayes se define de la siguiente manera: según la ecuación (1).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

La fórmula dice: "La probabilidad de que ocurra A, dado B, es igual a la probabilidad de que ocurra B si ocurre A, por la probabilidad de que ocurra A, dividida por la probabilidad de que ocurra B".

2.2.3.2 Decision Trees

Un árbol de decisión, es un método de aprendizaje supervisado no paramétrico para clasificación y regresión.

El objetivo es crear un modelo que prediga el valor de una variable a través de las reglas de decisión de los datos de la variable.

La ventaja de un árbol de decisión es que es:

- Excelente para trabajar con datos numéricos y categóricos.
- Los árboles se pueden visualizar. Es fácil de entender y explicar.
- Ideal para resolver múltiples problemas de salida.
- Requieren muy pocos datos para la preparación.

Las desventajas de los árboles de decisión son las siguientes:

- Dado que pequeños cambios en los datos que manejan dan como resultado árboles completamente diferentes.
- Pueden crear árboles demasiado complejos que generalizan mal a los datos. Esto se llama sobreajuste. Para evitar este problema, se necesitan mecanismos como la poda, el número mínimo de muestras necesarias para representar los nodos de las hojas o la profundidad máxima para representar un árbol.

2.2.3.3 Support Vector Machines:

Se define como un conjunto de técnicas para descubrir automáticamente patrones en conjuntos de datos y luego usar esos patrones para hacer predicciones (Giraldo Londoño, 2020).

El proceso de aprendizaje es adoptado por un conjunto de datos de entrenamiento y luego se utiliza un algoritmo para el reconocimiento de patrones. Finalmente, un modelo de clasificación con patrones encontrado.

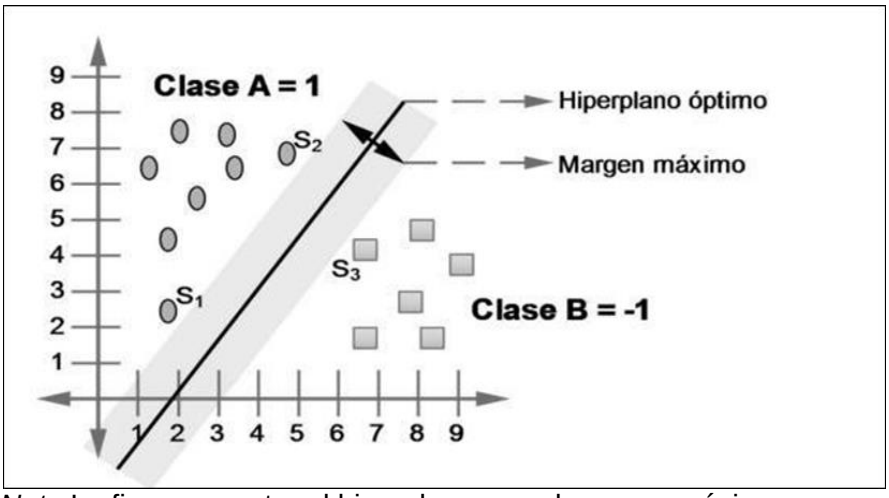
El rendimiento del modelo se mide por la exactitud, es decir, datos predichos correctamente clasificados. El aprendizaje automático se utiliza en varios campos para resolver problemas y es una técnica de clasificación muy utilizada. En reconocimiento facial, se utiliza en documentos, imágenes, etc. para la clasificación.

El propósito de la clasificación es generar una función de conjunto de datos f(x) para el entrenamiento, que predice la clase a la que pertenece un elemento.

Las máquinas de vectores de soporte son los modelos de aprendizaje automático de clasificación más utilizados. Desarrollado por Vladimirs Vapniks en 1995. tal como se puede apreciar en la Figura 5.

Figura 5.

Hiperplano con el margen máximo



Nota.La figura muestra el hiperplano con el margen máximo.

Fuente: Elaboración propia.

Entrenar el modelo permite encontrar el hiperplano que separa dos o más clases donde el hiperplano está desplazado por el margen máximo.

La separación de hiperplanos se logra mediante clasificación lineal. Como se muestra en la Figura 3. S1, S2 y S3 son vectores de soporte definidos por el mejor hiperplano.

El aprendizaje supervisado son la forma más efectiva y fácil de implementarlo. Support Vector Machine es un algoritmo de Machine Learning. El objetivo de la máquina vectores de soporte es encontrar los hiperplanos separados entre la clase. Estos puntos se denominan vectores de soporte; el límite máximo entre los puntos que separan el hiperplano (Reveles Gómez, 2021).

Dadas las dos categorías si se separan, por ejemplo, -1 y 1, busque el hiperplano separe las dos categorías del algoritmo. El hiperplano se calcula de acuerdo con las siguientes fórmulas. según la ecuación (2).

$$F(x) = x.w + b \qquad (2)$$

Las principales funciones de kernel utilizadas en las máquinas de vectores de soporte se muestran en la Tabla 1.

Tabla 1.

Tipos de kernel de Máquinas de Vectores de Soporte

Tipo de SVM	Kernel de Mercer	Descripción
Función de base radial (RBF) o gaussiana.	k (x1, x2) = exp (- [||x1-x2||] ^2/[2σ] ^2).	Aprendizaje de una clase. σ representa la anchura del kernel
Lineal	K (x1, x2) =xT1x2	Aprendizaje de dos clases.
Polinómica	K (x1, x2) =(xT1x2+1) ρ.	ρ representa el orden del polinomio.
Sigmoide	K (x1, x2) = tanh(β0xT1x2+β1).	Representa un kernel de Mercer solo para determinados valores β0 y β1.

Nota. La tabla muestra los Tipos de kernel de Máquinas de Vectores de Soporte. Fuente: Elaboración propia.

2.2.4 Métricas de desempeño para clasificadores

a. La Accuracy

La métrica de Accuracy expresa la relación entre el resultado clasificado correctamente y el resultado total.

La exactitud ("accuracy" en inglés) indica la proporción de resultados verdaderos, verdaderos positivos (VP), Verdaderos negativos (VN) dividido entre el total de verdaderos positivos, falsos positivos, verdaderos negativos, falsos negativos. Entonces, en la fórmula esta métrica se va calcular como se puede ver (Barrios Arce, 2019) según la ecuación (3).

$$Accuracy = VP + VN/VP + FP + FN + VN \quad (3)$$

b. La Precision

La precisión son las tasas de verdaderos positivos divididas por la proporción de verdaderos positivos y falsos positivos. La métrica se calculará como se muestra en la fórmula (Barrios Arce, 2019) según la ecuación (4).

$$Precision = \frac{VP}{VP} + FP \quad (4)$$

c. El Recall

La sensibilidad ("recordar" o "sensibilidad") representa la proporción de verdaderos positivos clasificados correctamente por el algoritmo. La métrica se calculará como se muestra en la fórmula (Barrios Arce, 2019) según la ecuación (5).

$$\text{Recall} = \frac{\text{VP}}{\text{VP}} + \text{FN} \quad (5)$$

d. F1-Score

Esta métrica combina precisión y sensibilidad en una sola métrica. se puede calcular como se muestra en la fórmula (Barrios Arce, 2019) según la ecuación (6).

$$PuntajeF1 = 2 * Precision * \frac{Sensibilidad}{Precision} + Sensibilidad \quad (6)$$

e. La curva ROC

Según la curva ROC general, representa 1 especificidad frente a sensibilidad para cada posible umbral o punto de corte en la escala de resultados de la prueba. Es decir, y = f(x), donde: (del Valle Benavides, 2017) según la ecuación (7).

$$ROC(c) = \begin{cases} y = S(c) \\ x = 1 - E(c) \end{cases} \quad (7)$$

Sin embargo, dada la dificultad de obtener datos de la población, podemos aproximarnos usando una curva ROC de muestra que representa la tasa de falsos positivos en las abscisas y la tasa de verdaderos positivos en las ordenadas según la ecuación (8).

$$ROC_P(c) = \begin{cases} y = FVP(c) \\ x = FFP(c) \end{cases} \quad (8)$$

Dado que tenemos probabilidades en ambos ejes, las curvas ROC de la muestra y la población estarán contenidas en el cuadrado [0, 1] × [0, 1]. También por convención, se considera que los pacientes tienen valores de x más altos que las personas sanas. Entonces la curva estará contenida en el triángulo: {(x, y) |0 ≤ x ≤ y ≤ 1}. Si por la naturaleza de la prueba se invierten los resultados (en promedio, los pacientes enfermos dan valores más bajos que los pacientes sanos), se debe reordenar.

2.3 Definición de términos

2.3.1 Conocimiento

Es información que se almacena y se pone a disposición de las partes interesadas para que puedan realizar y/o mejorar sus operaciones y para que puedan aprender (Gutiérrez Meléndez, 2012).

2.3.2 Información

Son un conjunto de datos significativos cuando están relacionados (Gutiérrez Meléndez, 2012).

2.3.3 Toma de decisiones

Un proceso de selección entre diferentes escenarios posibles (Amaya, 2010).

2.3.4 Bigdata

El proceso de recopilar y analizar grandes cantidades de datos de diversas fuentes, tanto en forma estructurada como no estructurada (Sánchez Guevara, 2014).

2.3.5 Knime

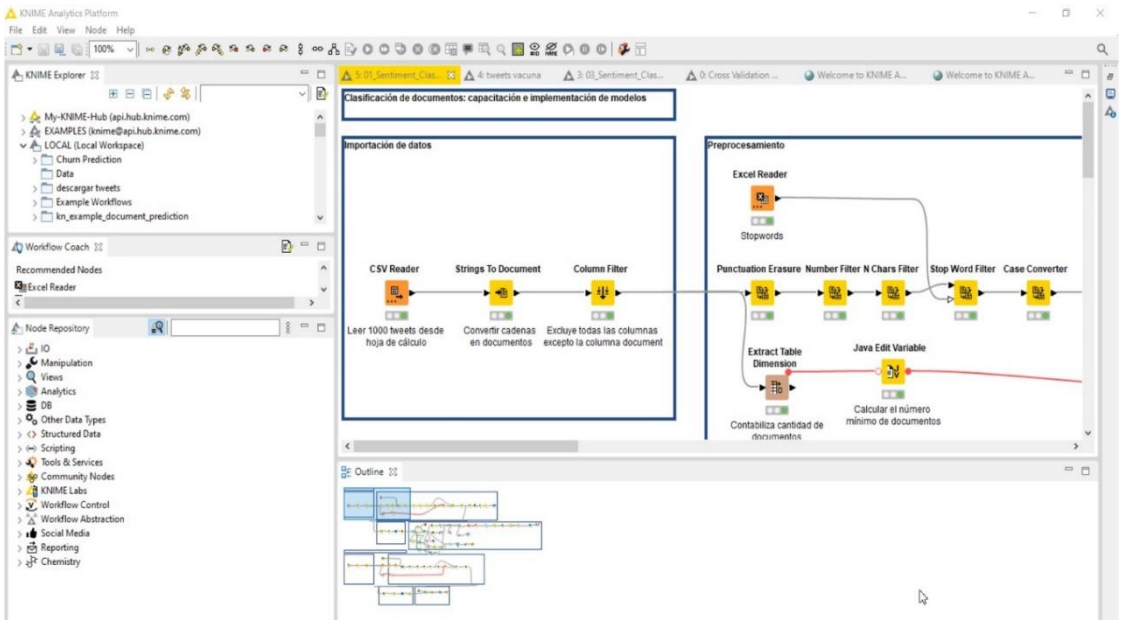
Knime (o Konstanz Information Miner) es un software de minería de datos para desarrollar modelos en un entorno visual. Está desarrollado sobre la plataforma Eclipse.

Los nodos realizan diferentes acciones y se pueden mostrar en tablas de datos.

- Proceso de filas, columnas, etc. mediante filtrado, preprocesamiento, agrupación, etc.
- Visualización de la curva ROC, nube de etiquetas, etc.
- Cree modelos para minería de datos, máquinas de vectores de soporte, árboles de decisión y más.
- Validación de modelos como curvas ROC, etc.
- Suma los puntos para ver los resultados.
- Aplicar el modelo al conjunto de datos. tal como se puede apreciar en la Figura 6.

Figura 6.

Interfaz de knime



Nota. La figura muestra un ejemplo de la Interfaz de knime. Fuente: Elaboración propia.

2.3.6 KDD

Las bases de datos Knowledge Discovery (KDD) tienen un proceso completo de obtención de información, preparación de datos e interpretación de los resultados obtenidos (Molina López y García Herrero, 2006).

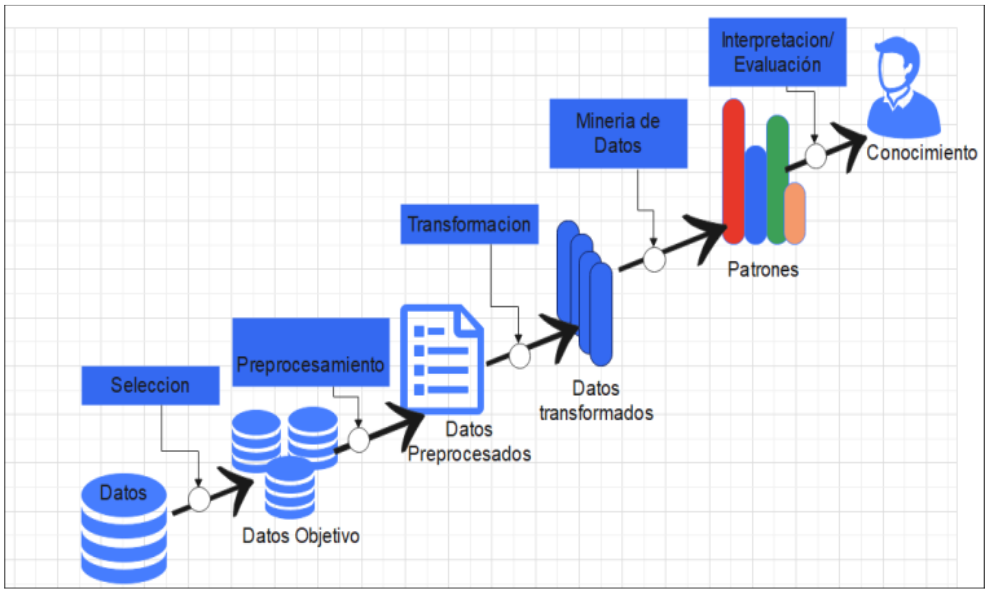
2.3.7 Knowledge Discovery in Databases

Son abstracciones de procesos o patrones automatizados que representan conocimiento implícitamente almacenado o recuperado en grandes bases de datos, almacenes de datos, redes, otros almacenes masivos de información o flujos de datos (Han y Kamber, 2006).

El descubrimiento de conocimiento en bases de datos se define como un proceso que incluye una serie de pasos bien definidos, cada uno de los cuales es esencial para la transformación de los datos en conocimiento (Quiroz Gil y Valencia, 2012). tal como se puede apreciar en la Figura 7.

Figura 7.

Pasos que componen el proceso KDD



Nota. La figura muestra los pasos que componen el proceso de Knowledge Discovery in Databases (KDD). Fuente: Elaboración propia.

2.3.8 Redes sociales

Una red social consta de dos elementos: conexiones y contactos. Las redes sociales son útiles cuando hay un gran número de personas conectadas (Rissoan, 2019).

Análisis de redes sociales (ARS), que parte del concepto básico de red como vínculo o sistema de vínculos entre nodos o entidades sociales, y como vínculo como estructura de comunicación interpersonal, aunque actualmente se cree que los nodos pueden ser agregaciones. (Como organizaciones o países) o puestos. Si bien se supone que la estructura de las relaciones sociales tiene mayor poder explicativo que las características de los actores que integran un determinado sistema, cualquier tipo de relación social puede ser considerada una red (Luna, 2014).

Con el tiempo, las redes sociales se han vuelto muy complejas. A veces es difícil identificar contactos y conexiones (Rissoan, 2019).

Una red social se trata de un grupo de personas. Una red consta de un conjunto de objetos y un mapa de conexiones llamados nodos. Una red social básica consta de dos objetos: Objeto 1 y Objeto 2, con una conexión entre ellos. Por ejemplo, el nodo 1 y el nodo 2 representan personas y las relaciones

son conexiones que representan a compañeros de clase. Si el nodo 1 y el nodo 2 pertenecen a la misma clase, entonces el nodo 2 y el nodo 1 pertenecen a la misma clase (Kadushin, 2013).

La red social es percibida no solo como un objeto que acompaña a las personas en el mar y en la tierra, sino también como una forma de organización social en la que constantemente se intercambian ideas, servicios, objetos y formas de hacer (Madariaga Orozco et al., 2014).

2.3.9 Twitter

Twitter es un servicio de microblogging que nació como una nueva forma de ver lo que sucede en el mundo, basado en la idea de que los usuarios pueden seguirse unos a otros (Kwak et al., 2010).

2.3.10 API de Twitter

Twitter cuenta con una interfaz de programación de aplicaciones, para acceder mediante programación a tweets usando condiciones de consulta, la API tiene parámetros específicos para la recuperación de tweets (Go et al., 2013).

- Autenticación: La API de Twitter extraen una gran cantidad de datos. La autenticación es un método de seguridad de datos para usuarios y desarrolladores. Hay varios tipos de métodos de autenticación, son:
- Contexto de usuario de OAuth 1.0a.
- Token de portador de OAuth 2.0.
- Autenticación básica.
- Flujo de código de autorización de OAuth 2.0 con PKCE.

2.3.11 Normalización de tweets

En los tweets, extraer puntuación, convertir todas las palabras a minúsculas, eliminar todas las palabras que son muy comunes pero que no agregan mucho valor sintáctico (también conocidas como palabras vacías), obtener la forma canónica de las palabras, convertirlas en raíces, encontrar palabras adversativas. en los tweets, las oraciones refutan claramente las

ideas anteriores, reducen la relevancia y se enfocan en las ideas posteriores (Acevedo Miranda et al., 2015; Garrido, 2015).

La normalización de los tweets se realiza limpiando el texto, eliminando la puntuación, convirtiendo el texto a minúsculas y eliminando palabras comunes que no contribuyen a la investigación (Acevedo Miranda et al., 2015; Garrido, 2015).

2.3.12 Clasificador Support Vector Machine

Se define generando un hiperplano que asigna correctamente los documentos a la clase c. Para ello, calcula la derivada de la fórmula de Lagrangiano Dual para encontrar el valor necesario para calcular el hiperplano Bosch (2013).

2.3.13 Máquina de soporte vectorial

Es un algoritmo de clasificación de aprendizaje automático simple que se puede usar en conjuntos de datos tanto lineales como no lineales. El hiperplano de separación óptima es simplemente una línea que representa la separación de dos clases de las características de diferentes clasificaciones.

2.3.14 Minería de texto

Es un análisis cualitativo de extracción de información de un texto para identificar claramente las ideas o conceptos principales contenidos en el texto, en el que se pueden agrupar varias categorías relevantes (Paez Guarnizo y Monroy, 2020).

2.3.15 Análisis de sentimientos

El procesamiento del lenguaje natural es un método para rastrear las emociones de los usuarios frente a un tema o producto específico (Paez Guarnizo y Monroy, 2020).

2.3.16 Conjunto de datos

Se utiliza para entrenar sistemas que pueden detectar varios patrones que consisten en instancias, características y atributos (Paez Guarnizo y Monroy, 2020).

2.3.17 Modelos de predicción

Un modelo de predicción contiene propiedades que definen el modelado y sus metadatos, que analizan el nombre, la descripción, la fecha de procesamiento, los permisos y los filtros utilizados para el procesamiento (Paez Guarnizo y Monroy, 2020).

2.3.18 Modelo supervisado

El aprendizaje supervisado implica hacer predicciones futuras basadas en comportamientos o características vistas en los datos almacenados. Esto nos permite buscar patrones en datos históricos relevantes (Paez Guarnizo y Monroy, 2020).

CAPÍTULO III: MARCO METODOLÓGICO

3.1 Tipo y Diseño de la investigación

3.1.1 Tipo de investigación

La presente investigación es correlacional se basa en medir el grado de asociación entre variables dependientes y1 y y2. ninguna de estas variables es fijada o manipulada por el investigador. Requiere hipótesis y aplicación de prueba estadística. Universidad Privada de Tacna (2017).

El objetivo de la investigación se basa en asociar variables a través de un patrón predecible para una población o un grupo. Hernández Sampieri, Fernández Collado, y Baptista Lucio (2014).

Este tipo de estudios tiene como objetivo conocer la relación o grado de asociación que exista entre dos o más conceptos, categorías o variables en una muestra o contexto en particular. En ocasiones sólo se analiza la relación entre dos variables, pero con frecuencia se ubican en el estudio vínculos entre tres, cuatro o más variables.

3.1.2 Nivel de Investigación

Es el nivel de investigación es comprensivo corresponde a una investigación cuyos objetivos implican explicar, predecir o proponer. Universidad Privada de Tacna (2017).

En su investigación el nivel busca explicaciones y relaciones entre eventos. En este nivel se generan las teorías, las predicciones y las propuestas. Barrera J. H., (2010).

3.2 Población y/o muestra de estudio

3.2.1 Población

La población está constituida por el conjunto de tweets publicados por las personas del Perú que interactúan con Twitter relacionados a la pandemia COVID-19. La población estará conformada por 5,000 tweets.

La población se ha adecuado 5,000 tweets, se evidencia porqué los

usuarios de Perú emitieron escasos comentarios de tweets a medida que transcurría los meses de enero, febrero y marzo hasta la actualidad disminuyó en menor número de tweets.

Por otro lado, se evidencia la escasez de tweets que hay en el Perú, por ejemplo, en la fecha 03 de abril del 2022 publicaron solo 49 tweets en ese día. Por día publicaban tweets menores o igual a 100.

Se evidencia que de los 24 departamentos y la Provincia Constitucional del Callao del Perú, la ciudad de Lima es quien más publico tweets, y hay departamentos que no publicaron ningún tweet, la ciudad de Tacna es quien publico menor o igual a 2 tweets en un día y hay días que no publico ningún tweet.

- Por el tiempo limitado de la investigación solo se logró recolectar la información de 5,000 tweets.
- En el mes de enero, solo realice la interacción con la herramienta Knime.
- Analíticos, para familiarizarme con el uso de los nodos.
- En el mes de febrero se investigó e interactuó con los diversos nodos para la limpieza de los tweets.
- En el mes de marzo se interactuó con los nodos de los algoritmos de predicción y se realizó la aplicación de los algoritmos, Árbol de decisión, Naive Bayes y Máquinas de vectores de soporte, y se eligió el mejor algoritmo es Máquinas de Vectores de soporte porque tuvo un mejor desempeño en exactitud y mejor predicción de sentimientos de tweets positivos y negativos, en comparación del algoritmo Naive Bayes predecía los tweets tenían mayormente sentimientos negativos. Por otra parte, el algoritmo Árbol de decisión predecía bastantes tweets tenían sentimientos positivos.
- En el mes de marzo se realizó la actividad de descarga de los tweets con la herramienta Knime Analíticos de los meses de marzo y abril. También se realizó la limpieza de los tweets, se eliminó los emoticones, los URL, los signos de puntuación, etc. Por otro lado, en los archivos Microsoft Excel se clasifico manualmente los 5000 tweets asignándole el sentimiento positivo o negativo. Por otra parte, en el proceso de descargar, mientras más tweets mayormente se demora en descargar. A continuación, se mencionan las fechas de descarga de los tweets. tal como se puede apreciar en la tabla 2.

Tabla 2.

Fechas de descarga de tweets

Fechas de descarga de tweets		
3/04/2022	21/04/2022	22/04/2022
23/04/2022	24/04/2022	25/04/2022
26/04/2022	28/04/2022	29/04/2022
14/04/2022	15/04/2022	7/03/2022
6/03/2022	5/03/2022	10/03/2022
8/03/2022	9/03/2022	11/03/2022
12/03/2022	14/03/2022	

Nota.La tabla muestra tweets de las fechas que se han descargaron. Fuente: Elaboración propia.

3.2.2 Muestra

El muestreo de conveniencia consiste en la elección de una muestra cuyas características sean similares a las de la población objetivo. En este tipo de muestreo la “representatividad” la determina el investigador de modo subjetivo. Casal y Mateu (2003).

Para la presente investigación, utilizamos 1,000 tweets y no 879 tweets porque el muestreo por conveniencia ocupa mayor cantidad de muestra que la calculada por la fórmula propiamente dicha.

Los criterios considerados para la obtención de la muestra de estudio, son los siguientes:

N = Tamaño de la población

p = 50 % probabilidad a favor

q = (1-p) = 50 % probabilidad en contraα =Error Alfa = 0,05

1-α = nivel de confianza = 0,95

Z(1-α) = Según tabla de distribución normal = 1,96

e = Margen de error = 3 % = 0,03

El tamaño de la muestra se obtiene de la siguiente aplicación de la fórmula:

$$n = \frac{N * {Z_{1-\propto}}^2 * p * q}{d^2 * (N-1) + {Z_{1-\propto}}^2 * p * q}$$

$$n = \frac{5000 * (1{,}96)^2 * (0{,}5) * (0{,}5)}{(0{,}03)^2 * (5000-1) + (1{,}96)^2 * (0{,}5) * (0{,}5)}$$

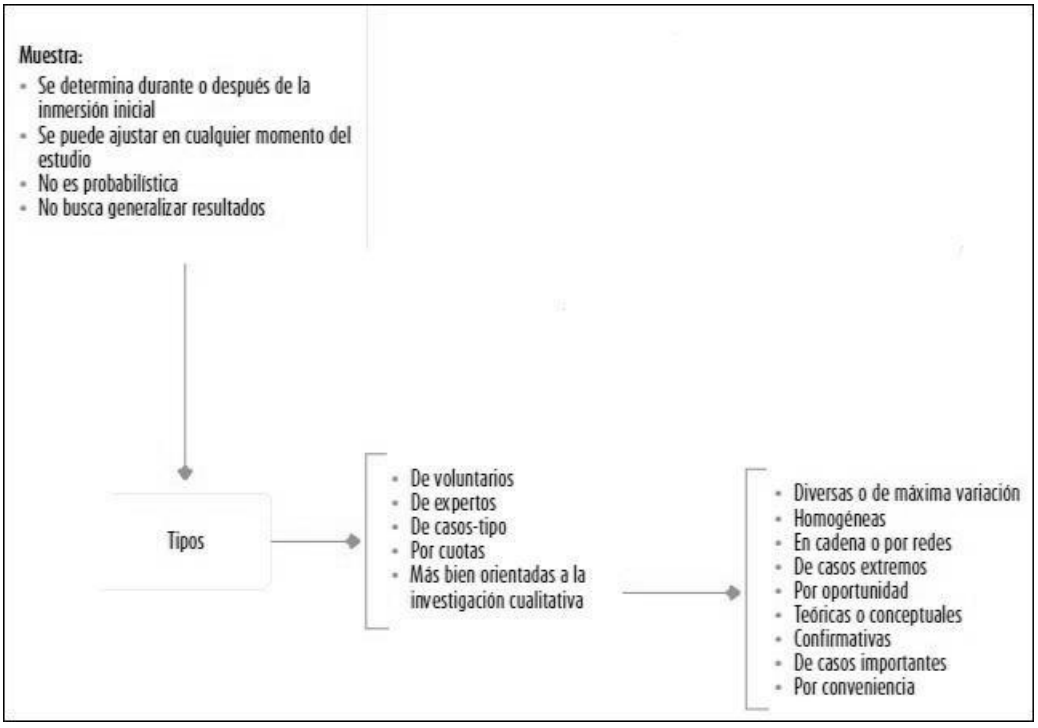
$$n = \frac{4{,}802}{5{,}4595}$$

$$n = 879$$

La presente investigación está orientada en una muestra por conveniencia, en la figura 8. Muestra por conveniencia, se muestra los distintos tipos de muestra.

Figura 8.

Muestra de Tipos Por conveniencia



*Nota.*La figura muestra tipos por conveniencia. Fuente: (Hernández Sampieri, Fernández Collado y Baptista Lucio, 2014).

Definen a la Muestras por conveniencia estas muestras están formadas por los casos disponibles a los cuales tenemos acceso. Hernández Sampieri et al. (2014) y Battaglia (2008a). Tal fue la situacion de Rizzo (2004), quien no pudo ingresar a varias empresas para efectuar entrevistas a profundidad en niveles gerenciales, acerca de los factores que conforman el clima organizacional, y entonces decidió entrevistar a compañeros que junto con ella cursaban un postgrado en desarrollo humano y eran directivos de diferentes organizaciones.

Para obtener la muestra aplicó un muestreo por conveniencia incluyendo los tweets que encontró entre el 03 hasta el 10 de abril y del 29 de mayo hasta 5 de junio, con respecto al candidato Pedro Pablo Kuczynski 22,154 y 53,284 tweets para la primera y segunda vuelta. (Poma Pancaya, 2020).

Mientras para la candidata Keiko Fujimori 13,860 y 27,347 tweets para la primera y segunda vuelta.

En definitiva, la presente investigación para obtener la muestra se aplicó un muestreo por conveniencia incluyendo dentro de las mismas los tweets que se encontraron entre el 04 de marzo, el 03 de abril, el 14 de abril hasta el 15 de abril, del 22 de abril hasta el 26 de abril y del 28 de abril hasta el 29 de abril, constituida por los 1,000 tweets los que se utilizaron para el “Análisis de comportamiento de tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos, 2022”.

Para la aplicación del clasificador la cantidad de tweets utilizados para la creación del modelo fue de 1,000 tweets. fraccionados en: 630 comentarios positivos y 370 comentarios negativos. Y para ejecutar el modelo utilizó 80 % tweets para entrenamiento y 20 % para la prueba (Mamani Coaquira et al.,2021).

Nuestro estudio para la aplicación del clasificador usamos 1000 tweets divididos de la siguiente manera: 700 tweets positivos y 300 tweets negativos. Y para la ejecución del modelo utilizamos el 900 de los tweets para el entrenamiento y 100 para la prueba.

Con relación a la muestra de los 1,000 tweets para el cross validation se utilizó 900 tweets para entrenamiento y 100 tweets para prueba. De los 100 tweets de prueba se obtuvo las métricas del algoritmo SVM con distintos resultados en cada una de las iteraciones. También es como un bucle para

correr, ejecutar en mismo tiempo y no estar repitiendo uno por uno y reseteando a cada momento el particionamiento.

3.3 Operacionalización de variables

Consiste en desglosar y analizar cada una de las variables que componen el estudio, desde las más generales hasta las más específicas. Su exposición completa nos mostrará la realización de teoría e investigación. Sus variables se definen conceptualmente, mostrando sus dimensiones y indicadores como se muestra en la Tabla 3.

Tabla 3.

Operacionalización de variables

Variable	***Definición operacional***	***Indicador***	***Dimensión***
Variable independiente			
Minería de textos	Es el proceso de derivar información nueva de textos.	-Número de tweets a ser analizados. -Rendimiento del algoritmo. -Eficiencia. de algoritmo.	Técnica seleccionada para el análisis de los tweets.
Variable dependiente			
Análisis del comportamiento de los tweets	Se refiere a la acción de reconocer y clasificar la conducta de los tweets en base a distintos aspectos	-Porcentaje de tweets positivos. -Porcentaje de tweets negativos.	Comportamiento evidenciado de los tweets relacionados a la pandemia COVID-19.

- Identificación de Variables

Es una propiedad que puede variar y cuya variación es susceptible de medirse. Sampieri (2014).

Las variables son los elementos que vamos a medir, controlar y estudiar dentro del problema formulado, de allí que se requiera la posibilidad real y cierta de que se puedan cuantificar. Ese trabajo de manejarlas, insertarlas encuadros, manipularlas en los instrumentos del caso se llama Operacionalización.

- Descripción de los indicadores

A continuación, se explicará los indicadores, según la variable de investigación.

a. Variable independiente "Minería de textos".

- Número de tweets a ser analizados.
- Cantidad de información.
- Disponibilidad de información.
- Velocidad de descarga de tweets.
- Velocidad de preprocesamiento.

b. Variable dependiente "Análisis del comportamiento de los tweets".

- Porcentaje de tweets positivos.
- Cantidad de tweets positivos.
- Rendimiento del algoritmo.
- Eficiente algoritmo.
- Porcentaje de tweets negativos.
- Cantidad de tweets negativos.
- Rendimiento del algoritmo.
- Eficiente algoritmo.

En el estudio de "Análisis del comportamiento de los tweets, de la participación electoral de los principales candidatos a la presidencia del Perú utilizando minería de textos en el año 2016". Se encontró las siguientes V.I y V.D y sus indicadores. Pancaya (2016).

- Variable independiente: Minería de textos y su Indicadores: Número de tweets a ser analizados.
- Variable dependiente: Análisis del comportamiento de los tweets y sus indicadores: Porcentaje de tweets positivos y Porcentaje de tweets negativos.

3.4 Técnicas e instrumentos para la recolección de datos

- Instrumentos:

En instrumentos para el efecto de la investigación los instrumentos son los siguientes:

- Herramientas de software Knime Analytics.
- API de Twitter.

3.4.1 Pruebas estadísticas

- Prueba z (zTest)

Para aplicar la Prueba Z los datos deben cumplir las siguientes condiciones:

El tamaño de la muestra debe ser mayor o igual a 30 unidades. De ser menor se utiliza la prueba t de student.

La prueba Z basa en la distribución Normal Estándar. El valor calculado de Z es según la ecuación (9).

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (9)$$

3.5 Procesamiento y análisis de datos

- Extraer tweets utilizando la API de Twitter.
- Realizar la limpieza de los datos.
- Escoger un conjunto de tweets para entrenar el modelo.
- Realizar la clasificación de tweets.
- Analizar e interpretar los resultados.

3.5.1 Metodología de análisis de sentimiento

3.5.1.1 Proceso de análisis de sentimiento

El análisis de sentimiento tiene la finalidad de extraer opiniones de textos, clasificar emociones expresadas en textos posee polaridad positiva y negativa. Pang y Lee (2008).

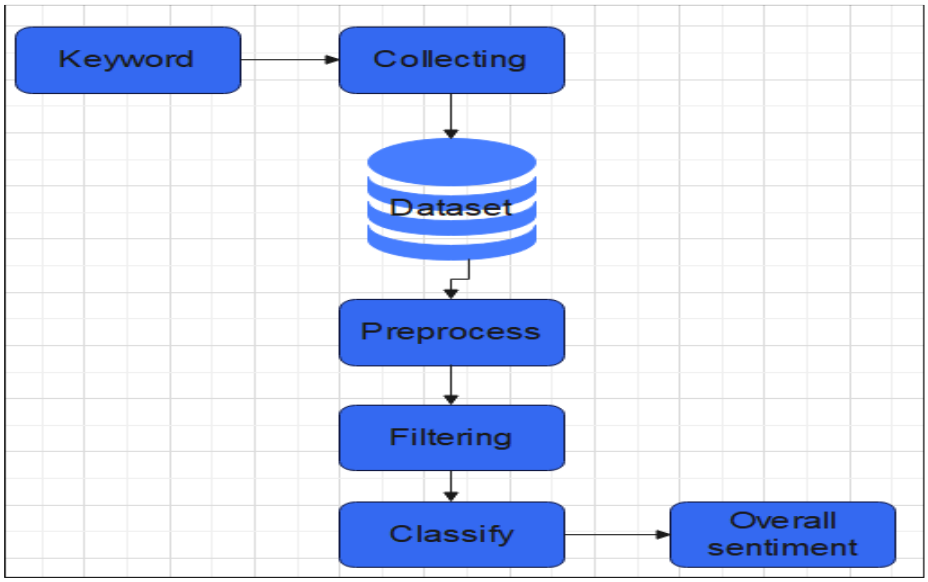
Define análisis de sentimiento es el proceso de inferir la información afectiva generalmente una etiqueta emocional. Cambria, Grassi, Hussain y Havasi (2012).

En el presente trabajo de investigación se empleó la metodología Knowledge Discovery in Databases (KDD) Esta propuesta de Análisis de sentimiento (AS) está constituida en 5 etapas: selección, preprocesamiento, transformación, minería de datos, interpretación/evaluación. Fayyad Piatetsky Shapiro y Smyth, (1996).

El coleccionismo, preprocesamiento, Filtración, clasificando y una etapa adicional de interpretación. En la figura 5. Se muestra un flujograma del proceso general de análisis de sentimiento en KDD para hacer el análisis de tweets, en el cual se puede apreciar que el presente trabajo se desarrolla de manera cascada, por otro lado, existe una etapa previa del proceso en el grafico se denomina keyword de acuerdo a esto se realiza una selección de palabras clave: de acuerdo a estas palabras se seleccionara los tweets a extraer. Alhumoud (2015). tal como se puede apreciar en la Figura 9.

Figura 9.

Flujograma del proceso general de análisis de sentimiento



Nota.La figura muestra el Flujograma del proceso general de análisis de sentimiento. Fuente: Elaboración propia.

Seguidamente se muestra las técnicas más empleadas para cada una de las etapas del proceso de análisis de sentimiento.

3.5.1.2 Etapas

a. Recolección

Para esta etapa se realiza la recolección de los tweets y lo realizamos a través de la plataforma Knime Analytics, para ello necesitamos una cuenta de Twitter del cual obtenemos la API key, API secret, Access token y Access token secret.

Para realizar esta tarea seguidamente se muestra los siguientes pasos:

Luego configuramos el Nodo Twitter API Conector y nos muestra la siguiente ventana en el cual completamos los datos requeridos API key, API secret, Access token y Access token secret.

Luego configuramos el Nodo Twitter Search y nos muestra la siguiente ventana en el cual escribimos la consulta pandemia Perú y seleccionamos la cantidad de filas es 100, seleccionamos todos los campos Tweet, User Location, Time, etc.

Luego configuramos el Nodo Row Filter y nos muestra la siguiente ventana en el cual seleccionamos la columna User-Location y seleccionamos un criterio coincidente en este caso es *Perú, este criterio es para recolectar todos los tweets de Perú.

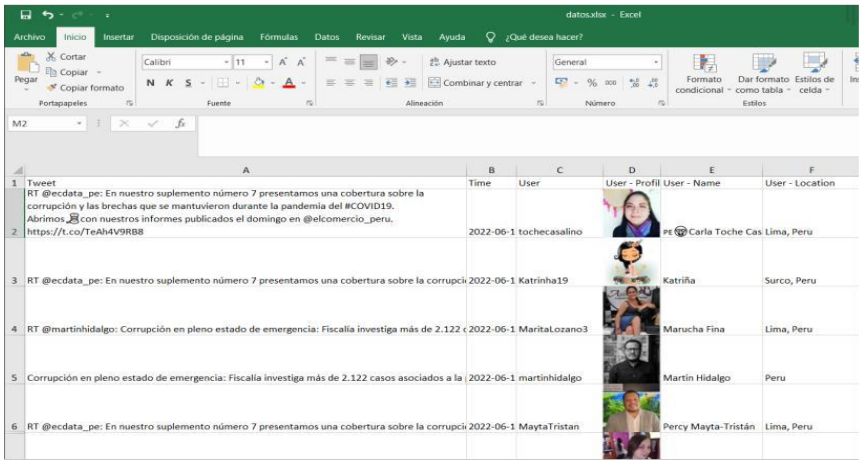
Luego configuramos el Nodo Excel Writer y nos muestra la siguiente ventana en el cual guardamos en la ubicación C:\Users\USUARIO\Desktop\etapas\datos.xlsx el archivo en formato Excel y le ponemos el nombre datos.xlsx y sobre escribimos.

Observamos el resultado de la recolección de tweets.

El siguiente paso es realizar el etiquetado de los tweets según su sentimiento positivo o negativo. tal como se puede apreciar en la Figura 10 y Figura 11.

Figura 10.

Ejemplo etiquetado de tweets

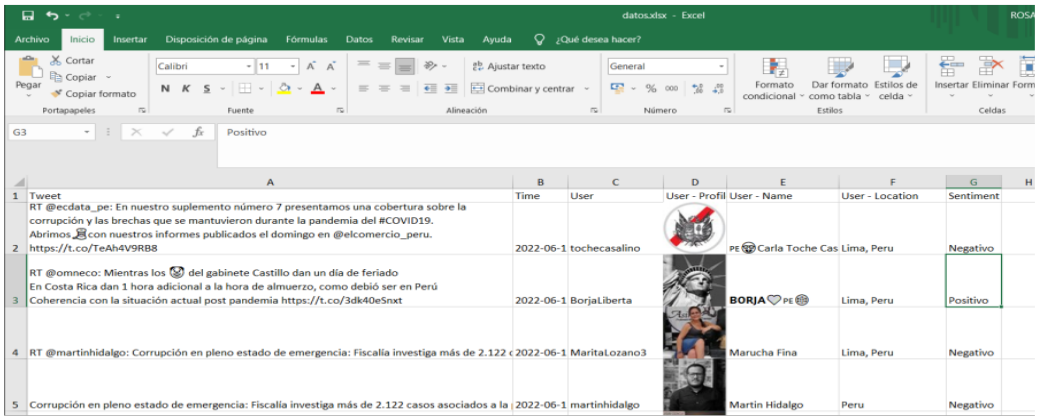


Nota.La figura muestra un ejemplo de etiquetas de tweets. Fuente:

Elaboración propia.

Figura 11.

Ejemplo etiquetado Manual de tweets



	A	B	C	D	E	F	G
1	Tweet	Time	User	User - Profil	User - Name	User - Location	Sentiment
2	RT @ecdata_pe: En nuestro suplemento número 7 presentamos una cobertura sobre la corrupción y las brechas que se mantuvieron durante la pandemia del #COVID19. Abrimos con nuestros informes publicados el domingo en @elcomercio_peru. https://t.co/TeAh4V9RB8	2022-06-1	tochecasalino		Carla Toche Cas	Lima, Peru	Negativo
3	RT @omneco: Mientras los del gabinete Castillo dan un día de feriado En Costa Rica dan 1 hora adicional a la hora de almuerzo, como debió ser en Perú Coherencia con la situación actual post pandemia https://t.co/3dk40eSnxt	2022-06-1	BorjaLiberta		BORJA	Lima, Peru	Positivo
4	RT @martinhidalgo: Corrupción en pleno estado de emergencia: Fiscalía investiga más de 2.122 c	2022-06-1	MaritaLozano3		Marucha Fina	Lima, Peru	Negativo
5	Corrupción en pleno estado de emergencia: Fiscalía investiga más de 2.122 casos asociados a la	2022-06-1	martinhidalgo		Martin Hidalgo	Peru	Negativo

Nota.La figura muestra un ejemplo de etiqueta manual de tweets positivo o negativo. Fuente: Elaboración propia.

b. Preprocesamiento

En esta etapa de preprocesamiento, debido es que en esta etapa después de haber extraído la data empezará a transformarse en información. “AS es el arte de la selección de características”, es por lo cual esta etapa es la más importante de AS. Bosch (2013).

En esta etapa de preprocesamiento se realiza la transformación y se crea vectores de bits 1 y 0 para todos los documentos 1 significa que existe ese término y 0 significa que no existe ese término en el documento y así sucesivamente para todos los demás documentos, estos términos se ordenan por columnas y las filas son los documentos.

Luego se colorea por sentimiento se pone una etiqueta de color rojo para negativos y una etiqueta verde para positivos.

- Punctuation Erasure: Este nodo Punctuation Erasure elimina todos los caracteres de signos de puntuación de los términos contenidos en los documentos.
- Number Filter: Este nodo de Number Filter filtra todos los términos contenidos en los documentos que constan de dígitos, incluidos los separadores decimales “,” o “.” Y “+” o”-”.
- N Chars Filter: Este Nodo de N Chars Filter filtra todos los términos contenidos en los documentos con menos del número N caracteres especificado.

- Stop Word Filter: Este Nodo de Stop Word Filter filtra todos los términos de los documentos, que están contenidos en la lista de palabras vacías especificadas, se puede seleccionar una lista de palabras vacías integrada o personalizada.
- Case Converter: Este Nodo Case Converter convierte caracteres alfanuméricos a mayúscula o minúsculas.
- Snowball Stemmer: Este Nodo Snowball Stemmer asigna distintas maneras de la misma palabra una “raíz” común; por ejemplo, el Stemmer español asigna conectado, conectando, conectivo, conexiones, conexión para conectar. Por lo tanto, una búsqueda de conectado también encontraría documentos que solo poseen las otras formas.
- Bag of Words Creator: Este nodo Bag of Words Creator crea una bolsa de palabras (BoW) de un conjunto de documentos. Un BoW constituye de al menos una columna que contiene los términos que aparecen en el documento correspondiente.
- Term To String: Este nodo Term to String convierte los términos de una columna en cadenas y adjunta una nueva columna que contenga estas cadenas. Las etiquetas de los términos se pierden.
- GroupBy: Este nodo GroupBy agrupa las filas de una tabla por los valores únicos en las columnas del grupo seleccionado.
- Row Filter: Este nodo Row Filter permite el filtrado de filas según ciertos criterios.
- Reference Row Filter: Este nodo Reference Row Filter permite filtrar de la primera tabla empleando segunda tabla como referencia.
- TF: Este nodo TF calcula la frecuencia relativa del término (TF) de cada término según cada documento y agrega una columna que contiene el valor de TF. El valor se calcula dividiendo la frecuencia absoluta de un término según un documento por el número de todos los términos de ese documento.

c. Filtrado

En esta etapa de filtrado se realiza la limpieza de la data, se enumera algunas actividades para esta etapa como son: Eliminación de palabras repetidas, eliminación de emoticones, eliminación de signos de puntuación, Eliminación de stop words o palabras que no tienen significancia al texto y

Normalización o intercambio de elementos no utilizables por otros que presten mayor ayuda a las posteriores etapas. Alhumoud (2015).

Los filtros más empleados en esta etapa es la eliminación de menciones @, hashtags #, Urls y signos de puntuación. R-Moreno, Cuesta y F. Barrero (2013).

d. Clasificación

En esta etapa se emplea el algoritmo de Máquinas de Vector de soporte. Primeramente, se realiza el particionamiento con la validación cruzada de 10 iteraciones 90 % de tweets para entrenamiento y 10 % de tweets para la prueba, luego se aplica el aprendizaje (Learner) SVM este modelo se guarda en un archivo Model Writer y la predicción (Predictor) SVM el cual obtiene como resultado una clasificación de la predicción.

- Support Vector Machine (SVM)

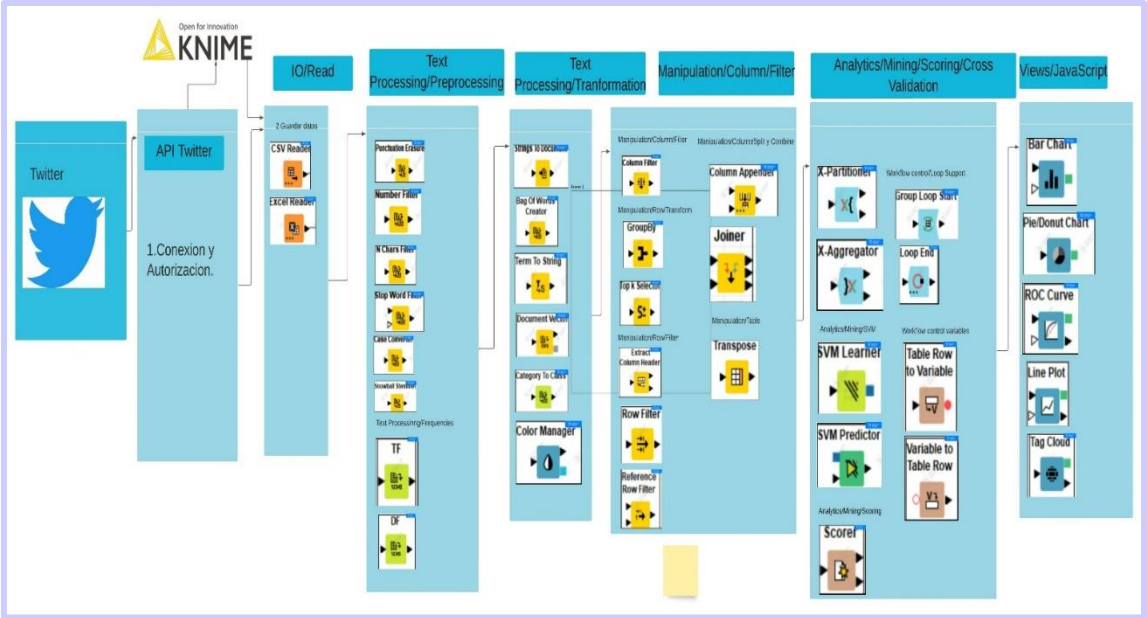
Se define en la generación de hiper planos capaz de asignar correctamente los documentos a una clase c. Para esto se calcula la derivada de la fórmula del Lagrangiano Dual para hallar los valores a que son necesarios para calcular los hiper planos. Bosch (2013).

Se obtienen mejores resultados en el proceso de clasificación empleando SVM. Bravo Marquez, Mendoza y Poblete (2013). tal como se puede apreciar en la Figura 12.

3.5.1.3 Diagrama de Arquitectura

Figura 12.

Diagrama de la arquitectura de Análisis de Sentimientos



Nota. La figura muestra el diagrama de la arquitectura de análisis de sentimiento.

Fuente: Elaboración propia.

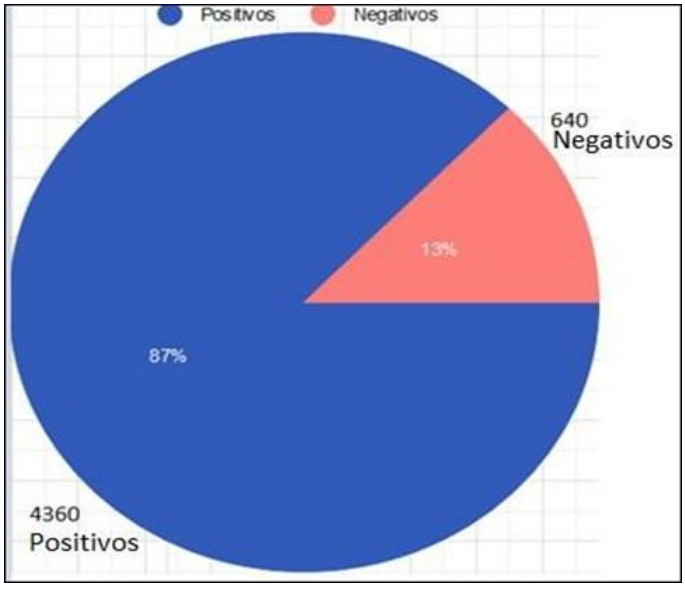
CAPÍTULO IV: RESULTADOS

4.1 Resultados

Se puede observar en la Figura 13 los resultados del análisis del comportamiento de los tweets relacionados a la pandemia COVID-19. Todas las figuras se pueden observar en el anexo 3.

Figura 13.

Comportamiento de tweets



Nota.La figura muestra el comportamiento de tweets.

Análisis e interpretación de la Figura 13

Como podemos observar en la Figura 13 el 87 % (4360) de los tweets son positivos, mientras que el 13 % (640) de los tweets son negativos.

4.1.1 Elaboración de clasificador de sentimientos

4.1.1.1 Desarrollo de extracción de los datos

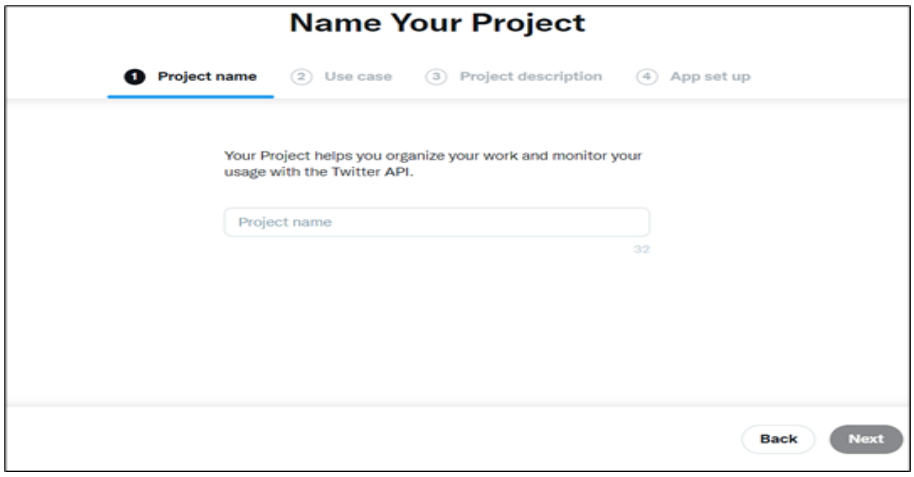
4.1.1.1.1 Recopilación de datos

Los comentarios de los tweets se encuentran almacenados en la base de datos de Twitter, para extraer los comentarios se empleó la API de Twitter. Para este paso es un requisito muy importante tener una cuenta de Twitter.

- Ingresamos al sitio web de Twitter a la siguiente URL https://apps.twitter.com/, luego creamos una cuenta, luego registramos los datos requeridos. Se puede observar en la Figura 14.

Figura 14.

Crear nuevo proyecto en Twitter

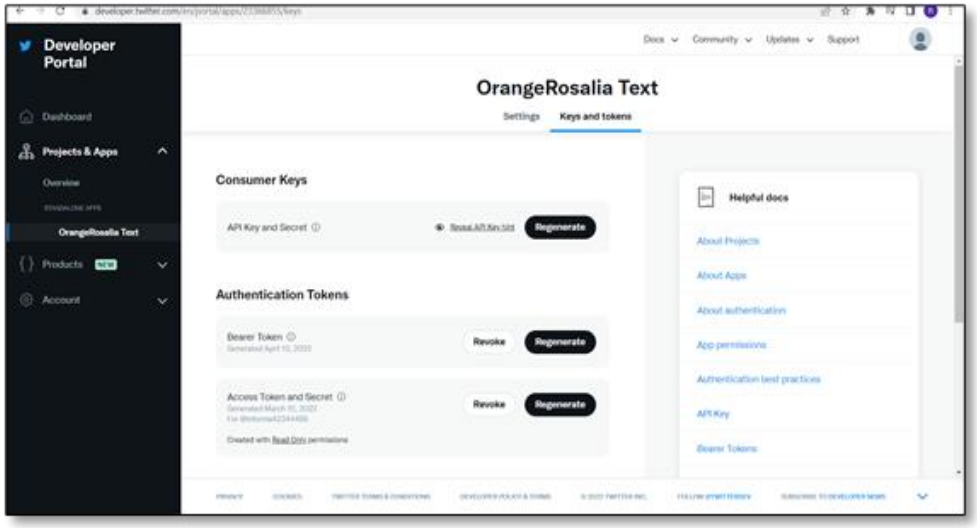


Nota.La figura muestra un formulario para crear un nuevo proyecto en twitter. Fuente: Elaboración Propia.

- Se puede observar en la Figura 15. En este caso mi proyecto se denomina OrangeRosalia Text, este API de Twitter es usada para la extracción de los tweets.

Figura 15.

Keys and tokens



Nota.La figura muestra una Interfaz para generar clave Keys and tokens. Fuente: Elaboración propia.

Estos son los Consumer Keys y Authentication Tokens:

- API key.
- API secret.
- Access token.
- Access token secret.

Estas cuatro API de Twitter sirven para conectarnos en la plataforma de Knime Analytics.

- Se puede observar en la Figura 16 el Access Token and Secret.

Figura 16.

Access token and secret

Nota.La figura muestra la Interfaz para generar claves Access Token and Secret. Fuente: Elaboración propia.

- Se puede observar en la Figura 17 el API key and secret.

Figura 17.

Creación de API key and secret

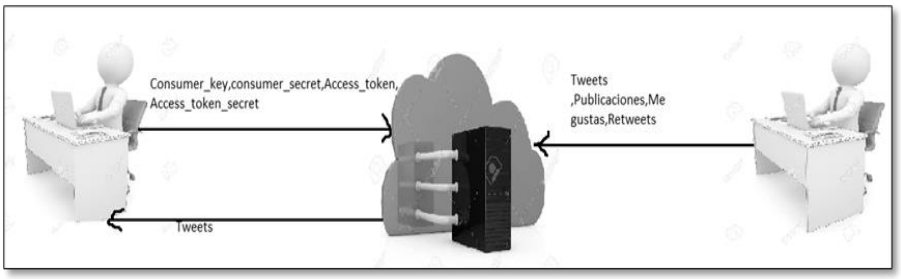
Nota.La figura muestra la clave API key and Secret.Fuente: Elaboración propia

4.1.1.1.2 Funcionamiento del API de Twitter

Se puede observar en la Figura 18 el funcionamiento de Search API Twitter.

Figura 18.

Search API Twitter

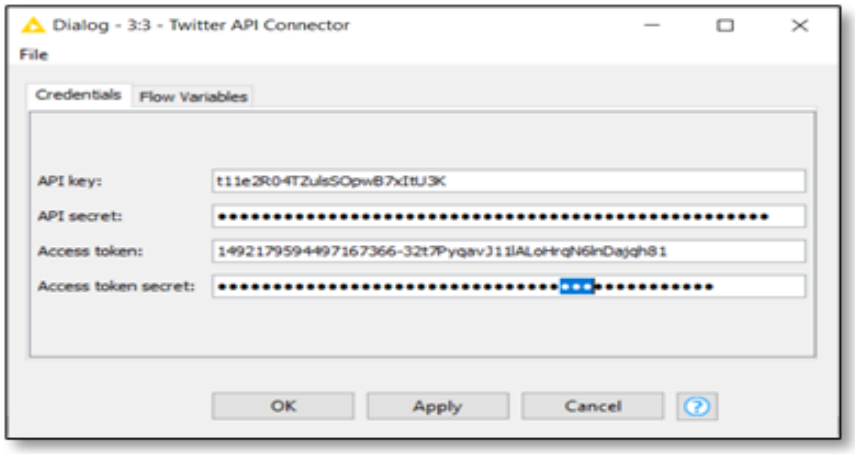


Nota.La figura muestra el funcionamiento de la API twitter para extraer tweets. Fuente: Elaboración propia

- Podemos observar en la Figura 19 la conexión de la API para extraer tweets.
- Todos los nodos para extraer tweets se pueden observar en el anexo 2.

Figura 19.

Conector de API Twitter



Nota.La figura muestra una Interfaz de claves de API de twitter para conectarse a Twitter. Fuente: Elaboración propia.

4.1.1.1.3 Proporción de la muestra

Se descargaron los tweets desde el 4/03/2022 hasta 30/04/2022. tal como se puede apreciar en la tabla 4, tabla 5 y tabla 6.

Tabla 4.

Fechas de descarga de tweets

Tema	Fecha de descarga de tweets
Pandemia	Del 04 de marzo hasta 30 de abril 2022

Nota.La tabla muestra una fecha 04/03/2022 que se descargaron tweets. Fuente: Elaboración propia

Tabla 5.

Cantidad de tweets

Cantidad de *tweets*	**Fecha de descarga de *tweets***
5000	Del 04 de marzo hasta el 29 de abril 2022

Nota.La tabla muestra el rango de fechas 04/03/2022 - 29/04/2022 que se descargaron tweets. Fuente: Elaboración propia

Tabla 6.

Tweets de prueba

Cantidad de *tweets*	**# De tweets de prueba**	**# De *tweets* de entrenamiento**
De 1000 *tweets* 5000	100	900

Nota.La tabla muestra 100 tweets para prueba y 900 tweets para entrenamiento. Fuente: Elaboración propia

4.1.1.2 Desarrollo de limpieza de datos

Los tweets tienen emoticones, signos de puntuación, números, URL, palabras vacías, asimismo todos estos se deben eliminar y dejar limpio el texto. Se deben excluir todo lo que no contribuya a la investigación. La limpieza se realizó en los nodos de Knime.

4.1.1.3 Procedimiento de limpieza

- Eliminar todos los URL.
- Eliminar todos los usuarios de Twitter.
- Eliminar todos los acentos, para evitar problemas.
- Eliminar emoticones.
- Eliminar signos de puntuación.
- Convertir todas las palabras a minúsculas.
- Reducir las palabras.

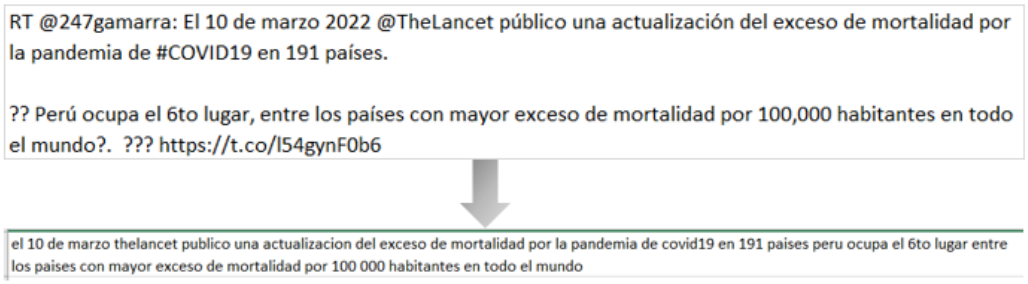
- Eliminar los caracteres Unicode.
- Eliminar los stopwords o palabras vacías.
- Stemmar las palabras de tweets.
- Eliminar todos los números.

Podemos observar en la Figura 20 un ejemplo de limpieza de un tweet.

El nodo de limpieza de los tweets se encuentra en el anexo 2. tal como se puede apreciar en la Figura 20.

Figura 20.

Limpieza de tweets



Nota.La figura muestra un ejemplo de limpieza de tweets. Fuente: Elaboración propia.

4.1.1.4 Desarrollo de aprendizaje

En este trabajo de investigación se empleó el algoritmo Máquinas de Vectores de Soporte. Estos algoritmos usan métodos de clasificación y regresión para construir modelos de predicción.

4.1.1.5 Algoritmo de Máquinas de Vector Soporte

a. Beneficios de las Máquinas de Vectores de Soporte estos son:

- Brindan una buena precisión.
- Ejecutan predicciones muy rápidas.

- Emplean menos memoria.
- Desempeñan muy bien con un claro margen de separación.
- Luego de haber entrenado el clasificador de los tweets de entrenamiento analizó la matriz de confusión en el cual encontramos los resultados.
  Hay dos tipos de equivocación:
- El Algoritmo de Máquinas de Vector de Soporte predice valores falsos positivos, los valores reales en realidad son negativos. De la misma manera predice valores falsos negativos, los valores reales en realidad son positivos. Se puede observar en la Figura 21.
- Los errores se muestran en la diagonal de color rojo y la diagonal de color verde muestra los aciertos.

Figura 21.

Matriz de confusión

Nota.La figura muestra una matriz de confusión. Fuente: (Health Big Data, 2019)

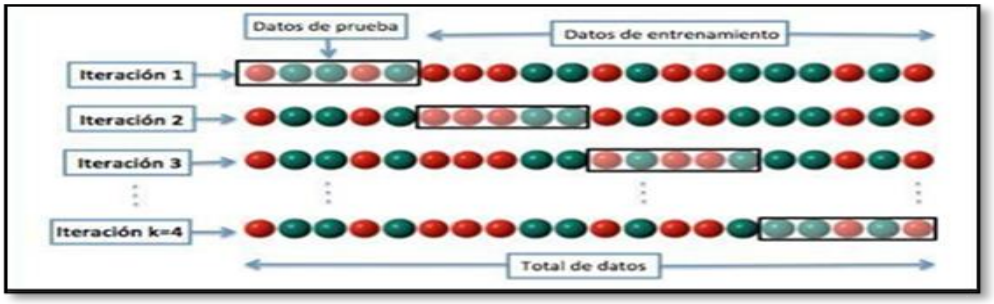
4.1.1.5.1Cross Validation

Para este trabajo de investigación se seleccionaron 1000 tweets de los cuales 700 son positivos y 300 son negativos, se asignó el 90 % de estos tweets para training y el 10 % de estos tweets para testing. El proceso se realizó en 10 iteraciones usando el método cross validation de k iteraciones, se puede observar en la Figura 22.

Luego de haber entrenado los tweets, se observó la matriz de confusión donde se observan todos los resultados, los tweets correctamente clasificados (Correct classified), incorrectamente clasificados (Wrong classified), el error (Error), la exactitud (Accuracy) y la Kappa de cohens (Cohen's kappa).

Figura 22.

Cross Validation



Nota. La figura muestra validación cruzada de k iteraciones. Fuente: (domenech91, 2011)

4.1.1.5.2 Característica Operativa del Receptor (ROC)

La curva ROC se calcula de la sensibilidad o la razón de verdaderos positivos (VPR) y el ratio o razón de falsos positivos (FPR).

- Sensibilidad o razón de verdaderos positivos (VPR).
- Ratio o razón de falsos positivos (FPR) según la ecuación (10) y la ecuación (11).

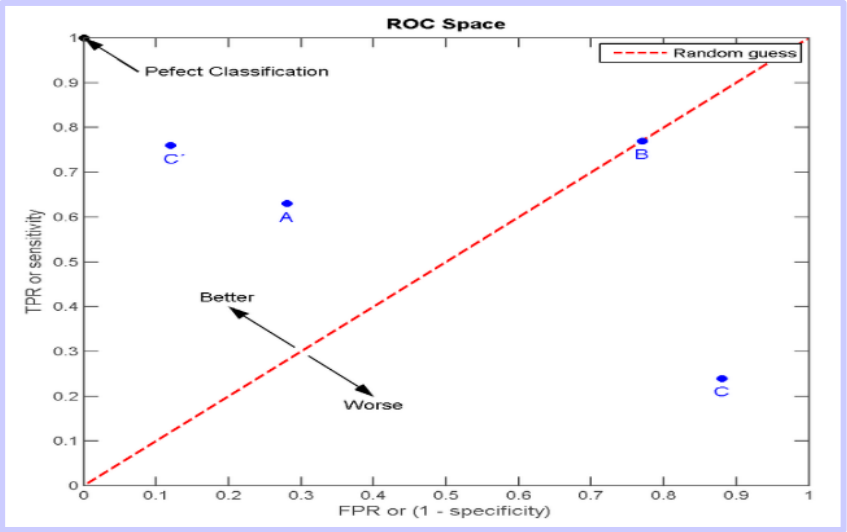
$$VPR = \frac{VP}{P} = \frac{VP}{VP+FN} \quad (10)$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP+VN} \quad (11)$$

Si el área posee un valor 1, este es un valor perfecto. Podemos observar en la Figura 23.

Figura 23.

Curva ROC



Nota.La figura muestra la curva Roc. Fuente: Elaboración propia.

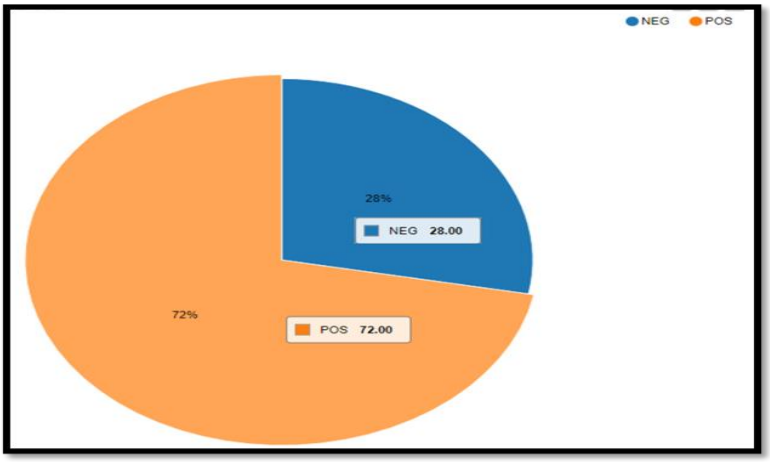
4.1.2 Análisis de las métricas de evaluación del clasificador

Para el análisis de las métricas de evaluación del clasificador, se hizo uso de 1000 tweets, de estos se seleccionaron 900 tweets para entrenamiento y la diferencia para la prueba.

Se pueden observar en la Figura 24. Los valores de tweets positivos y negativos obtenidos del Algoritmo de Maquina de Vectores de Soporte. Se obtuvo como resultado de la última iteración N°10, el 72 % tweets positivos y 28 % negativos.

Figura 24.

Tweets de prueba 100



Nota.La figura muestra los 100 tweets de prueba. Fuente: Elaboración Propia

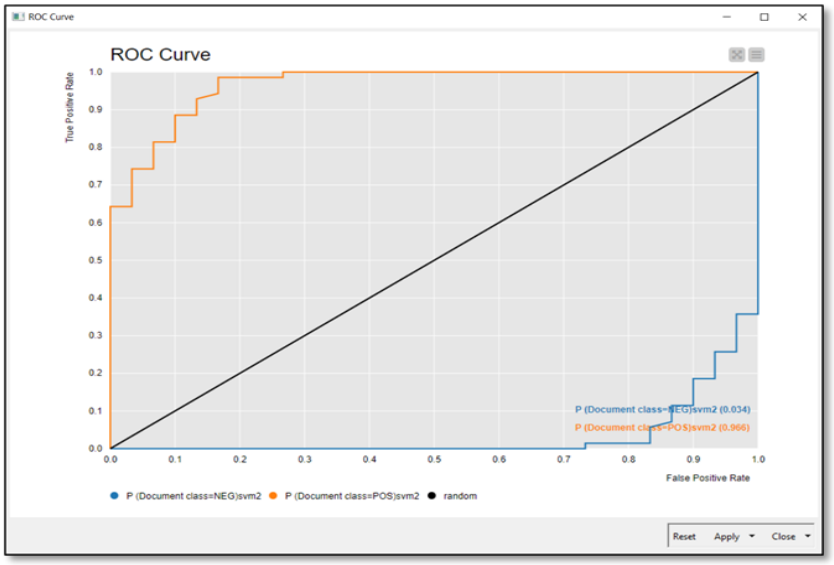
Análisis e Interpretación de la Figura 24

Se observan en la Figura 24. El modelo predictivo realizó una clasificación del total de los tweets, con mayor frecuencia hay 72 tweets positivos su porcentaje es 72 %, a diferencia con menor frecuencia hay 28 tweets negativos su porcentaje es 28 %.

En base a la matriz de confusión en cada iteración de cross validation se obtuvo las razones de verdaderos positivos (VPR) o sensibilidades y las razones de falsos positivos (FPR), se puede observar en la Figura 25.

Figura 25.

Curva ROC



Nota.La figura muestra la Curva Roc.Fuente: Elaboración Propia

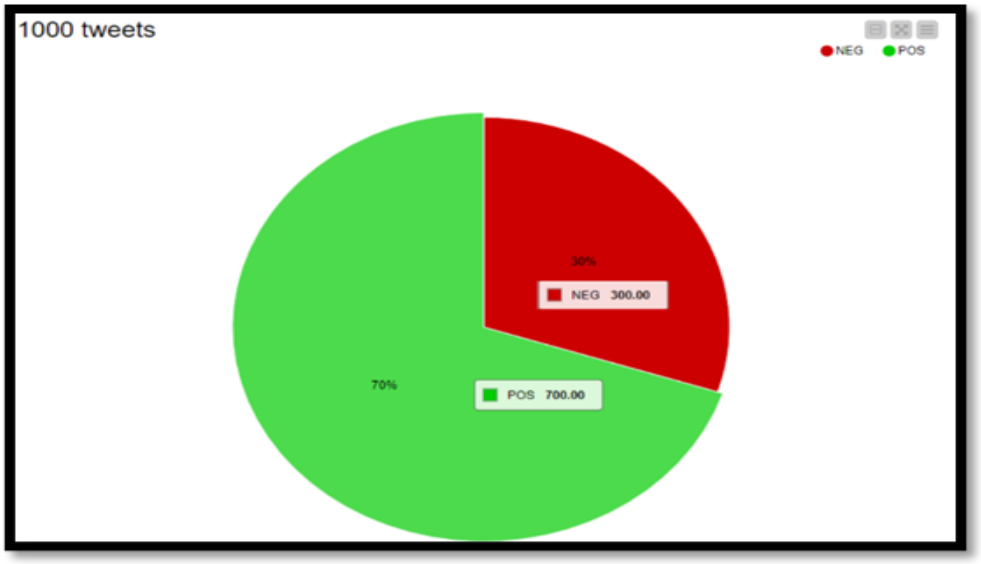
Análisis e Interpretación de la Figura 25

Se puede observar en la Figura 25. La curva ROC de la última iteración en el eje “y” la línea de color naranja tiene un área bajo la curva de 0,966 en porcentajes equivale 96,6 % y este valor es muy cercano a 1, esto quiere decir que es un valor perfecto.

Luego de realizar la transformación de los 1000 tweets estos se encuentran listos para el modelo predictivo, se puede observar en la Figura 26.

Figura 26.

Entrenamiento del modelo



Nota.La figura muestra 1000 tweets para entrenamiento del modelo. Fuente: Elaboración Propia

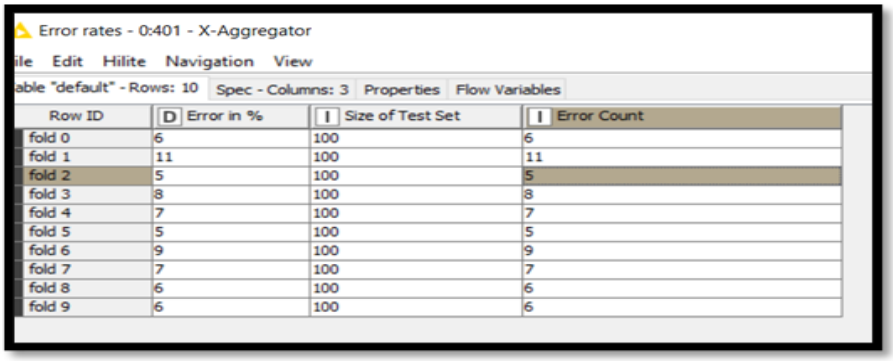
Análisis e Interpretación de la Figura 26

Se puede observar que luego de la transformación se identificó del total de los tweets, al 70 % (700) son positivos, mientras que el 30 % (300) son negativos.

Se puede observar en la Figura 27. La tasa de error de cada una de las iteraciones, de cross validation.

Figura 27.

Tasas de error



Row ID	Error in %	Size of Test Set	Error Count
fold 0	6	100	6
fold 1	11	100	11
fold 2	5	100	5
fold 3	8	100	8
fold 4	7	100	7
fold 5	5	100	5
fold 6	9	100	9
fold 7	7	100	7
fold 8	6	100	6
fold 9	6	100	6

Nota.La figura muestra Tasas de error. Fuente: Elaboración Propia.

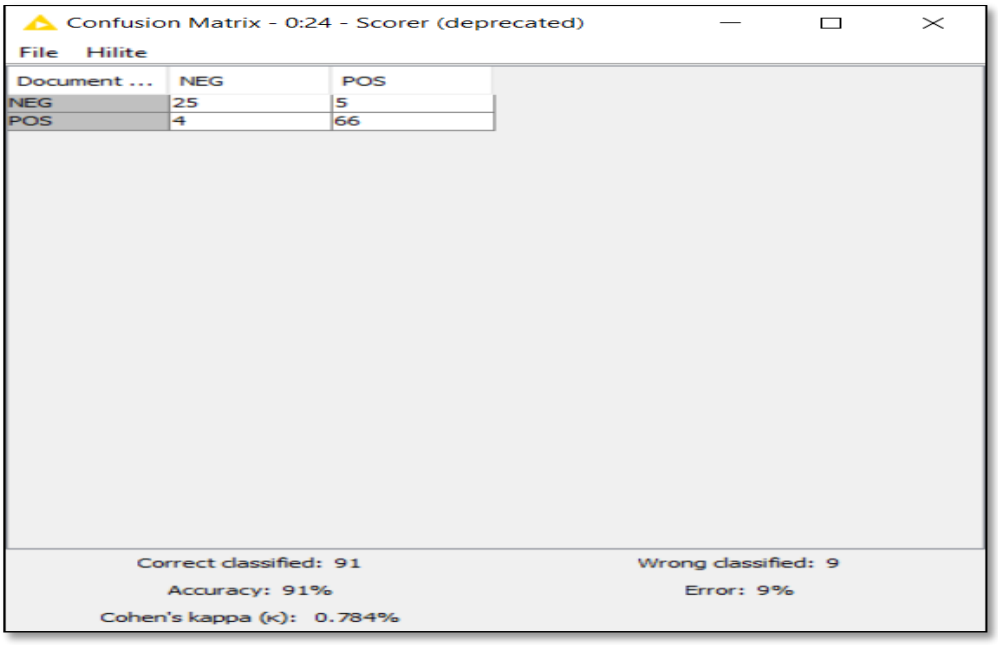
Análisis e Interpretación de la Figura 27

Luego de que se identificaron 700 tweets positivos y 300 tweets negativos, además de ello se aprecia la tasa de error de cada uno de ellos generadas en las iteraciones de cross validation.

Se puede observar en la Figura 28. El registro de los valores de las matrices de confusión de cada una de las iteraciones, también se pueden observar los porcentajes de exactitudes (Accuracy), los tweets correctamente clasificados (Correct classified) y los tweets incorrectamente clasificados (Wrong classified).

Figura 28.

Matriz de confusión de la iteración N° 10



Nota.La figura muestra la matriz de confusión de la iteración Nro. 10. Fuente: Elaboración Propia.

Análisis e Interpretación de la Figura 28

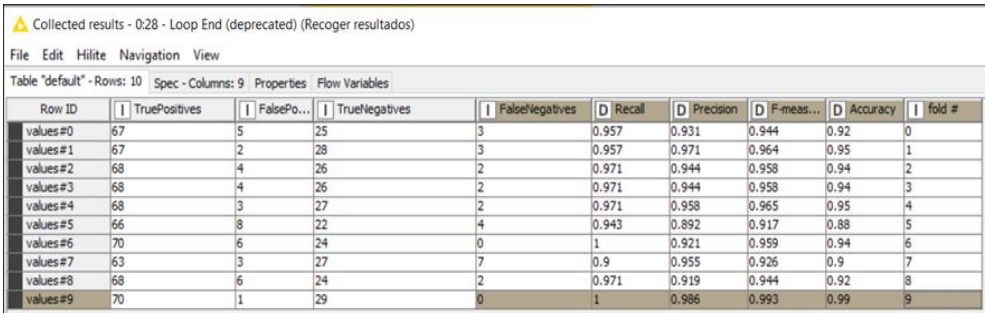
Se puede observar en la Figura 28. Una matriz de confusión de la iteración N° 10, estos valores se obtuvieron de cross validation. Los tweets correctamente clasificados (Correct classified) y los tweets incorrectamente clasificados (Wrong classified), la exactitud (Accuracy), los errores (Error) y la kappa cohens (Cohen's kappa).

- 91 tweets clasificados correctamente (Correct classified).
- 9 tweets clasificados incorrectamente (Wrong classified).
- 91 % tweets de exactitud (Accuracy).
- 9 % tweets de error (Error).
- 0,784 % tweets de kappa de Cohens (Cohen's kappa).

Se puede observar en la Figura 29. El registro de los valores de cada una de las iteraciones y sus respectivas exactitudes (Accurracy) y Precisión (Precisión), en este caso son 10 iteraciones generadas de cross validation.

Figura 29.

Determinación de las métricas de evaluación en cada iteración



Collected results - 0:28 - Loop End (deprecated) (Recoger resultados)

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 9 Properties Flow Variables

Row ID	TruePositives	FalsePo...	TrueNegatives	FalseNegatives	Recall	Precision	F-meas...	Accuracy	fold #
values#0	67	5	25	3	0.957	0.931	0.944	0.92	0
values#1	67	2	28	3	0.957	0.971	0.964	0.95	1
values#2	68	4	26	2	0.971	0.944	0.958	0.94	2
values#3	68	4	26	2	0.971	0.944	0.958	0.94	3
values#4	68	3	27	2	0.971	0.958	0.965	0.95	4
values#5	66	8	22	4	0.943	0.892	0.917	0.88	5
values#6	70	6	24	0	1	0.921	0.959	0.94	6
values#7	63	3	27	7	0.9	0.955	0.926	0.9	7
values#8	68	6	24	2	0.971	0.919	0.944	0.92	8
values#9	70	1	29	0	1	0.986	0.993	0.99	9

Nota.La figura muestra las métricas de evaluación de cada iteración. Fuente: Elaboración Propia.

Análisis e Interpretación de la Figura 29

Se puede observar en la Figura 29. La determinación de las métricas de evaluación en cada iteración, los valores se han obtenido de cross validation.

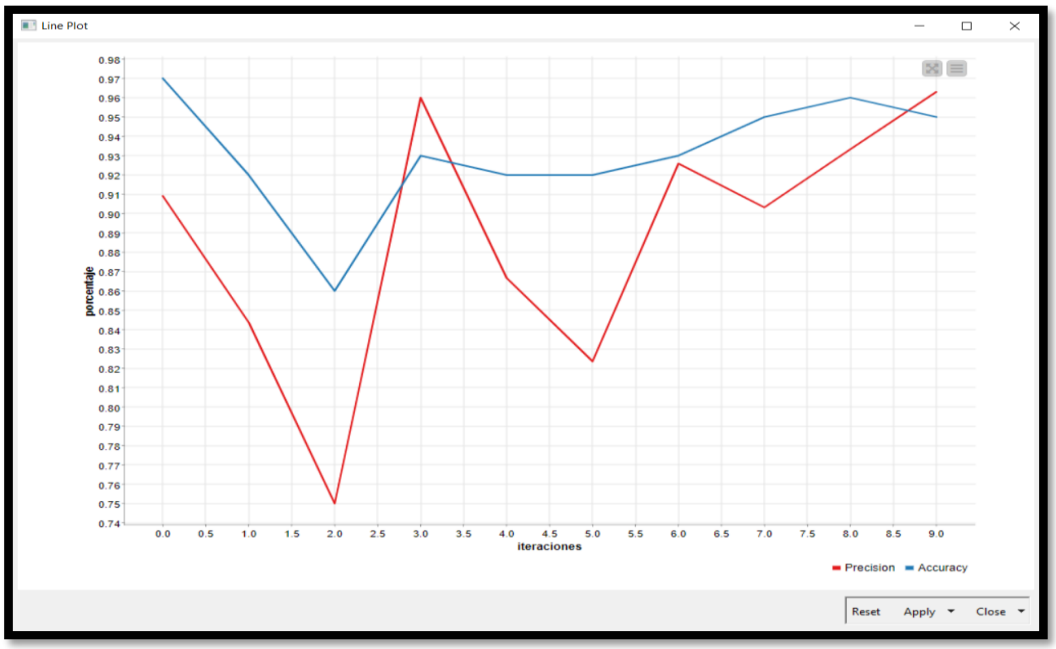
- 92 tweets clasificados correctamente (Correct classified) en laiteración 1.
- 8 tweets clasificados incorrectamente (Wrong classified) en laiteración 1.
- 95 tweets clasificados correctamente (Correct classified) en laiteración 2.
- tweets clasificados incorrectamente (Wrong classified) en iteración 2.
- 94 tweets clasificados correctamente (Correct classified) en la iteración 3.
- tweets clasificados incorrectamente (Wrong classified) en la iteración 3.
- 94 tweets clasificados correctamente (Correct classified) en la iteración 4.
- tweets clasificados incorrectamente (Wrong classified) en la iteración 4.
- 95 tweets clasificados correctamente (Correct classified) en iteración 5.
- tweets clasificados incorrectamente (Wrong classified) en la iteración 5.
- 88 tweets clasificados correctamente (Correct classified) en iteración 6.
- 12 tweets clasificados incorrectamente (Wrong classified) en la iteración 6.

- 94 tweets clasificados correctamente (Correct classified) en la iteración 7.
- tweets clasificados incorrectamente (Wrong classified) en la iteración 7.
- 90 tweets clasificados correctamente (Correct classified) en iteración 8.
- 10 tweets clasificados incorrectamente (Wrong classified) en iteración 8.
- 92 tweets clasificados correctamente (Correct classified) en iteración 9.
- tweets clasificados incorrectamente (Wrong classified) en la iteración 9.
- 99 tweets clasificados correctamente (Correct classified) en la iteración 10.
- 1 tweets clasificados incorrectamente (Wrong classified) en la iteración 10.

Se observa en la Figura 30. Los valores de exactitud (Accuracy) y Precisión (Precisión), de las 10 iteraciones generados por el método de cross validation.

Figura 30.

Cross validation, exactitud(azul) y precisión(rojo)



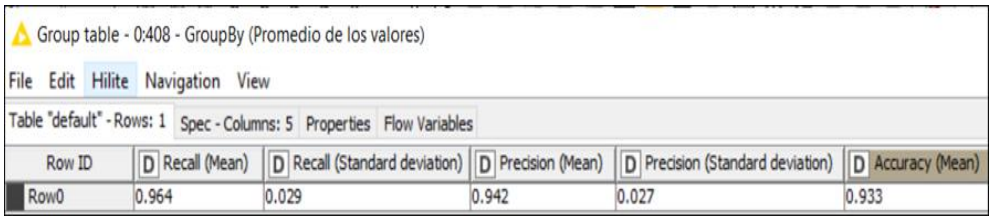
Nota.La figura muestra cross validation la exactitud de color azul y la precisión de color rojo. Fuente: Elaboración Propia

Análisis e Interpretación de la Figura 30

Se observa en la Figura 30. La línea de color rojo es la precisión (Precisión) y la línea de color azul es la exactitud (Accuracy).

Figura 31.

Promedio de valores



Nota.La figura muestra el promedio de valores. Fuente: Elaboración Propia

Análisis e Interpretación de la Figura 31

Se observa en la Figura 31. El promedio de la media aritmética de la precisión (precisión) es 0,942 y el promedio de la media aritmética de la exactitud (Accuracy) es 0,933. Estos valores se obtuvieron de cross validation generadas en cada una de las iteraciones, en este caso son 10 iteraciones.

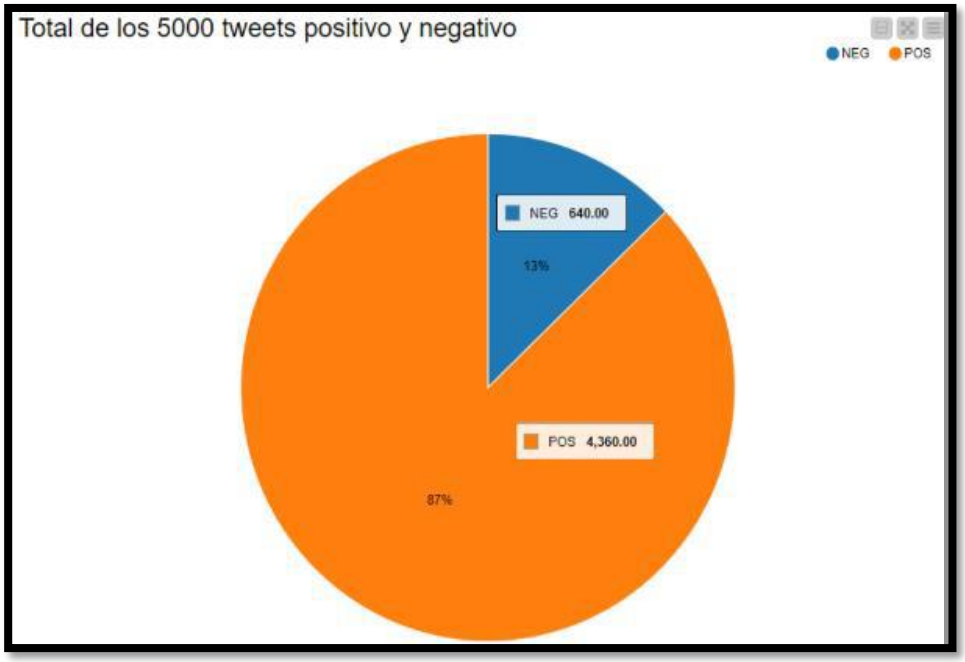
Análisis de los resultados

Luego de haber entrenado el modelo, se hizo la aplicación del modelo entrenado a los 5000 tweets.

Se realizó el proceso de recolección de datos, preprocesamiento, transformación, luego se realizó la predicción utilizando el modelo entrenado y los 5000 tweets.

Figura 32.

Aplicación de 5000 tweets



Nota. La figura muestra la aplicación de 5000 tweets.

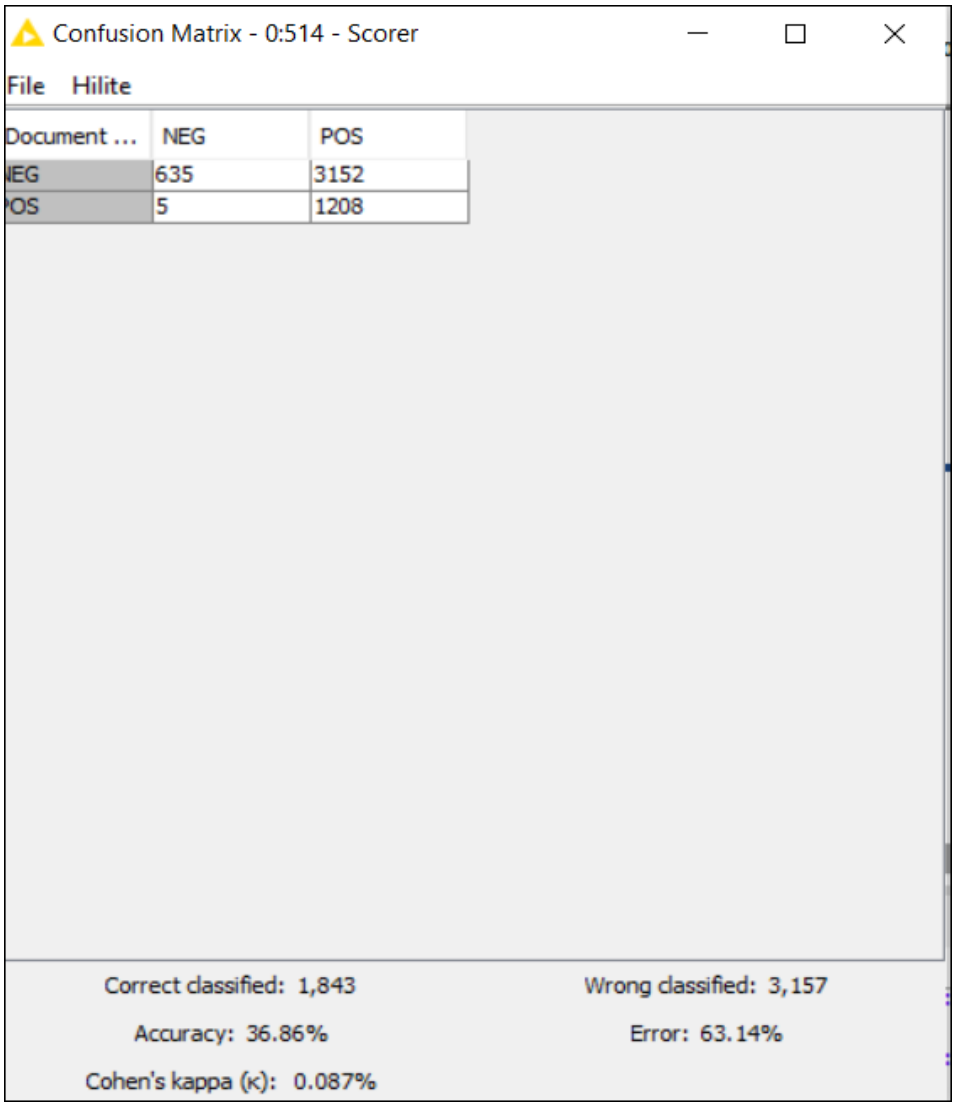
Fuente: Elaboración Propia.

Análisis e Interpretación de la Figura 32

Se observa que el clasificador identificó del total de los tweets, el 87 % (4360) representan a los tweets positivos, mientras que al 13 % (640) representan a los tweets negativos.

Figura 33.

Matriz de confusión de la Aplicación de 5000 tweets



Nota. La figura muestra la Matriz de confusión de la Aplicación de 5000 tweets. Fuente: Elaboración Propia

Análisis e Interpretación de la Figura 33

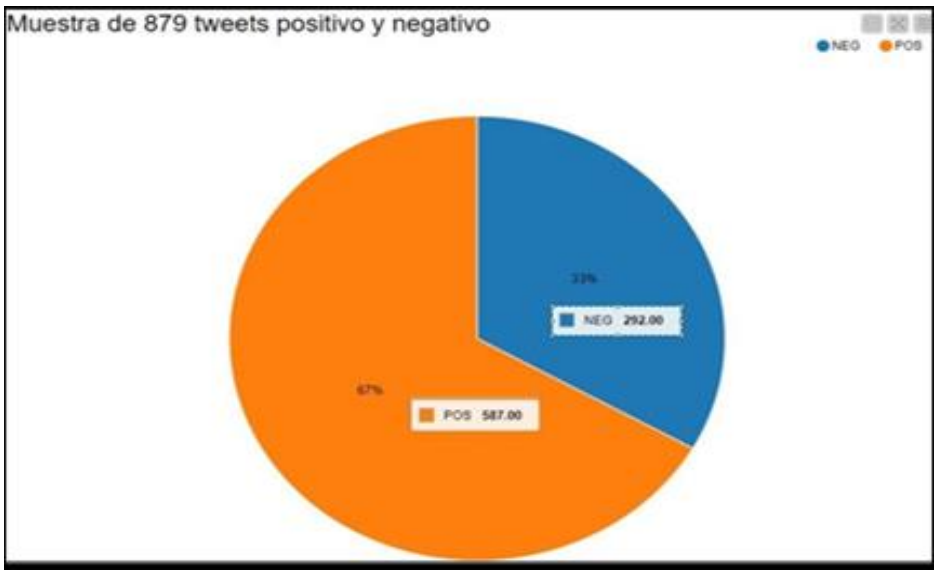
Se puede observar en la Figura 33. Una matriz de la aplicación de los 5000 tweets. Los tweets correctamente clasificados (Correct classified) y los tweets incorrectamente clasificados (Wrong classified), la exactitud (Accuracy), los errores (Error) y la kappa cohens (Cohen's kappa).

- 1843 tweets clasificados correctamente (Correct classified).

- 3,157 tweets clasificados incorrectamente (Wrong classified).
- 36,86 % tweets de exactitud (Accuracy).
- 63,14 % tweets de error (Error).
- 0,087 % tweets de kappa de Cohens (Cohen’s kappa).

Figura 34.

Gráfico de la población de 879 tweets



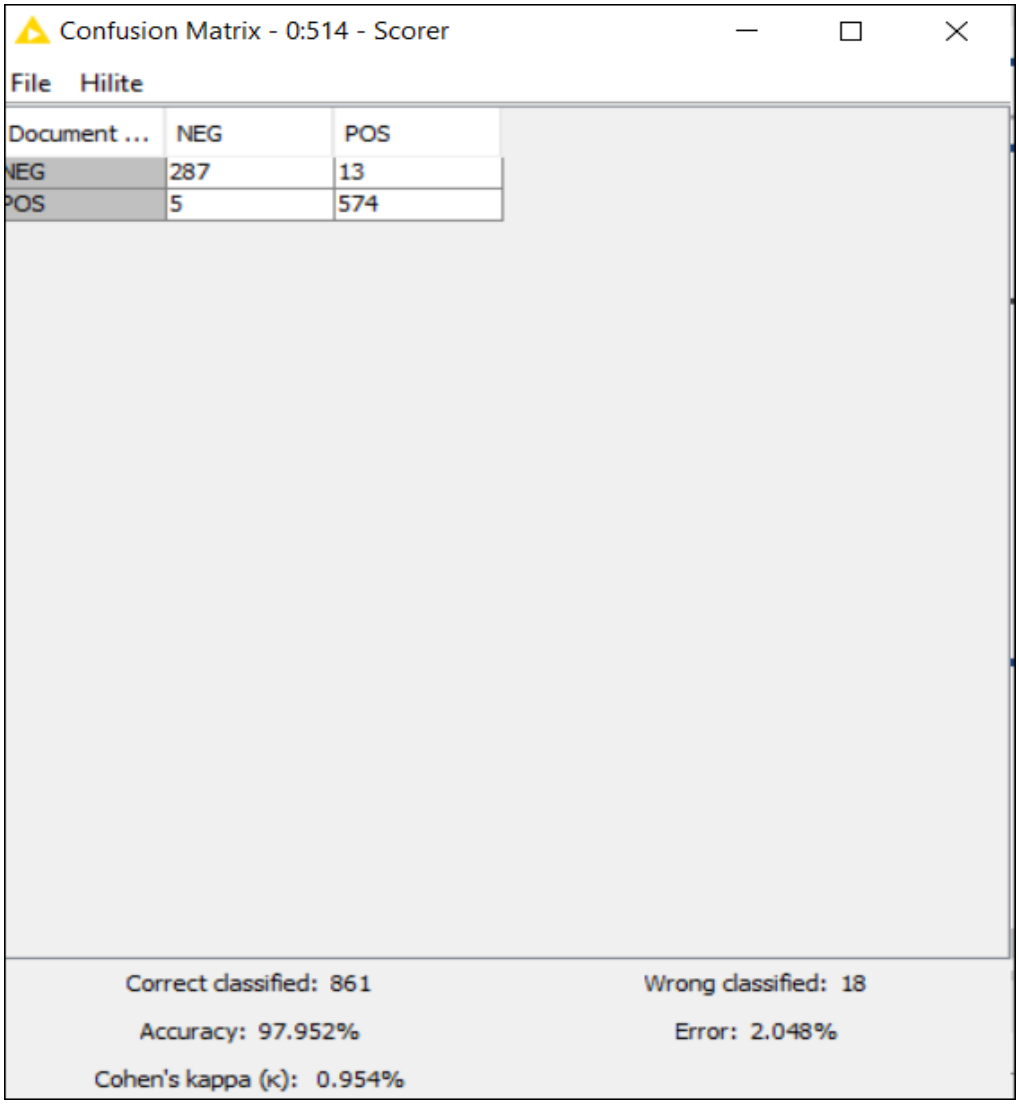
Nota. Gráfico de la población de 879 tweets. Fuente: Elaboración propia

Análisis e Interpretación de la Figura 34

Se observa que el clasificador identificó del total de los tweets, el 67 % (587) representan a los tweets positivos, mientras que al 33 % (292) representan a los tweets negativos.

Figura 35.

Gráfico de Matriz de confusión de 879 tweets de la Muestra



Nota. Muestra un Gráfico de Matriz de confusión de 879 tweets de la Muestra. Fuente: Elaboración propia.

Análisis e Interpretación de la Figura 35

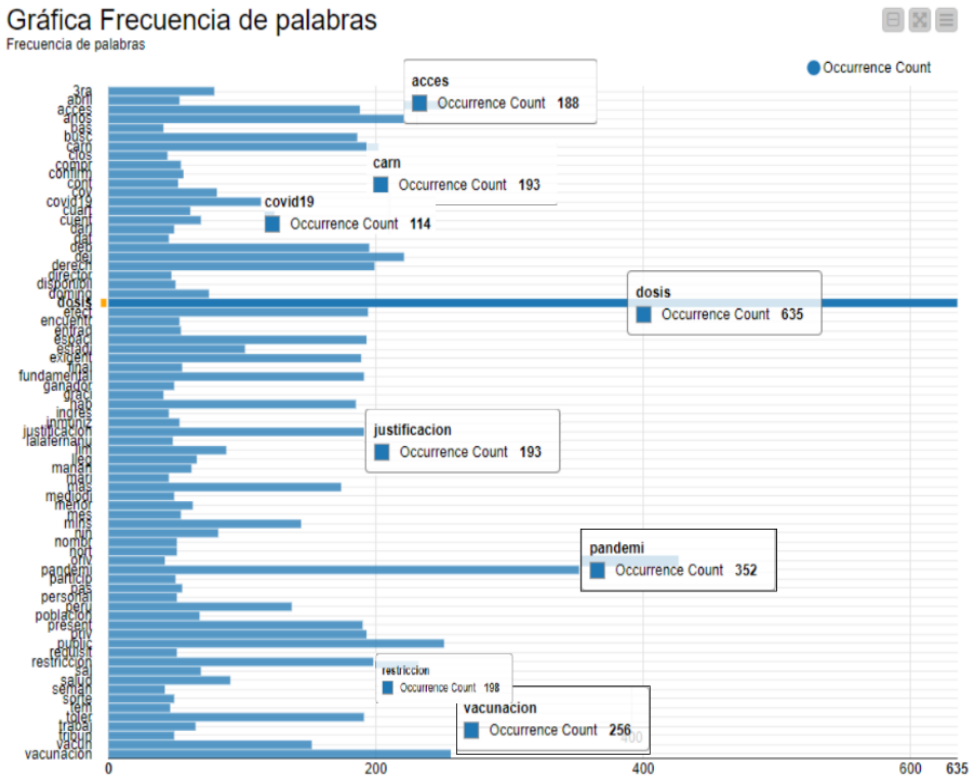
Se puede observar en la Figura 35. Una matriz de confusión, estos valores se obtuvieron. Los tweets correctamente clasificados (Correct classified) y los tweets incorrectamente clasificados (Wrong classified), la exactitud (Accuracy), los errores (Error) y la kappa cohens (Cohen's kappa).

- 861 tweets clasificados correctamente (Correct classified).
- 18 tweets clasificados incorrectamente (Wrong classified).

- 97,952 % tweets de exactitud (Accuracy).
- 2,048 % tweets de error (Error).
- 0,954 % tweets de kappa de Cohens (Cohen's kappa).

Figura 36.

Frecuencia de palabras



Nota.La figura muestra Frecuencia de palabras. Fuente: Elaboración Propia

Análisis e Interpretación de la Figura 36

Se puede observar en la Figura 36 la palabra "pandemia" aparece 352 veces, esta palabra genera un sentimiento de miedo, angustia y enojo. El cual indica que esta palabra pandemia está relacionado al sentimiento de los tweets, también esta palabra se empleó para la búsqueda de los tweets es por esta razón que aparece mayormente.

La segunda palabra es "Dosis" se repite 635 veces esta palabra genera un sentimiento de miedo, miedo a la muerte, fatiga y malestar físico y mental, está asociada al sentimiento de los tweets.

Figura 37.

Nube de palabras

Nota. Nube de palabras. Fuente: Elaboración Propia

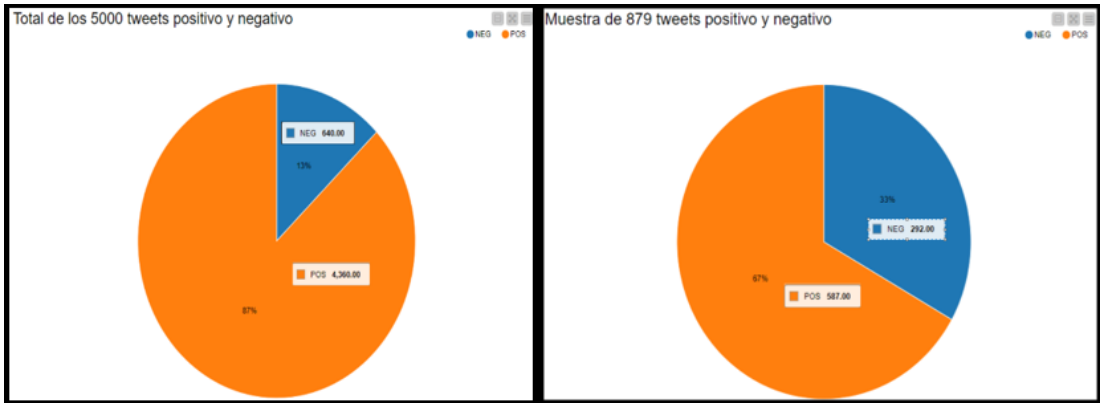
Análisis e Interpretación de la Figura 37

Se observa en la Figura 37 que las palabras más frecuentes son: Pandemia, Dosis, Vacunación, presidente, Efecto, Acceso, Espacio, COVID-19, Vacuna, Crisis, Casa, Mascarilla, Banco, Muerte, Virus, Contagio. Con respecto a las palabras obtenidas anteriormente se obtiene la palabra “Pandemia”. La pandemia de coronavirus COVID-19 es la crisis de salud global que define nuestro tiempo y el mayor desafío que hemos enfrentado. La palabra “Dosis” está asociado a la cantidad de dosis primera, segunda, tercera y cuarta dosis que se aplica a las personas contra la COVID-19. La palabra “vacunación” está asociado a la campaña de vacunación para mayores de 12 años. La palabra “presidente” hace referencia al gobierno que toma decisiones en epidemia de COVID-19. La palabra “Efecto” hace referencia a Los efectos de la COVID-19 en la salud de las personas. La palabra “Acceso” hace referenciaal acceso restringido a los usuarios ya que no recibirán visitas debido a que en estos lugares se encuentran personas vulnerables de la enfermedad de Coronavirus. La palabra “espacio” está asociado a los lugares donde se vacunará contra la COVID-19. La palabra COVID-19 este asociado al aumento de casos positivos de COVID-19. La

palabra “vacuna” está asociado vacunas seguras y eficaces para poner fin a la pandemia de COVID-19. La palabra “Crisis” está asociado la mayor crisis económica y sanitaria que Perú ha enfrentado en los últimos años. La palabra “Casa” este asociado a la COVID-19 es muy contagioso y se extiende por todo el país, la única manera de contenerlo es que todos nos quedemos en casa para protegernos de la enfermedad. La palabra “Mascarilla” está asociado a que son muy útiles para evitar el contagio del virus de la COVID-19. La palabra “Banco” hace referencia a los peruanos deben presentar el certificado de la vacuna de anticovid en los Bancos para evitar contagios de COVID-19. La palabra “muerte” hace referencia a la muerte por COVID-19. La palabra “virus” hace referencia a la enfermedad COVID-19 que infecta a persona de cualquier edad. La palabra “contagio” está asociado a lávese las manos con agua y jabón antes de comer y después de toser para prevenir el contagio del coronavirus.

- Resultados del trabajo de estadística inferencial

Comparación de la Aplicación de 5000 tweets y la muestra de 879 tweets. tal como se puede apreciar en los gráficos.



- Hipótesis Especifica Variable Dependiente

Con relación a la variable dependiente son los resultados clasificar el sentimiento de los tweets relacionados a la pandemia de COVID-19.

Para demostrar el clasificar el sentimiento de los tweets relacionados a la pandemia de COVID-19. de acuerdo a los resultados, se planteó la prueba de hipótesis nula:

a. Hipótesis nula

Ho: El clasificar el sentimiento de los tweets relacionados a la pandemia de COVID-19. Según la población de la aplicación de 5000 tweets y 879 tweets no tienen los mismos resultados.

$$H0: \bar{x}Aplicación \neq \mu Muestra$$

b. Hipótesis alterna

H1: El clasificar el sentimiento de los tweets relacionados a la pandemia de COVID-19. Según la población de la aplicación de 5000 tweets y 879 tweets tienen los mismos resultados.

$$H1: \bar{x}Aplicación = \mu Muestra$$

c. Nivel de Significancia

Para una población de 879 tweets, el estadístico más conveniente es la prueba Z. Para un valor de significancia de α=0,02 o 2 % de error.

$$\propto = 0{,}02 = 2\,\%\; y\; GL = 878$$

d. Se utilizó la distribución t

$$t \propto = t0{,}02 = -540{,}987\; y\; 540{,}987$$

Zona de rechazo y regla de decisión:

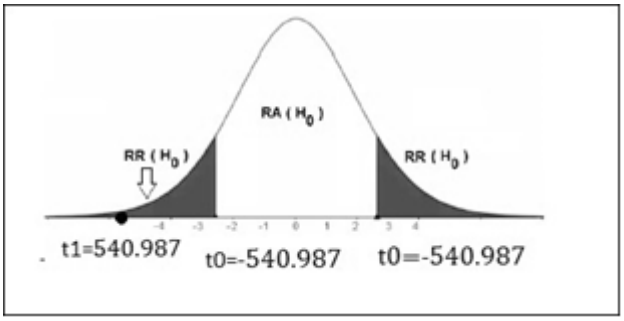
Se usó el estadístico de prueba para n=879; n < 30; Donde n=879 es el número de características de un dataset:

e. Estadística de Prueba

$$t1 = \frac{5000 - 879}{\frac{225{,}844969497}{\sqrt{879}}}$$

$$t1 = 540{,}987.$$

Se ha usado los datos proporcionados por la gráfica en el remplazo de la ecuación y el diagrama con las zonas de rechazo laterales derecha e izquierda, el nivel de significancia y t1:



Luego de realizada la prueba estadística, rechazamos la Hipótesis nula por lo tanto se acepta la Hipótesis alterna, así que clasificar el sentimiento de los tweets relacionados a la pandemia de COVID-19. Tiene las mismas características estándar de los tweets “Aplicación” de análisis de opinión de Twitter, por lo que es adecuado para utilizarla en modelos de clasificación de texto. tal como se puede apreciar en el gráfico.

- Con respecto a la hipótesis específica 02:

Calcular porcentaje de aciertos de la herramienta empleada para la realización de minería de textos.

Se planteo la siguiente hipótesis estadísticas:

a. Hipótesis Nula

H0: Calcular porcentaje de aciertos de la herramienta empleada para la realización de minería de textos. Según la Exactitud (Accuracy) de la Aplicación de los 5000 tweets es menor a la exactitud de (Accuracy) de la muestra de 879 tweets.

$$H0: \bar{x}AplicaciónEx \leq \mu MuestraExa$$

b. Hipótesis Alternativa

H1: Calcular porcentaje de aciertos de la herramienta empleada para la realización de minería de textos. Según la Exactitud (Accuracy) de la Aplicación de los 5000 tweets es mayor a la exactitud de (Accuracy) de la muestra de 879 tweets.

$$H1: \bar{x}AplicaciónEx > \mu MuestraExa$$

c. Nivel de Significancia

Se eligió el nivel de significancia de 0,05 o 5 % de error

$$\propto = 0{,}05 = 5\ \%$$

d. Se utilizó la distribución Z

$$Z \propto = Z\ 0{,}05 = 29{,}5$$

e. Zona de rechazo y regla de decisión

Se usó el estadístico de prueba para n =879; n > 30 según la ecuación (12).

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (12)$$

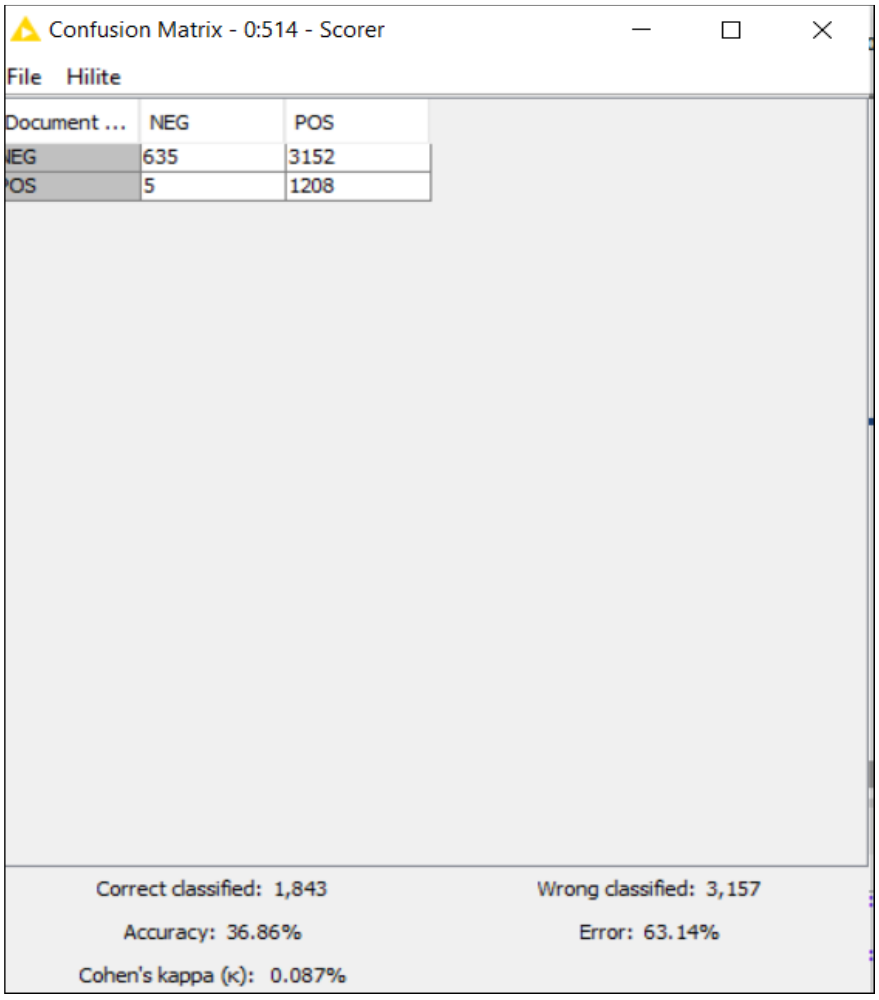
f. Estadística de Prueba

$$Z = \frac{1843 - 861}{\frac{29{,}5}{\sqrt{879}}} = 986{,}99322$$

Se empleó los datos proporcionados por las figuras en el remplazo de la ecuación. tal como se puede apreciar en la figura 38 y figura 39.

Figura 38.

Matriz de confusión de la Aplicación de 5000 tweets

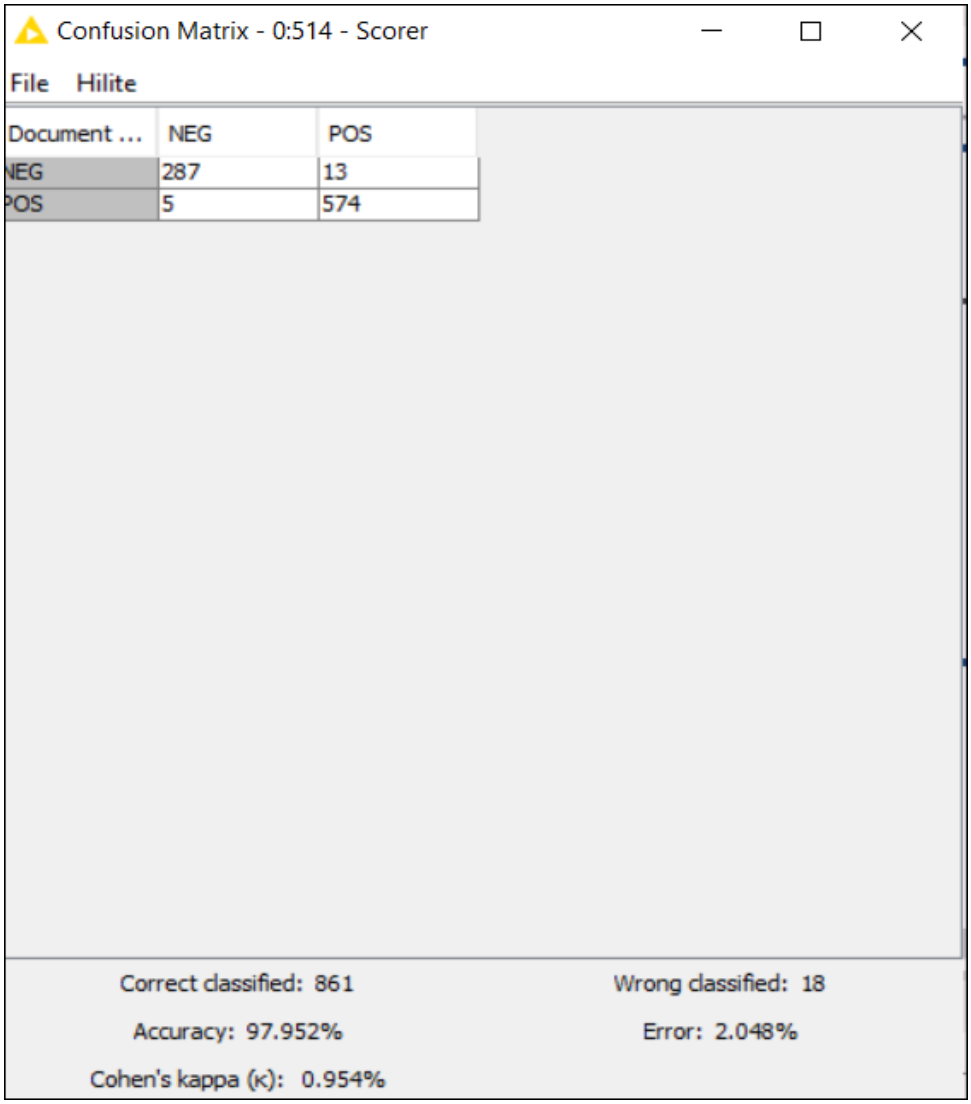


Nota. Matriz de confusión de la Aplicación de 5000 tweets.

Fuente: Elaboración Propia

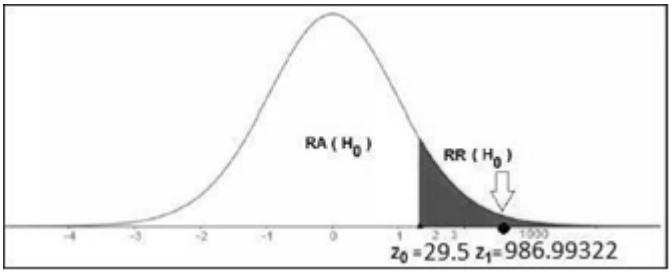
Figura 39.

Gráfico de Matriz de confusión de 879 tweets de la Muestra



Nota. Gráfico de Matriz de confusión de 879 tweets de la Muestra. Fuente: Elaboración propia.

Y el diagrama de zona de rechazo derecha y el nivel de significancia es igual:



Posterior de realizada la prueba estadística, rechazamos la Hipótesis nula por lo tanto se acepta la Hipótesis alterna, así que la exactitud de la Aplicación es menor a la exactitud de la muestra de 879 tweets: por lo tanto, el desempeño es adecuado. tal como se puede apreciar en el gráfico.

CAPÍTULO V: DISCUSIÓN

5.1 Discusión

- Discusión 1

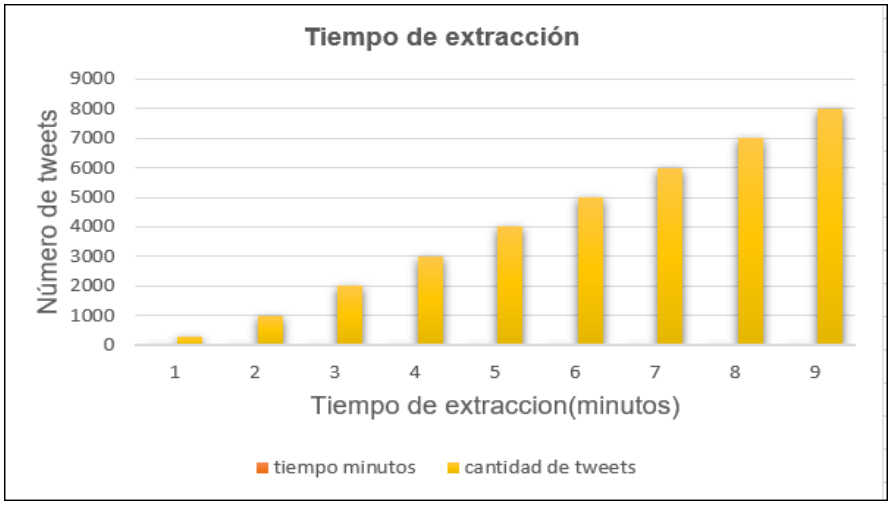
Según Paez y Monroy, en su proyecto denominado “Implementación de un modelo de análisis de sentimientos con respecto a la Jurisdicción Especial para la Paz (JEP) basado en minería de datos en Twitter”, su trabajo de investigación midió el desempeño de los algoritmos Random Forest, Máquinas de Vectores de Soporte y Naïve Bayes. Y como resultado se obtuvo que el algoritmo Random Forest tiene una mayor predicción. La determinación de las métricas es de 74,56 % de precisión, la exactitud es 70,15 %, recall es 70,15 % y con F1-Score de 68,10 %.

Además, se concluye que el tiempo estimado para la extracción de los tweets, la mejor herramienta es “R” según su proyecto.

Conforme a la investigación realizada, luego de haber realizado la extracción de los tweets en la plataforma Knime Analytics, se puede observar los tiempos de extracción en minutos y la cantidad de tweets, esto se puede observar en la Figura 40.

Figura 40.

Tiempos de extracción en minutos



Nota. Tiempos de extracción en minutos. Fuente: Elaboración propia

En efecto, mi investigación y el proyecto de Paez y Monroy, llega a resultados cercanos haciendo uso de algoritmos más idóneos para el contexto requerido, a la vez la extracción de los tweets, se usan herramientas al igual de acordes y se desarrollan para un uso eficiente. En mi caso el tiempo de extracción es óptima.

- Discusión 2

Según Mamani Coaquira, en su artículo denominado “Identificar Sentimientos en Cuarentena por la COVID-19 mediante Clasificador Léxico y Aprendizaje Supervisado”, concluye indicando la necesidad de publicar información relacionado al aislamiento social, estas informaciones deben estar relacionadas con las palabras positivas o negativas conseguidos en esta investigación.

Además, Mamani Coaquira concluye que el clasificador Léxico de palabras y el DLLE3 ayudó al algoritmo de clasificación Máquinas de Vectores de Soporte a mejorar la precisión hasta un 91,5 %, y si se opera con más datos conseguiría mejorar la clasificación. A través de evaluaciones nos da por entender que el algoritmo Máquinas de Vectores de Soporte empleando Python y Knime Analytics tiene mejor clasificación para datos de sentimientos, asimismo al emplear análisis léxico permite comprender mejor el contexto de las frases.

Asimismo, conforme a la presente investigación, después del experimento de clasificador Máquinas de Vectores de Soporte en el software Knime, se obtuvo una exactitud media del 93,3 % y una precisión media del 94,2 %, mientras que en la aplicación del modelo predictivo de los 5000 tweets relacionados a la pandemia COVID-19 se logró el 87 % de tweets positivos y el 13 % negativos.

En efecto, se puede destacar que mi investigación y el artículo de Mamani Coaquira son semejantes ya que su proyecto se realizó acerca del COVID-19 con la distinción de los años. Por otra parte, puedo señalar que mi Exactitud y Precisión son ligeramente más altos ya sea por una extracción de datos más asertiva y por el desarrollo del algoritmo más efectivo.

- Discusión 3

Según Poma Pancaya, en su tesis denominada “Análisis del comportamientode los Tweets, de la participación electoral de los principales candidatos a la presidencia del Perú utilizando minería de textos en el año 2016”.

El candidato Pedro Pablo Kuczynski Godard en la primera vuelta obtuvo 58,2 % tweets positivos y 41,8 % tweets negativos. El 47,3 % tweets positivos y 52,7 % tweets negativos en la segunda vuelta. La candidata Keiko Sofia Fujimori Higuchi en la primera vuelta de la obtuvo 52,4 % tweets positivos y 47,6 % tweets negativos. De igual manera el 47,5 % tweets positivos y 52,5 % tweets negativos en la segunda vuelta.

Para determinar el comportamiento de los dos candidatos presidenciales, desarrollo un clasificador de sentimientos empleando el algoritmo Naïve Bayes.

El candidato Pedro Pablo Kuczynski Godard se obtuvo el 88,2 % de tweets negativos y 85 % de tweets positivos. Para la candidata Keiko Fujimori Higuchi el 86,7 % de tweets negativos y 90,6 % de positivos. De esta forma calculó el porcentaje de aciertos con la herramienta usada de minería de textos.

Además, Poma Pancaya explica la mayor cantidad de tweets se encuentra en días del debate presidencial, a través de la herramienta empleada para minería de datos, y de acuerdo con la investigación realizada, se observó que el día 3 de abril había menos cantidad de tweets, debido a que estamos en el fin de la tercera ola de la pandemia en el Perú, esto se puede apreciar en la Figura 41.

Figura 41.

Ejemplo de cantidad de tweets 3 de abril del 2022

Nota. La figura muestra un ejemplo de cantidad de tweets 3 de abril del 2022.Fuente: Elaboración propia.

En efecto, se puede expresar que el desarrollo de la investigación realizada y del trabajo de tesis de Poma Pancaya, llega a un resultado cercano, aunque el uso del algoritmo difiere, usando Máquinas de Vectores de Soporte en la presente investigación y la contraparte usando el algoritmo Naïve Bayes. Pero debido a las mismas métricas e iteraciones, podemos reconocer las mismas matrices que determinan la veracidad de los resultados.

- Discusión 4

Según Condor, Loa, Huarcaya y Castro, en su Actas del III Congreso Internacional de Ingeniería de Sistemas denominado “Minería de datos en Twitter: Análisis del sentimiento del desempleo en la población hispanohablante en tiempos del COVID- 19”, en su trabajo de investigación como resultado obtuvo seis mil tweets expresados por los usuarios, los términos que más aparecen son “desempleo” 5998 coincidencias, “pobreza” 704 coincidencias, “pandemia” 637 coincidencias, “país” 521 coincidencias, “economía” 468 coincidencias, “trabajo” 408 coincidencias ,esto quiere decir que estos términos se encontraban en los textos de los tweets. Su resultado es el 79,21 % posee sentimiento neutral, el 19,42 % posee sentimiento negativo y 1,37 % posee sentimiento positivo.

Así mismo, de acuerdo con la investigación realizada, luego de haber realizado el preprocesamiento de los tweets, se observó que hubo con mayor frecuencia la palabra “dosis” 635 veces, se puede observar en la Tabla 7.

Tabla 7.

Palabras Frecuentes

Palabras	Cantidad
Acceso	188
Pandemia	352
COVID19	114
Dosis	635
Vacunación	256
Restricción	198

Nota.La tabla muestra las palabras más frecuentes. Fuente: Elaboración Propia

En efecto, se puede expresar que el desarrollo de la investigación realizada y del trabajo de Condor, Loa, Huarcaya y Castro, tienen una semejanza en la obtención de frecuencia de palabras como una métrica notable para determinar la relación con la búsqueda, en esta investigación es acerca de la pandemia como tema principal. Así mismo, se obtiene las palabras relevantes para la recolección de datos y desarrollo de esta investigación.

CONCLUSIONES

Se logró analizar el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos de manera correcta, obteniendo como resultados el 87 % de positivos y el 13 % de tweets negativos.

Se logró elaborar un clasificador de sentimientos de los tweets relacionados a la pandemia de COVID-19 de forma correcta, obteniendo como resultados el 70 % como tweets positivos y 30 % como tweets negativos.

Se logró calcular el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos de manera correcta, obteniendo como resultados, la exactitud (Accuracy) del 36,86 % de la Aplicación de 5000 tweets y la exactitud (Accuracy) del 97,952 % de la muestra de 879 tweets.

Esta investigación tuvo como objetivo calcular el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos. de forma correcta, obteniendo como resultados, con kernel polinomial la exactitud (Accuracy) del 93 %, con kernel hipertangente la exactitud (Accuracy) del 91 % y con el kernel función de base radial la exactitud (Accuracy) del 92 %. Según las evaluaciones del tipo de kernel muestran que el tipo de kernel polinomial tiene mejor rendimiento para la clasificación para tweets de sentimientos. Asimismo, se concluye que el tipo de kernel polinomial ayudo a mejorar al algoritmo de clasificación SVM hasta un 99 % de exactitud (Accuracy). Al analizar el algoritmo de clasificación SVM ha demostrado cómo el tipo de kernel optimiza el rendimiento de SVM en los clasificadores de esta investigaciónes el POLINOMIAL.

Al analizar los diferentes algoritmos de clasificación ha demostrado cómo el tipo de tipo de algoritmos para clasificación puede influir en el rendimiento del algoritmo. Asímismo se concluye utilizando el algoritmo SVM, obteniendo como resultado 95 % de exactitud (Accuracy), para el algoritmo Decision tree obteniendo como resultado 97 % de exactitud (Accuracy) y para el algoritmo Naive Bayes obteniendo como resultado 30 % de exactitud (Accuracy).

Con los resultados obtenidos de los tweets positivos y negativos, el

modelo predictivoes útil para investigadores y profesionales que tienen objetivos de contribuir en futuras emergencias para que tomen medidas de prevención respecto al COVID-19.

En definitiva, la presente investigación ayuda a contribuir a la toma de decisiones en el aspecto socioeconómico, sector público y en nuestra comunidad en general para el desarrollo del campo investigativo del Perú, ya que concluye con resultados que puede contribuir para determinar medidas con relación al COVID-19. Se desea que se interpreten estos resultados para el desarrollo de investigaciones de diferentes áreas como en el sector de Salud, Educación, Información, etc.

RECOMENDACIONES

Se recomienda particionar el conjunto de tweets que se utilizara depende del algoritmo, porque cada uno de estos algoritmos tienen distinto funcionamiento. Analizar el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos, con el objetivo de obtener mejores resultados el 87 % de positivos y el 13 % de tweets negativos.

Se recomienda realizar la comparación de los algoritmos y que se puedan implementar con metodologías porque hay algoritmos que no predicen correctamente y tienen bajo rendimiento con respecto a la exactitud (Accuracy). También realizar un clasificador de sentimientos de los tweets relacionados a la pandemia de COVID-19 prediciendo de forma correctamente clasificados como resultados el 70 % como tweets positivos y 30 % como tweets negativos.

Implementar una partición con Cross Validation de k iteraciones y un particionamiento de training y test para identificar como cambian los resultados. Se logró calcular el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos de manera correcta, obteniendo como resultados, la exactitud (Accuracy) del 36,86 % de la Aplicación de 5000 tweets y la exactitud (Accuracy) del 97,952 % de la muestra de 879 tweets.

Se recomienda para la clasificación del algoritmo SVM el tipo de kernel Polinomial porque ayudó a mejorar la clasificación de tweets de sentimientos un 99 % de exactitud (Accuracy). calculó el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos. de forma correcta, obteniendo como resultados, con kernel polinomial la exactitud (Accuracy) del 93 %, con kernel hipertangente la exactitud (Accuracy) del 91 % y con el kernel función de base radial la exactitud (Accuracy) del 92 %.

Implementar los diferentes algoritmos de clasificación para luego realizar una comparación del rendimiento del algoritmo. Así mismo se concluye utilizando el algoritmo SVM, Decision tree y Naive Bayes, obteniendo como resultado 95 %, 97 % y 30 % de exactitud (Accuracy).

Estudiar los resultados obtenidos usando la curva ROC.

Adaptar el modelo de clasificación para trabajar con otras redes sociales.

Incorporar los emoticones obtenidos de la consulta “Pandemia” en los tweets.

REFERENCIAS BIBLIOGRÁFICAS

Apaza Delgado, S. H. (2016). *Modelo computacional de Minería de Microblogs para el Análisis del Comportamiento del Consumidor de Telefonía Celular.*

Barrios Arce, J. I. (2019). *Health Big Data*. Obtenido de https://www.juanbarrios.com/la-matriz-de- confusión-y-sus-metricas/.

Baviera, T. (2016). *Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength*.

Chanchi G., G. E., Campo M., W. Y., & Sierra M., L. M. (2019). *Estudio del atributo satisfacción en pruebas de usabilidad, mediante técnicas de análisis de sentimientos*.

del Valle Benavides, A. R. (2017). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones*.

Gavilanes, J., Rio, R. M., Cilleruelo, E., & Garechana, G. (2011). *Aplicación de la minería de textos. Análisis de patentes*.

Godoy Viera, A. F. (2017). *Técnicas de aprendizaje de máquina utilizadas para la minería de texto*.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*.

Henríquez, C., Pla, F., Hurtado, L. F., & Guzmán, J. (2017). *Análisis de sentimientos a nivel de aspecto usando ontologías y aprendizaje automático*.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a Social Network or a News Media*?.

Molina López, J. M., & García Herrero, J. (2006). *Técnicas de Análisis de Datos*.

Osorio Zuluaga, G. A. (2009). *Análisis de características del ambiente creativo en empresas de Manizales con técnicas KDD*.

Poma Pancaya, A. J. (2020). *Análisis del Comportamiento de los Tweets, de la Participación Electoral de los Principales Candidatos a la Presidencia del Perú Utilizando Minería de Textos en el año 2016*.

Rissoan, R. (2019). *Redes Sociales Comprender y dominar las nuevas herramientas de comunicación*.

Senso, J. A., & Eíto Brun, R. (2004). *Minería textual*.

Torres Ovalle, S., Domínguez Lugo, A. J., Campos Posada, R., Campos Posada, G., Arzola Garza, O., Vázquez de los Santos, L., & Valdez Menchaca, A. (2011). *Análisis de Contenidos en Áreas Prioritarias*.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques*.

Zhai, C., & Massung, S. (2016). *Text Data Management and Analysis A Practical Introduction to Information Retrieval and Text Mining*.

Amaya, J. (2010). *Toma de desiciones gerenciales*.

Angulo Cuentas, G., Charris Polo, M., Arregocés Rodríguez, Y. F., & Cantillo Bolaño, C. (2009). *Desarrollo e Implementación de una Aplicación Basada en Minería de Textos para la Extracción y Formateo de Información de la Base de Datos de la USPTO que Permita Conocer Tendencias Tecnológicas*.

Arribas López, I. (2008). *Data warehouse de soporte a datos de GSA*.

Bosch, R. (2013). *Análisis de los sentimientos: Aprendizaje incremental para construir modelos de dominio*. Obtenido de https://dokumen.tips/documents/sentiment-analysis-incremental-learning-to-build-domain-sentiment-analysis.html?page=1.

Botta Ferret, E., & Cabrera Gato, J. E. (2007). *Minería de textos: una herramienta útil para mejorar la gestión del bibliotecario en el entorno digital*.

cambria, e., Grassi, M., hussain, a., & havasi, c. (2012). *Sentic Computing para marketing en redes sociales*.

Condor Tinoco, E. E., Loa Navarro, E., Huarcaya Ccoicca, J. A., & Castro Buleje, C. Y. (2020). *Minería de datos en Twitter: análisis del sentimiento del desempleo en la población hispanohablante en tiempos del COVID-19*.

Contreras Barrera, M. (2014). *Minería de texto: una vision actual Text mining: a current view*.

Dubiau, L., & Ale, J. M. (2013). *Analisis de sentimientos sobre un corpus en español*.

Flores González, R., Contreras M., P., & Andrade del Cid, P. (2020). *Comportamiento de las comunidades digitales en Twitter durante las elecciones México 2018*.

Garcés Chaparro, T. I. (2019). *Análisis de sentimientos en redes sociales orientado a la percepción de la calidad de servicios de internet, redes móviles, tv cable y electricidad*.

Gil Pascual, J. A. (2021). *Minería de texto con R.: Aplicaciones y técnicas estadísticas de apoyo*.

Giraldo Londoño, M. A. (2020). *Diseño de una arquitectura de clasificación biclase basada en las máquinas de vectores de soporte, MapReduce y Spark*.

Go, A., Bhayani, R., & Huang, L. (2013). *Twitter Sentiment Classification using Distant Supervision*.

Gutiérrez Meléndez, P. (2012). *Metodología de uso de Herramientas de Inteligencia de Negocios como Estrategia para Aumentar la Productividad y Competitividad de una PyME*.

Joyanes Aguilar, L. (2013). *Big Data. Análisis de grandes volúmenes de datos en organizaciones*.

Kadushin, C. (2013). *Comprender las redes sociales Teorías, conceptos y hallazgos*.

Karamibekr, M., & Akbar Ghorbani, A. (2013). *Lexical-Syntactical Patterns for Subjectivity Analysis of Social Issues*.

Luna, M. (2014). Redes sociales.

Madariaga Orozco, C., Abello Llanos, R., & Sierra García, O. (2014). *Redes Sociales infancia, familia y comunidad*.

Mamani Coaquira, Y., J. Ibarra, M., Mamani Vilca, E., Ordoñez Ramos, E., & Aquino Cruz, M. (2021). *Identificar Sentimientos en Cuarentena por la COVID-19 mediante Clasificador Léxico y Aprendizaje Supervisado*.

Martínez Cámara, E., Martín Valdivia, T., & Alfonso Ureña, L. (2011). *Analisis de Sentimientos*.

Mejova, Y. A. (2012). *Sentiment analysis within and across social media streams*.

Paez Guarnizo, E. P., & Monroy, A. F. (2020). *Implementación de un Modelo de Análisis de Sentimientos con respecto a la JEP basado en Minería de Datos en Twitter*.

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*.

Pérez Abelleira, M. A., & Cardoso, C. A. (2010). *Minería de texto para la categorización automática de documentos*.

Pineda Briseño, A., & Chire Saire, J. E. (2020). *Minería de texto para identificar las principales preocupaciones de los usuarios de Twitter durante COVID-19 en la Ciudad de Mexico*.

Quiroz Gil, N. L., & Valencia, C. A. (2012). *Aplicación del proceso de KDD en el contexto debibliomining: El caso Elogim*.

Ramon Saura, J., Reyes Menendez, A., & Palos Sanchez, P. (2018). *Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de sentimiento sobre las ofertas de #BlackFriday*.

Reveles Gómez, L. C. (2021). *Minería de Opinión: Un Análisis en tiempo real de tweets para Zacatecas*.

Sánchez Guevara, O. A. (2014). *Modelo de inteligencia de negocio para la toma de decisiones en la empresa San Roque S.A*.

Tapia Perales, M. R., Ruiz Montalvo, O. N., & Chirinos Mundaca, C. A. (2014). *Modelo de Clasificación de Opiniones Subjetivas en Redes Sociales*.

Torres Samboni, L. A. (2015). *Análisis de Sentimientos sobre el Posconflicto Colombiano Utilizando Herramientas de Mineria de Texto*.

Acevedo Miranda, C., Clorio Rodriguez, R., Zagal Flores, R., & García Mendoza, C.V. (2014). *Arquitectura Web para análisis de sentimientos en Facebook con enfoque semántico*. México. Obtenido de Repositorio institucional UNJBG: http://repositorio.unjbg.edu.pe/handle/UNJBG/4225.

Aguilar, L. J. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones*. Alfaomega Grupo Editor.

Alhumoud, S. O. (2015). *Survey on Arabic Sentiment Analysis in Twitter*.

Andrade del Cid, P. F. (2020). *Comportamiento de las comunidades digitales en Twitter durante las elecciones México 2018*. Revista de Comunicación.

Barrera, J. H. (2010). *Metodología de la investigación Guía para la comprensión holística de la ciencia*. Bogotá-Caracas: Quirón S. A.

Barrera, M. C. (2014). *Minería de texto: Una Visión Actual*. Biblioteca Universitaria,129-138.

Bosch, R. (2013). *Sentiment Analysis: Incremental learning to build domain models*.

Bravo Marquez, F., Mendoza, M., & Poblete, B. (2013). *Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis*.

Casal, J., & Mateu, E. (2003). *Tipos de Muestreo*. Barcelona.

Chaparro, T. I. (2019). *Análisis de sentimientos en redes sociales orientado a la percepción de la calidad de servicios de internet, redes móviles, tv cable y electricidad*. Santiago de Chile.

Diario Oficial el Peruano. (8 de 3 de 2022). Normas Legales. Lima: Perú. Obtenido de https://busquedas.elperuano.pe/download/url/modifican-

documento-normativo- denominado-disposiciones-para-resolucion-ministerial-n-108- 2022-minedu-2045838-1.

Diario Oficial el Peruano. (2022). Normas Legales. Lima: Editora Perú.

domenech91, J. (8 de 12 de 2011). Validación cruzada de k iteraciones. Obtenido de Validación cruzada de k iteraciones: https://commons.wikimedia.org/wiki/File:K-fold_cross_validation.jpg Editorial.

Etecé. (16 de 7 de 2021). Obtenido de https://concepto.de/redes-sociales/.

Fayyad, U., Piatetsky Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*.

García, V. Á. (1999). *La normalización industrial*. (1. Universitat de València, Ed.)

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*.

Hernández Sampieri, C. R., Fernández Collado, C., & Baptista Lucio, P. (1997). *Metodología de la Investigación*. Colombia: MCGRAW-HILL.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. (2014). *Metodología de la investigación*. México: McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V.

Jiakang Chang (EMBL-EBI), C. O.-E.-E. (febrero de 2018). Obtenido de https://universoabierto.org/2018/02/22/que-es-la-mineria-de-textos-como-funciona-y-por-que-es-util/.

Knime. (2018). *Text Mining Course for Knime*.

Londoño, M. A. (2020). *Diseño de una arquitectura de clasificación biclase basada en las máquinas*. Medellín, Antioquia, Colombia.

Luciana Dubiau, J. M. (2013). *Análisis de sentimientos sobre un, Corpus en Español: Experimentación con un Caso de Estudio*. Argentina.

Makice, K. (2009). *API de Twitter: en funcionamiento: aprenda a crear aplicaciones con la API de Twitter*. (I. O'Reilly Media, Ed.) Estados Unidos: O'Reilly Media, Inc.

MINSA. (2022). Obtenido de https://covid19.minsa.gob.pe/sala_situacional.asp.

MINSA. (2022). Sala Situacional Vacunación COVID-19. Lima.

MINSA. (2022). Vacuna COVID-19 en el Perú. Lima.

Montes y Gómez, M. (2005). *Mineria de texto empleando la semejanza entre estructuras semánticas*. México.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: *Machine Learning in Python*.

Ramirez C., T. A. (1997). *Como hacer un proyecto de investigación*. Caracas: Tulio A. Ramírez C.

R-Moreno, M., Cuesta, A., & F. Barrero, D. (2013). *Twitter Stream Analysis in Spanish*.

Sampieri, R. H. (2014). *Metodologia de la investigacion*. México: S.A. DE C.V.

Suárez, E. J. (2016). *Tutorial sobre Máquinas de vectores de soporte (SVM)*.

Tan, A.-H. (1999). *Text mining: The state of the art and the challenges*. Citeseer.

Universidad Privada de Tacna. (2017). *Manual para la Presentación de Planes e Informes de Investigación*. Tacna.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Witten, I., Don, K., Dewsnip, M., & Tablan, V. (2004). *Text mining in a digital library*. International Journal on Digital Libraries, IV.

Zhai, C., & Massung, S. (2016). *Text data management and analysis: a practical introduction to information retrieval and text mining*: Morgan & Claypool.

ANEXOS

Anexo 1.

Matriz de consistencia

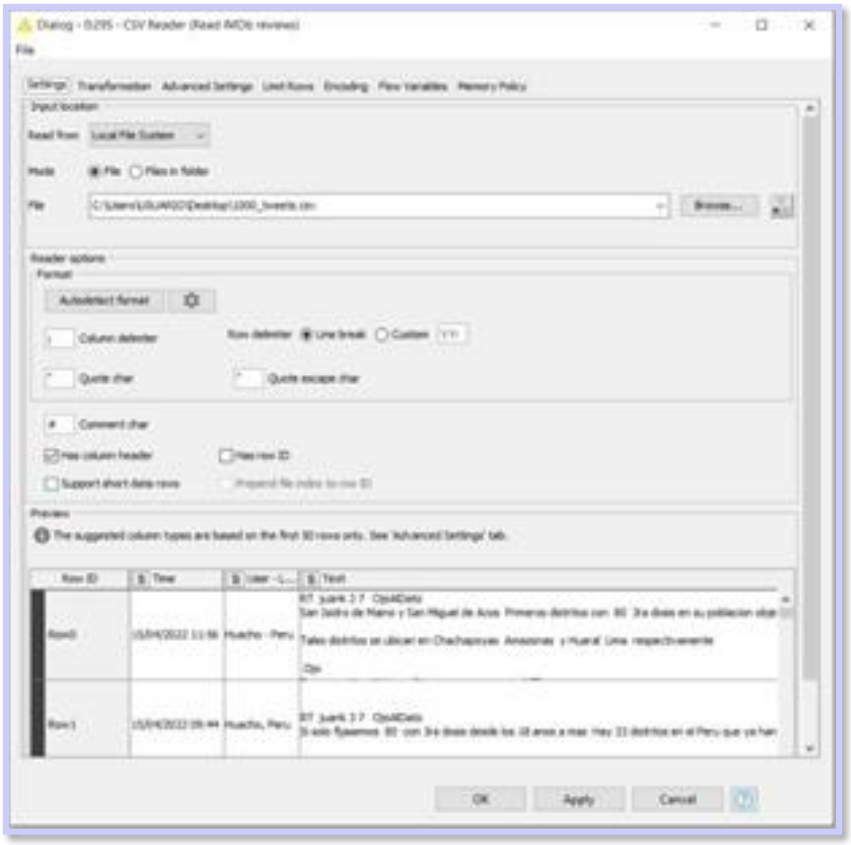
INTERROGANTE DEL PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	INDICADORES	MÉTODOLOGÍA
Interrogante general ¿Cómo es el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos, año 2022?	**Objetivo general** Analizar el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos, año 2022.	**Hipótesis general** Se podrá analizar el comportamiento de los tweets relacionados a la pandemia de COVID-19 en el Perú empleando minería de textos,2022.	**Variable independiente** Minería de textos.	- Número de tweets para ser analizados. - Rendimiento del algoritmo. -Eficiencia de algoritmo.	**Tipo de investigación** Correlacional. **Nivel de investigación:** Comprensivo **Diseño de Investigación:** Documental. **Población:** La población está constituida por el conjunto de tweets publicados por las personas del Perú que interactúan con el Twitter relacionados a la pandemia COVID-19.
Interrogantes específicas 1. ¿Cómo clasificar el comportamiento de los tweets relacionados a la pandemia de COVID-19?	**Objetivos específicos** - Elaborar un clasificador de sentimientos de los tweets relacionados a la pandemia de COVID-19 en el Perú, año 2022.	**Hipótesis específicas** - Se podrá elaborar un clasificador de sentimientos de los tweets relacionados a la pandemia de COVID-19.	**Variable dependiente** Análisis del comportamiento de los tweets.	- Porcentaje de tweets positivos	
2. ¿Cuál es el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos?	- Calcular el porcentaje de aciertos de la herramienta usada para la realización de minería de textos.	- Se podrá calcular el porcentaje de aciertos de la herramienta empleada para la realización de minería de textos.			

3. ¿Cuáles serán los resultados obtenidos por la herramienta empleada para minería de textos?	- Analizar los resultados obtenidos por la herramienta usada para minería de textos.	-		- Porcentaje de tweets negativos.	**Muestra:** Por conveniencia. **Técnicas:** Recopilación de datos. **Instrumentos:** -Plataforma de Knime Analytics. -API de Twitter.

Anexo 2.

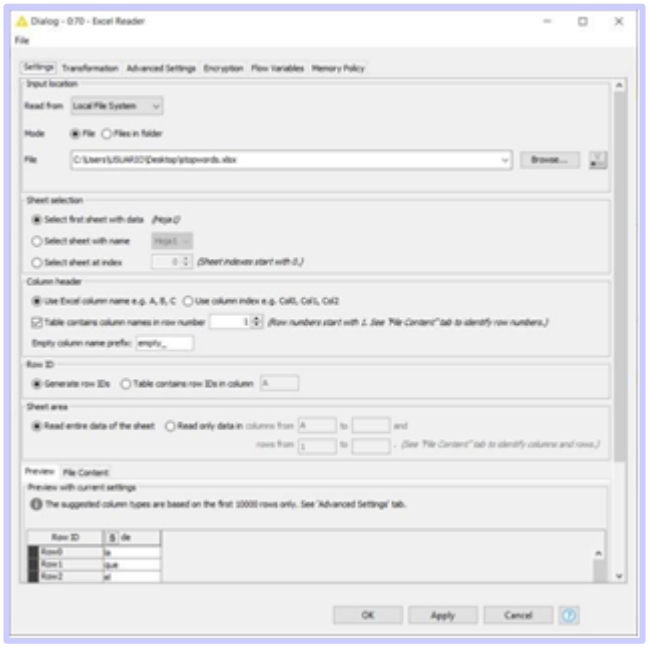
Procedimiento de la configuración de los nodos. Clasificación de documentos: Implementación de modelos

- Configuración: Nodo CSV Reader



Nota. En el Nodo CSV Reader cargamos el archivo de 1000 tweets Aplicamos y ya está listo para visualizar los datos.

- Configuración: Nodo Excel Reader

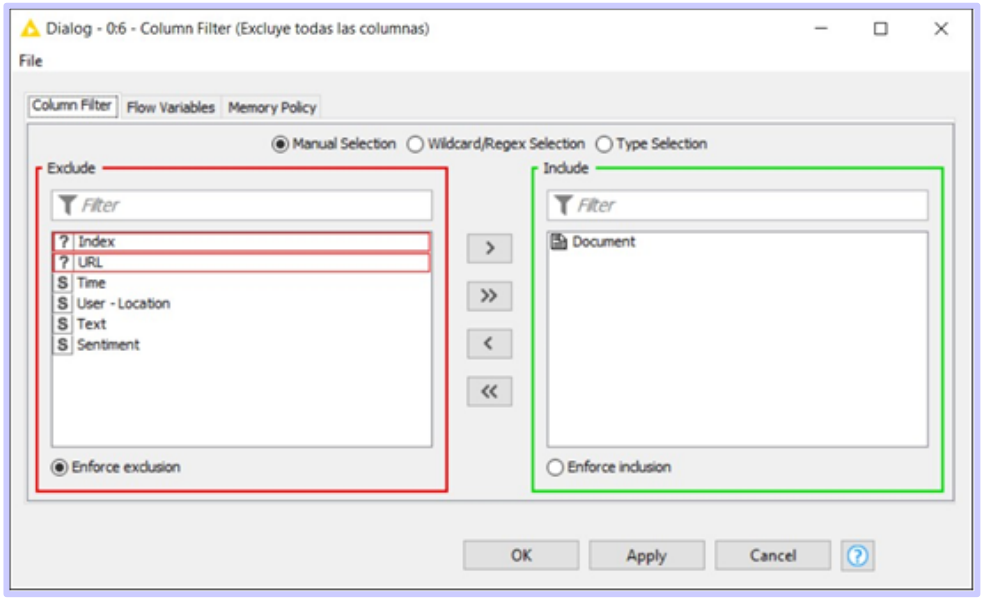


Nota. En el Nodo Excel Reader sirve para leer y cargamos el archivo stopwords.xlsx, estás son palabras como artículos, preposiciones, palabras que no ayudan a la investigación aplicamos y aceptamos.

- Configuración: Nodo String to Document

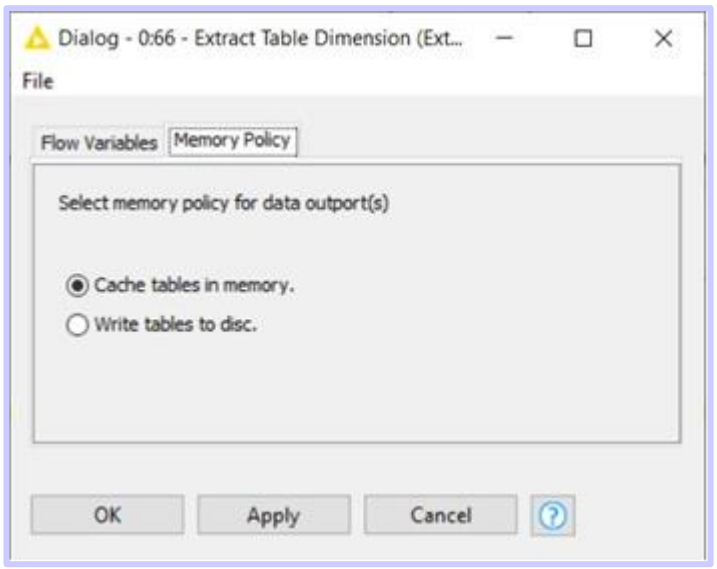
Nota. En el Nodo String to Document (Cadenas para documentar) seleccionamos Text, seleccionamos la categoría sentiment, tokenization seleccionamos en español. aplicamos y aceptamos. Este nodo sirve para convertir cadenas en Documentos. Para cada fila se creará un documento.

- Configuración: Nodo Column Filter



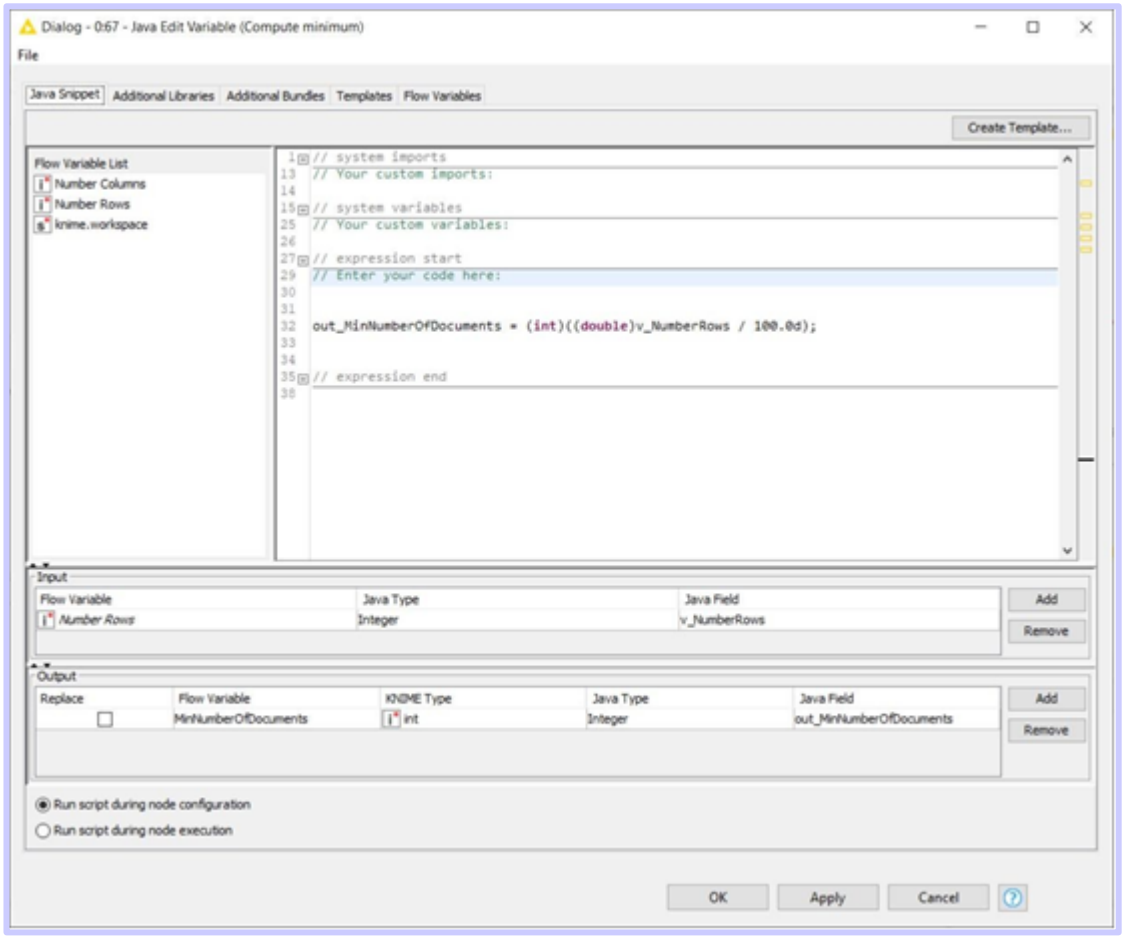
Nota. Este Nodo Column Filter sirve para filtrar columnas, excluye todas las columnas y solo incluye la columna Document aplicamos y aceptamos.

- Configuración: Nodo Extract Table Dimension



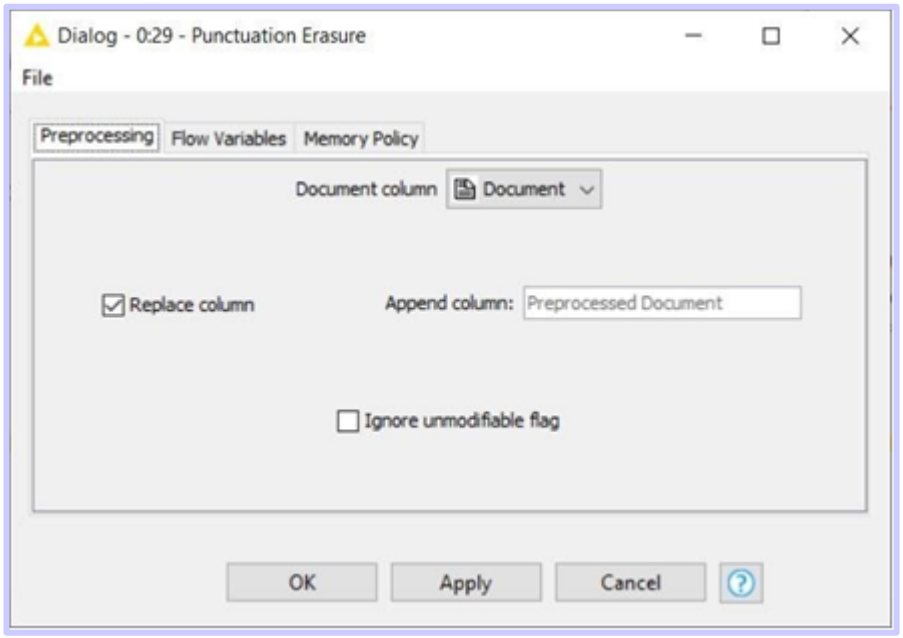
Nota. El Nodo Extract Table Dimension (Extraer dimensión de la tabla) sirve para extraer el número de filas y columnas de una tabla. Aplicamos y aceptamos.

- Configuración: Nodo Java Edit Variable



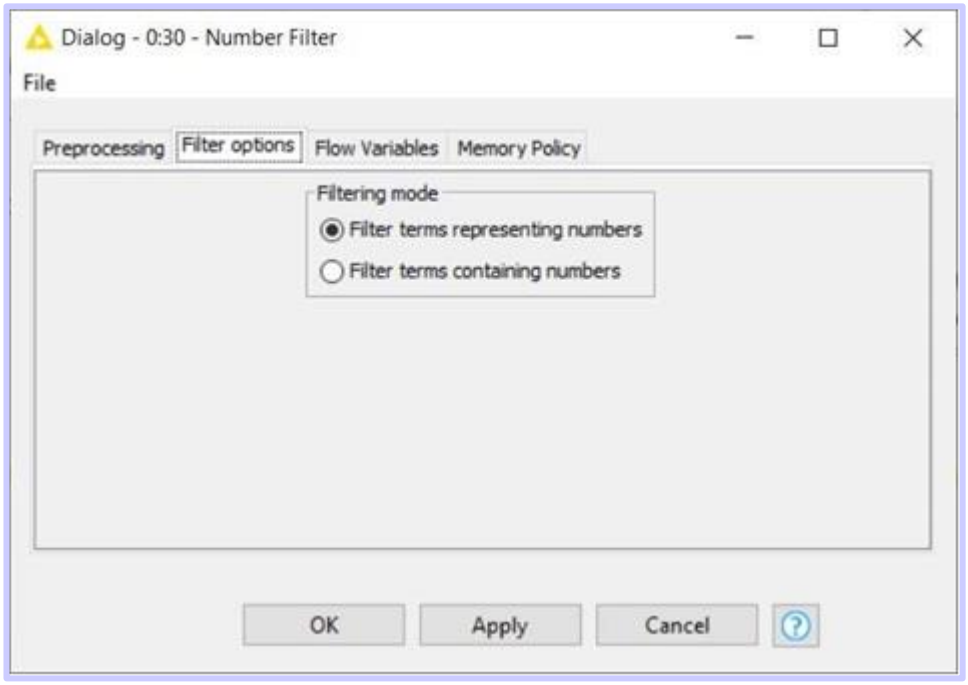
Nota. El Nodo Java Edit Variable (Variable de edición de Java) realiza el cálculo divide número de filas entre 100 y da un resultado. Aplicamos y aceptamos.

- Configuración: Nodo Punctuation Erasure



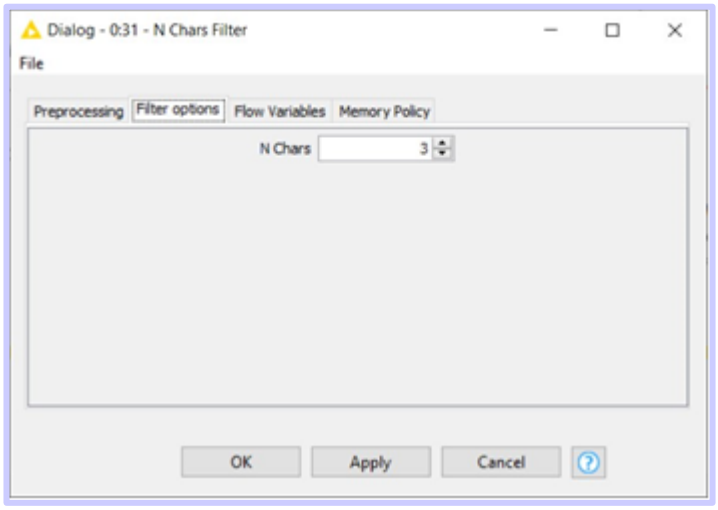
Nota. El Nodo Punctuation Erasure (Borrado de puntuación) elimina todos los caracteres de puntuación de los términos contenidos en los documentos, seleccionamos la columna Documento y reemplazamos la columna con el mismo. Aplicamos y aceptamos.

- Configuración: Nodo Number Filter



Nota. El Nodo Number Filter (Filtro de número) elimina todos los números contenidos en el documento.

- Configuración: Nodo N Chars Filter



Nota. Nodo N Chars Filter (Filtro de caracteres N) elimina palabras que contengan 2 caracteres.

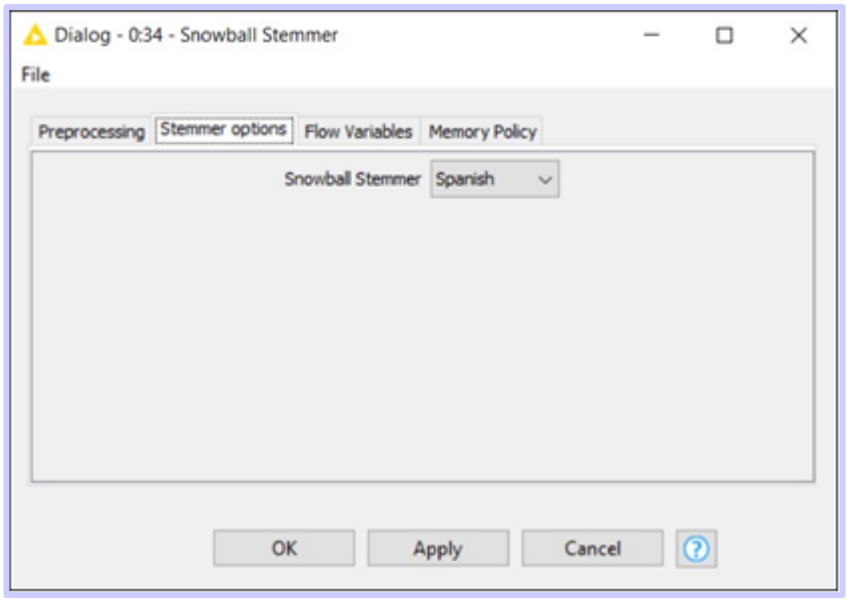
- Configuración: Nodo Stop Word Filter

Nota. El Nodo Stop Word Filter (Filtro de palabras de detención) seleccionamos una lista de palabras vacías integradas en español y hacer check y personalizamos otra lista de palabras vacías que es un archivo stopwords.xlsx. Aplicamos y aceptamos.

- Configuración: Nodo Case Converter

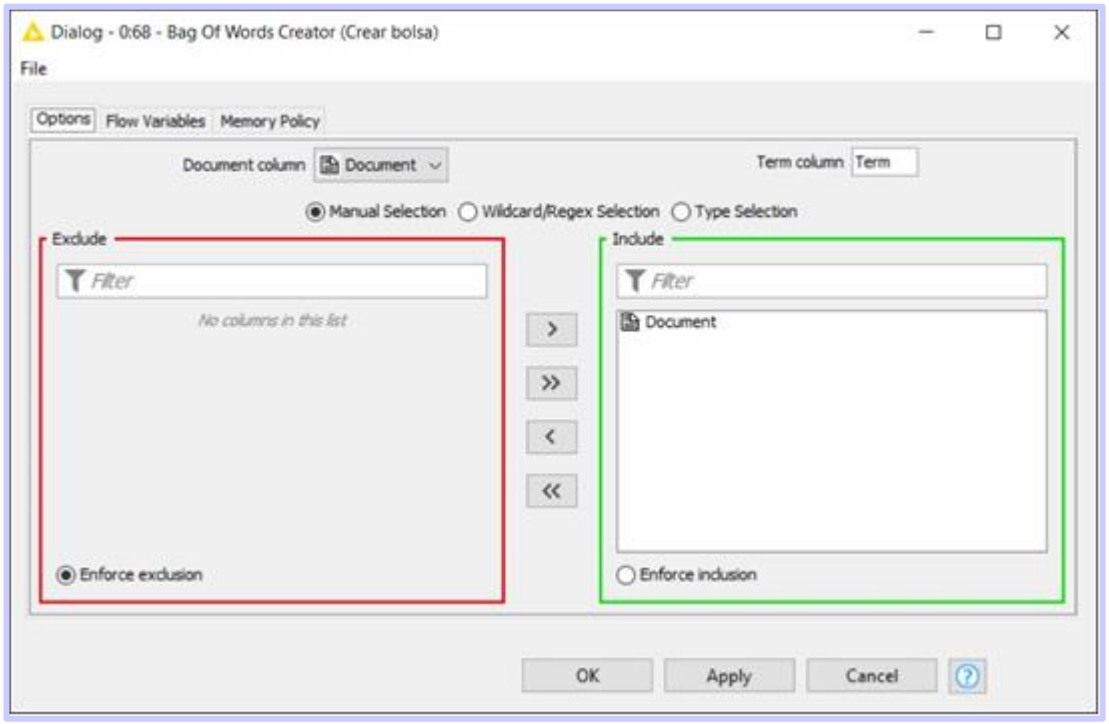
Nota. El Nodo Case Converter (Convertidor de casos) convierte caracteres a minúsculas. Aplicamos y aceptamos.

- Configuración: Nodo Snowball Stemmer



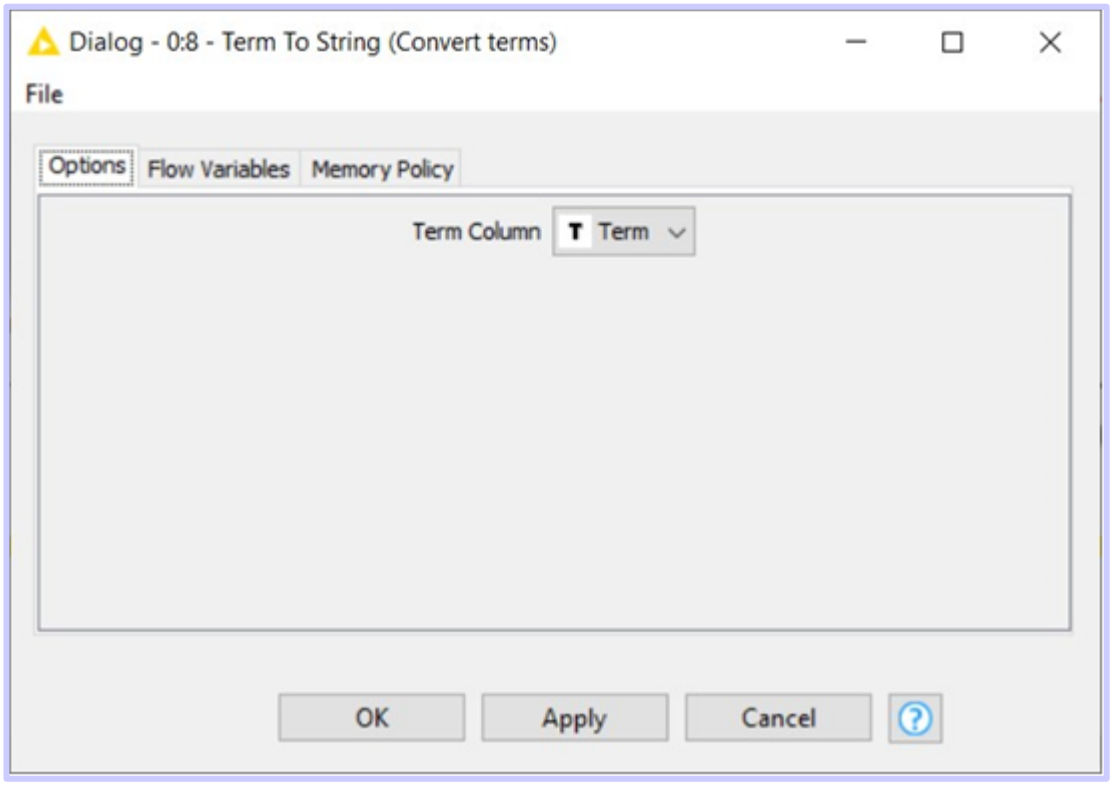
Nota. El Nodo Snowball Stemmer (Talador de bolas de nieve) reduce una palabra a su palabra base o raíz de modo que las palabras similares se encuentren en la raíz común. Seleccionamos el idioma español. Aplicamos y aceptamos.

- Configuración: Nodo Bag Of Words Creator



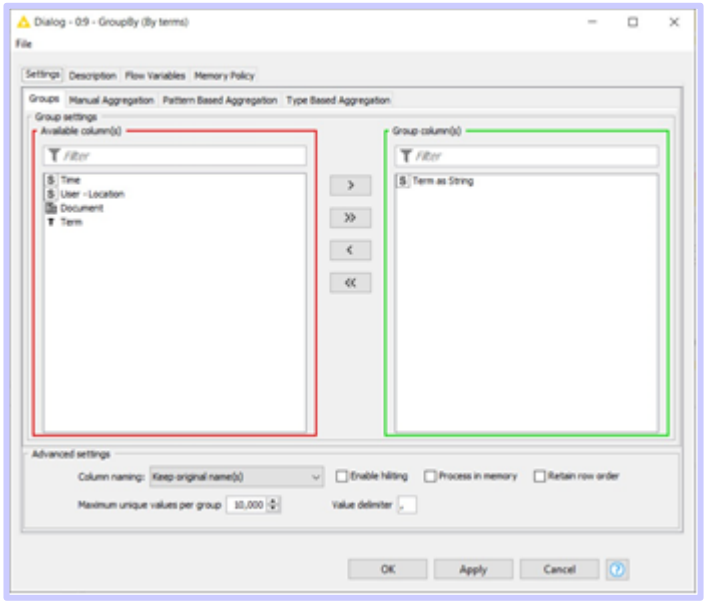
Nota. El Nodo Bag Of Words Creator (Creador de Bolsa de Palabras) crea una bolsa de palabras de un conjunto de documentos. Una Bolsa de palabras consta de una columna que contiene los términos que aparecen en el documento. Seleccionamos el Documento aplicamos y aceptamos.

- Configuración: Nodo Term to String



Nota. El Nodo Term to String (Término a cadena) convierte los términos de una columna en cadenas y adjunta una nueva columna con estas cadenas. se pierden las etiquetas de los términos. Aplicamos y aceptamos.

- Configuración: Nodo GroupBy

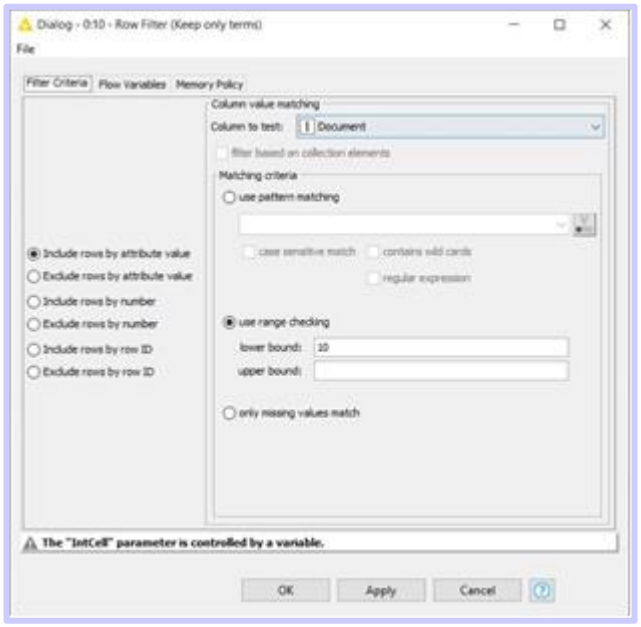


Nota. El Nodo GroupBy (Agrupar por) Agrupa las filas de una tabla por los valores únicos en columnas del grupo seleccionado. Aplicamos y aceptamos.

- Configuración: Nodo Tag Cloud

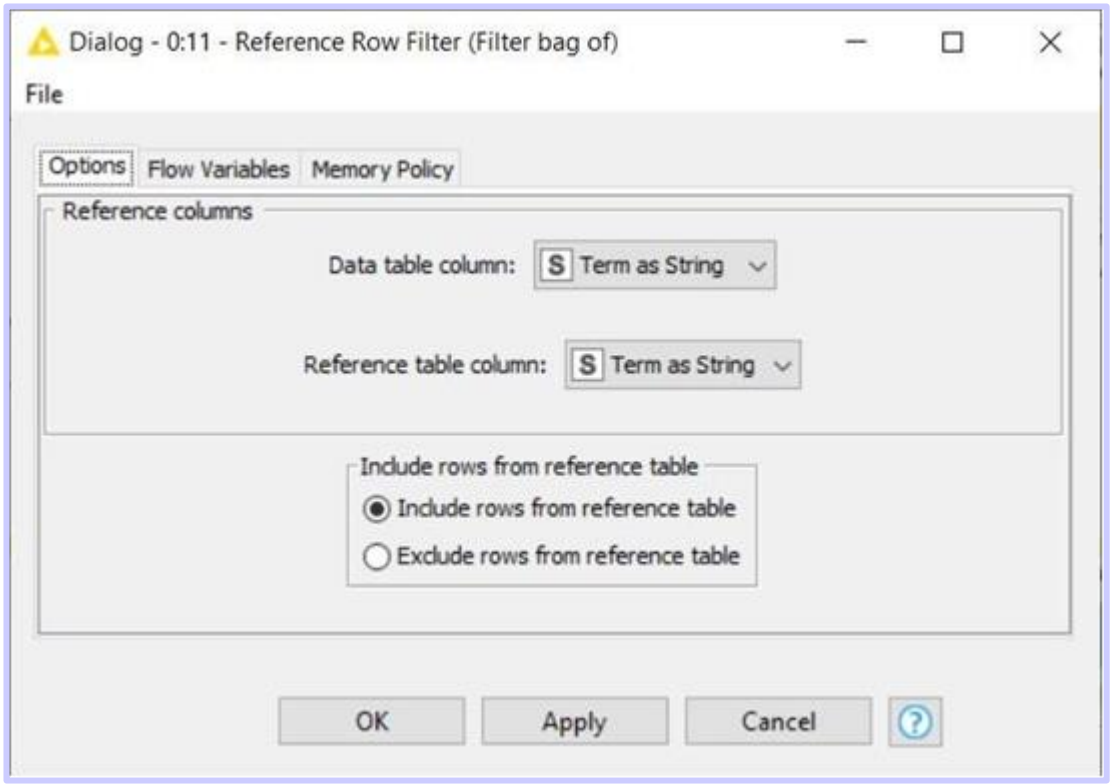
Nota. El Nodo Tag Cloud (Nube de etiquetas) Muestra una vista de nube de etiquetas seleccionamos términos como cadenas y la cantidad de etiqueta es 250. OK.

- Configuración: Nodo Row Filter



Nota. El Nodo Row Filter (Filtro de fila) hace filtrado de filas según ciertos criterios. Mantiene términos que aparezcan en un número mínimo de documentos. Aplicar y aceptar.

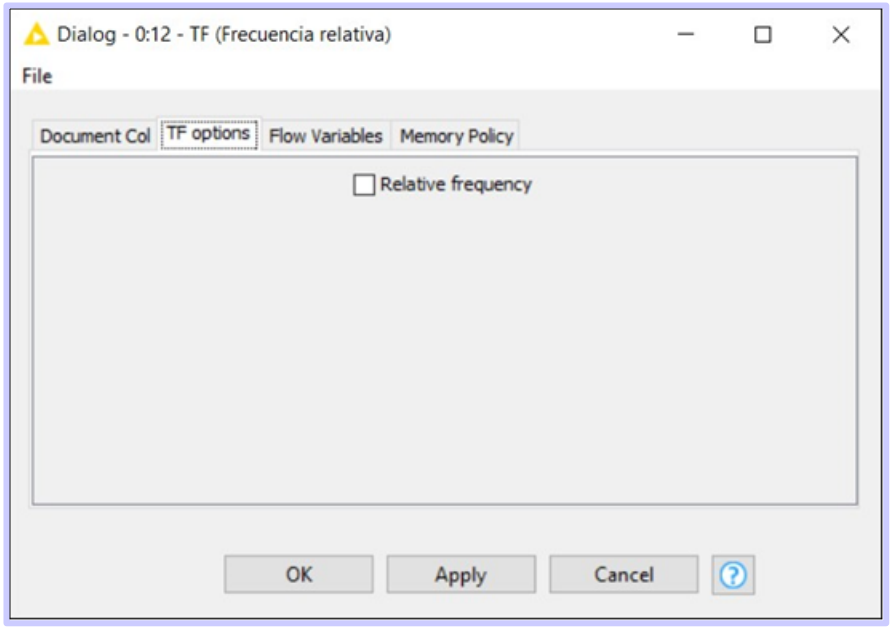
- Configuración: Nodo Reference Row Filter



Nota. El Nodo Reference Row Filter (Filtro de fila de referencia) filtra filas de la

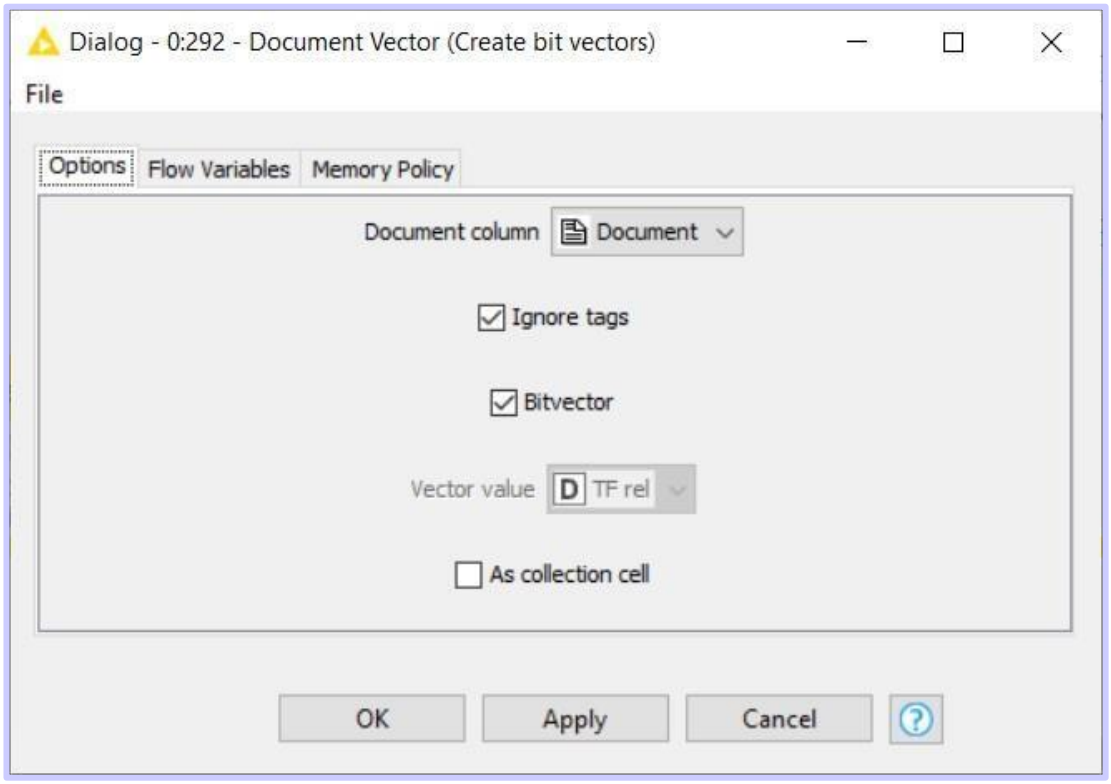
primera tabla utilizando la segunda tabla como referencia, Bolsa de palabras filtrado. Aplicamos y aceptamos.

- Configuración: Nodo TF



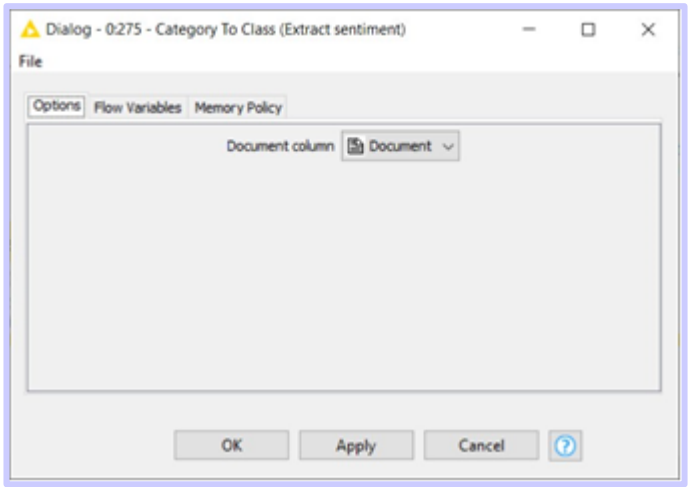
Nota. El Nodo TF Calcula la frecuencia relativa del término (TF) de cada término según cada documento y añade una columna que contiene el valor de TF. aplicamos y aceptamos.

- Configuración: Nodo Document Vector



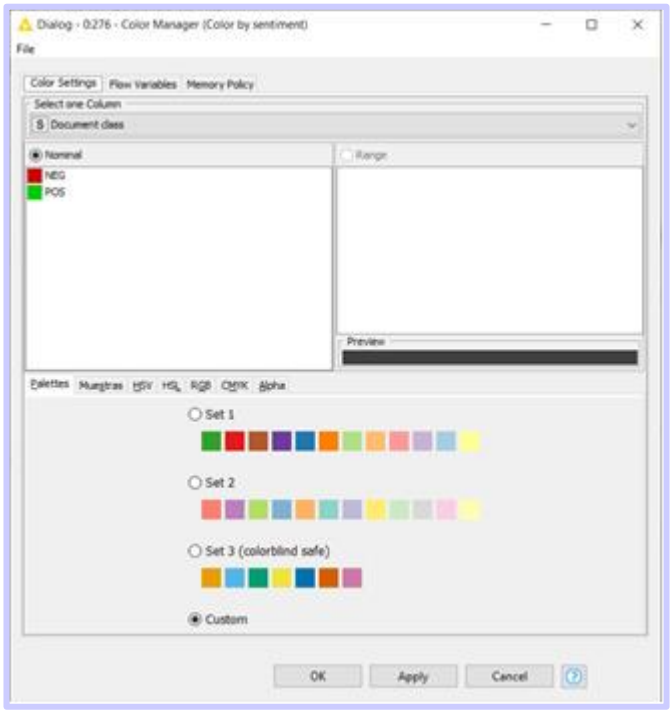
Nota. El Nodo Document Vector (Vector de documento) Crea vectores de bits para documentos. aplicamos y aceptamos.

- Configuración: Nodo Category to Class



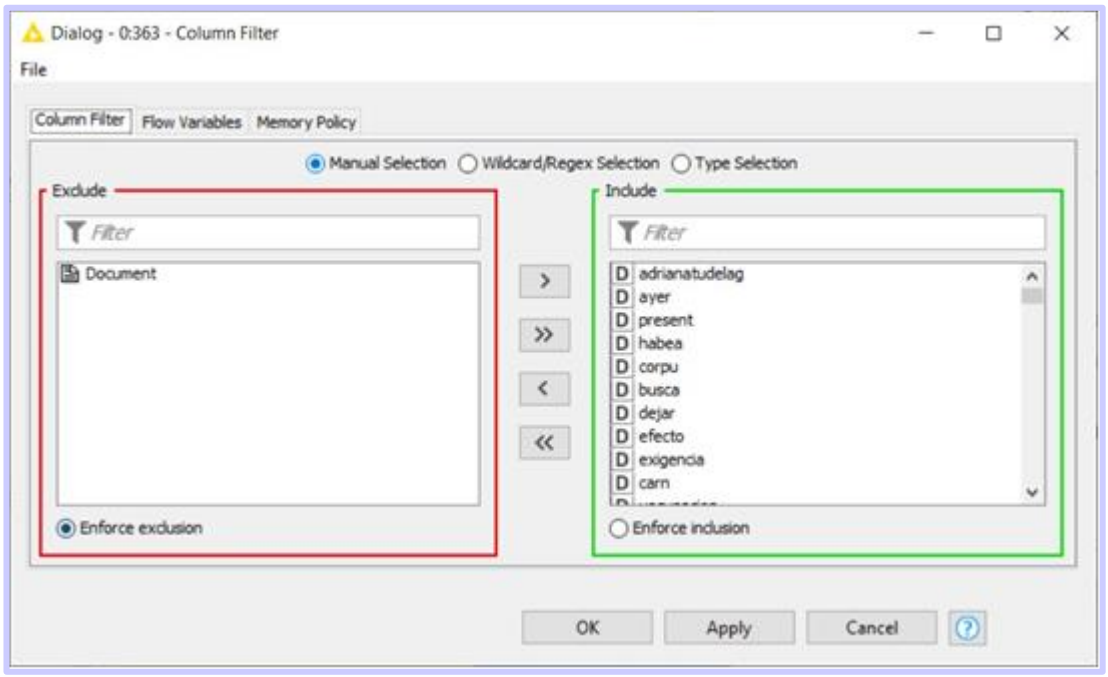
Nota. El Nodo Category to Class (Categoría a clase) Excluye la columna sentiment, sirve para agregar una columna de clase (cadena) a cada fila que contiene una celda de documento. El valor de la clase es la categoría del documento como cadena. Aplicamos y aceptamos.

- Configuración: Nodo Color Manager



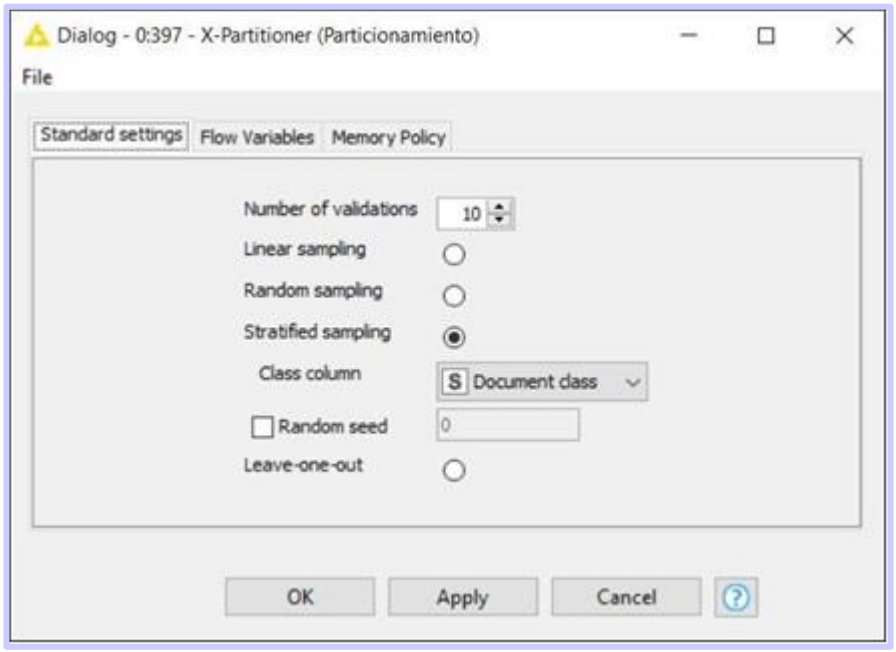
Nota. El Nodo Color Manager (Administrador de color) colorea cada categoría verde para positivo y rojo para negativo se asigna a columnas nominales. Aplicamos y aceptar.

- Configuración: Nodo Column Filter



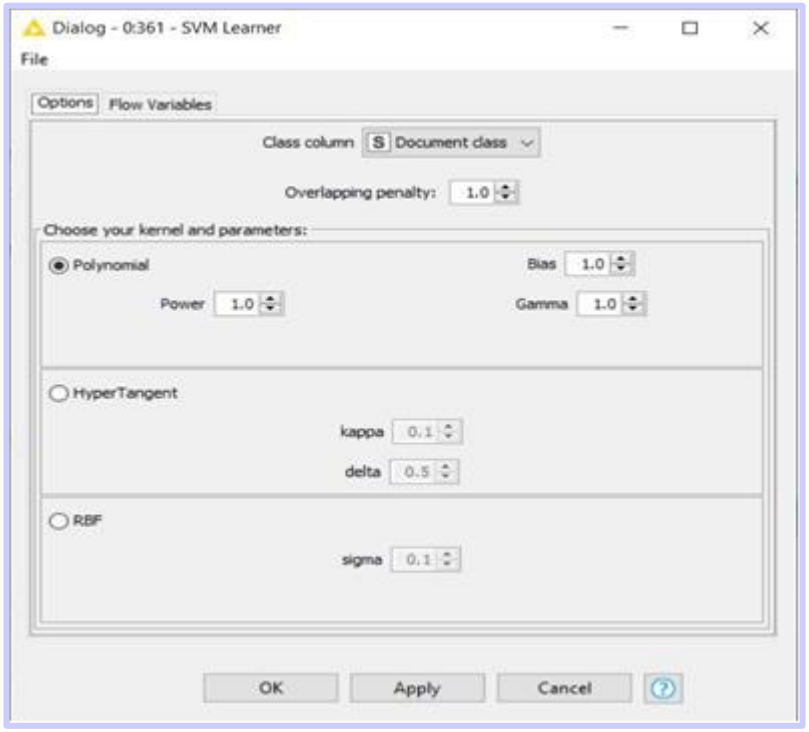
Nota. El Nodo Column Filter (Filtro de columna) filtra las columnas de la tabla mientras que solo las columnas restantes se pasan a la tabla. Aplicamos y aceptamos.

- Configuración: Nodo X-Partitioner



Nota. El Nodo X-Partitioner (X-particionador) es un bucle de validación cruzada. Al final del bucle debe haber un agregador X para recopilar los resultados de cada iteración. En este caso será 10 iteraciones. Aplicamos y aceptamos.

- Configuración: Nodo SVM Learner



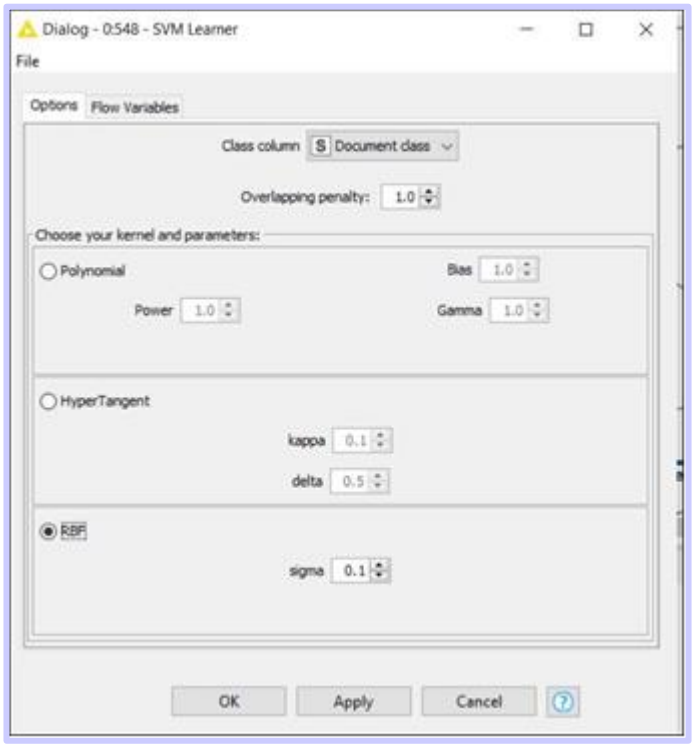
Nota. El Nodo SVM Learner (Aprendiz de SVM) entrena una máquina de vectores de soporte con los datos de entrada. Posee varios núcleos diferentes estos son: HyperTangent, Polynomial y RBF. El alumno de SVM permite problemas de múltiples

clases y calcula el hiperplano entre cada clase. Elegimos el tipo de kernel Polynomial Aplicamos y aceptamos.

- Configuración: Nodo SVM Learner

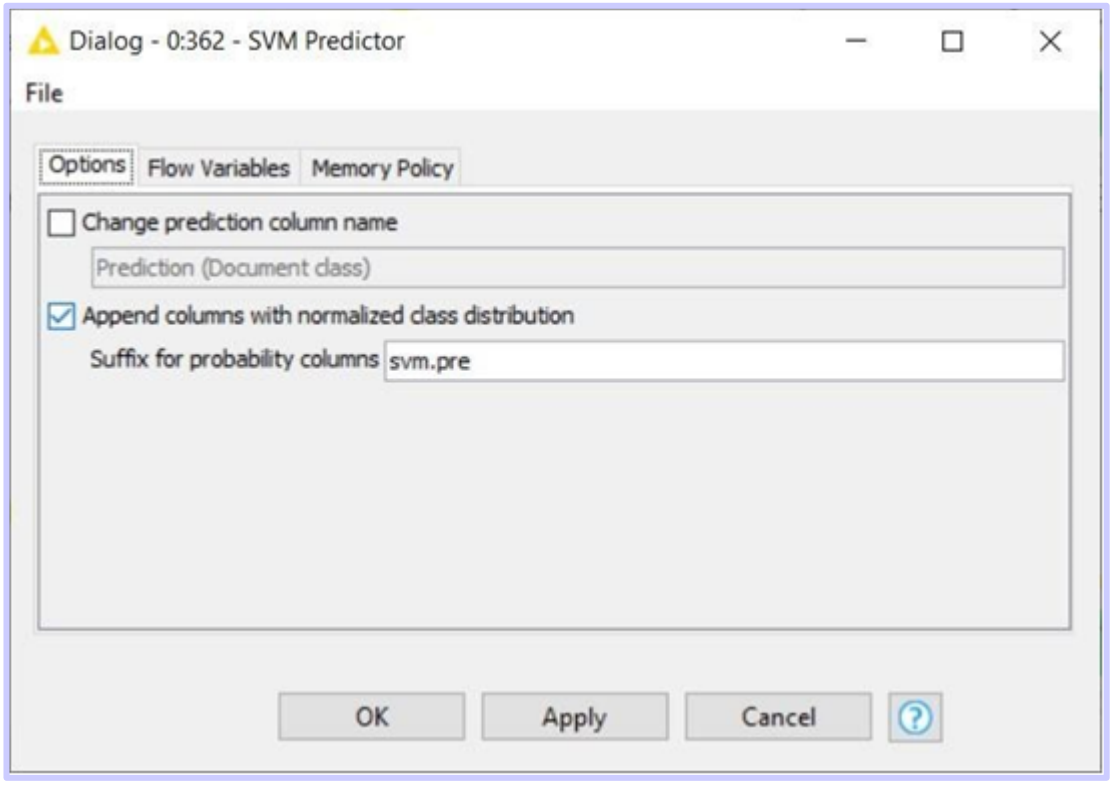
Nota. El Nodo SVM Learner (Aprendiz de SVM) entrena una máquina de vectores de soporte con los datos de entrada. Elegimos el tipo de kernel HyperTangent Aplicamos y aceptamos.

- Configuración: Nodo SVM Learner



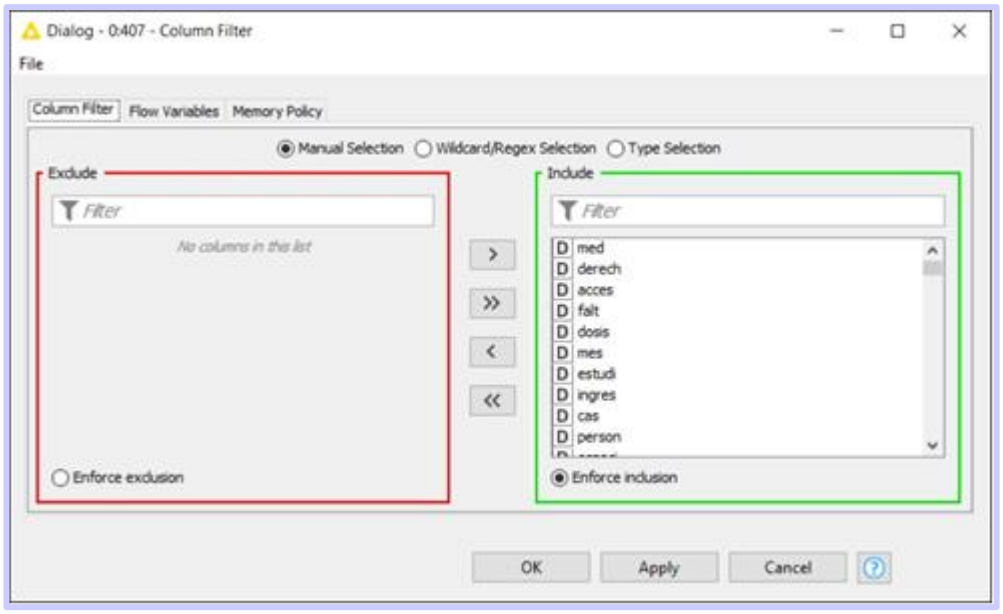
Nota. El Nodo SVM Learner (Aprendiz de SVM) entrena una máquina de vectores de soporte con los datos de entrada. Elegimos el tipo de kernel RBF Aplicamos y aceptamos.

- Configuración: Nodo SVM Predictor



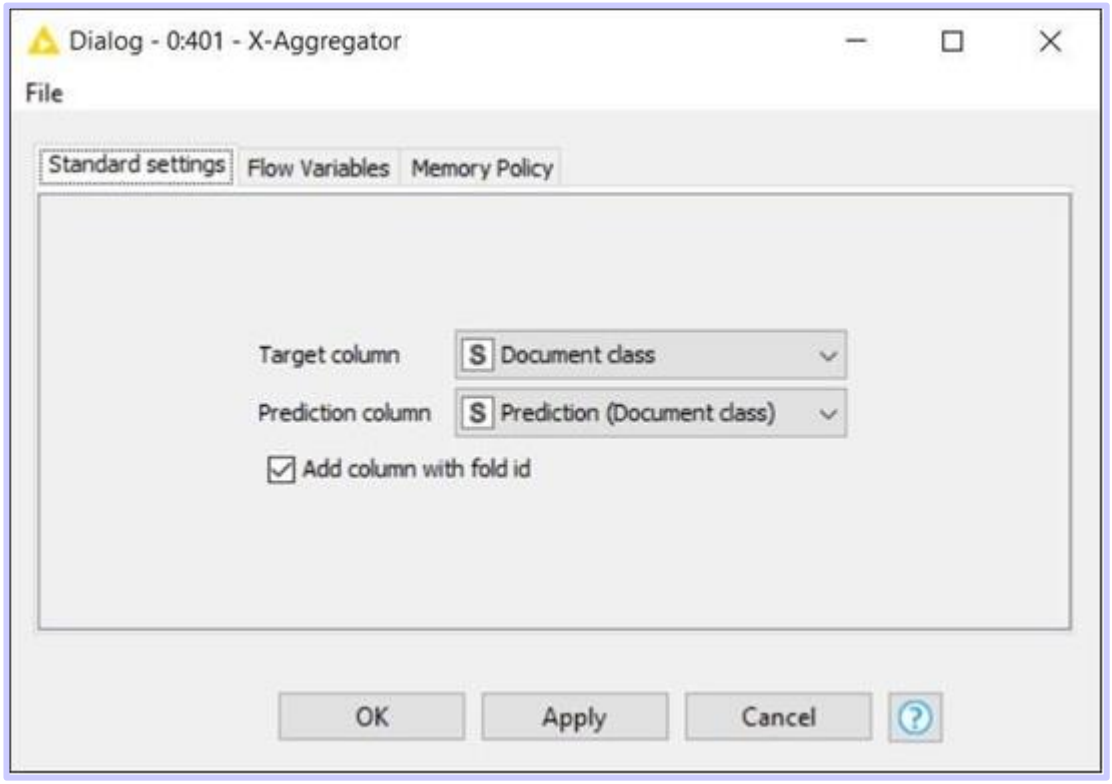
Nota. El Nodo SVM Predictor (Predictor de SVM) emplea un modelo de SVM generado por el nodo de aprendizaje de SVM para predecir el resultado de sentimiento positivo o negativo. Aplicamos y aceptamos.

- Configuración: Nodo Column Filter



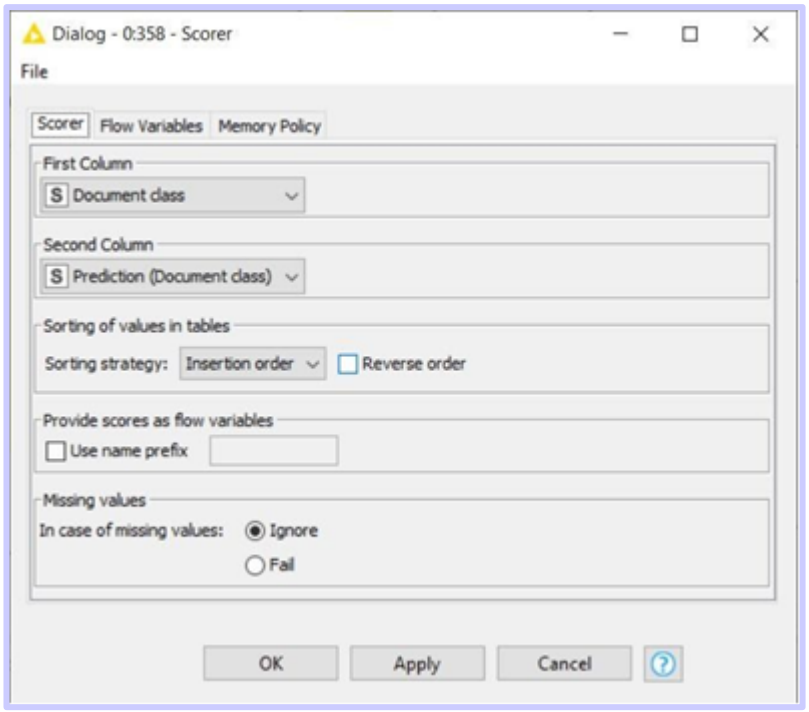
Nota. Nodo Column Filter (Filtro de columna) filtra las columnas de la tabla mientras que solo las columnas restantes se pasan a la tabla. Aplicamos y aceptamos.

- Configuración: Nodo X-Aggregator



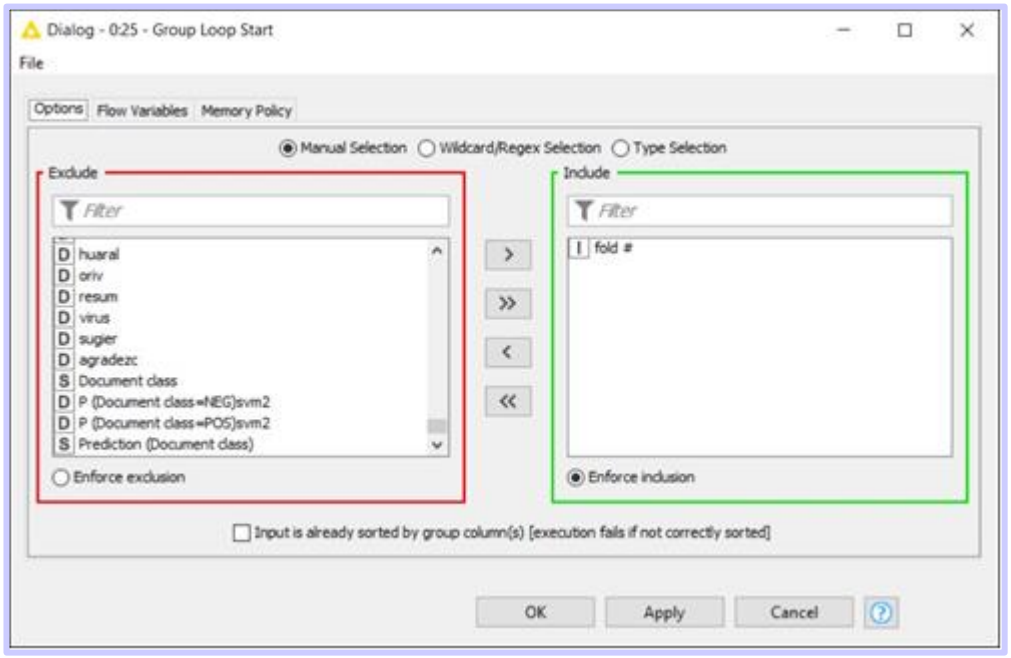
Nota. El Nodo X-Aggregator (Agregador X) Recopila el resultado del nodo predictor, compara la clase predicha y la clase real y genera las predicciones para las filas. Aplicamos y aceptamos.

- Configuración: Nodo Scorer



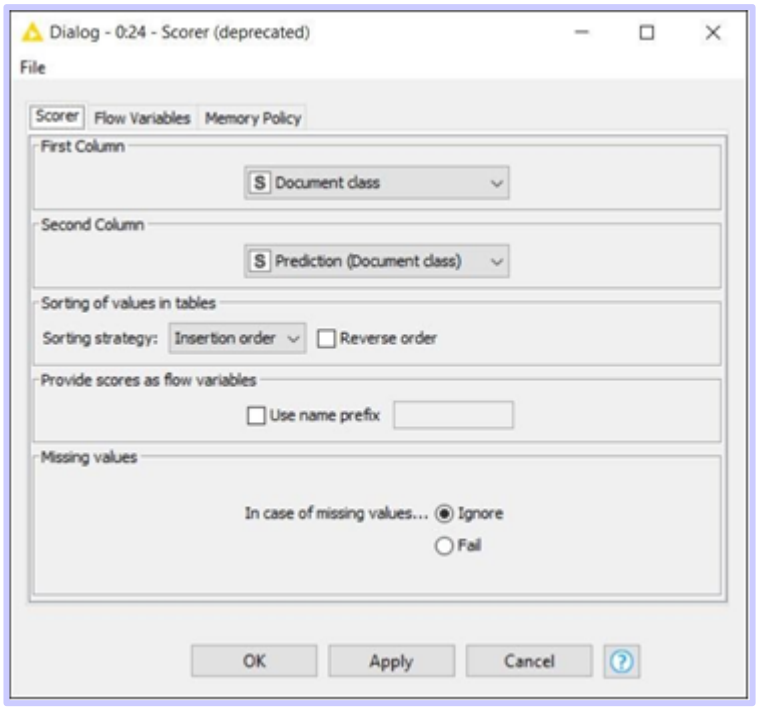
Nota. El Nodo Scorer (Goleador) como resultado muestra una matriz de confusión. Aplicamos y aceptamos.

- Configuración: Nodo Group Loop Start



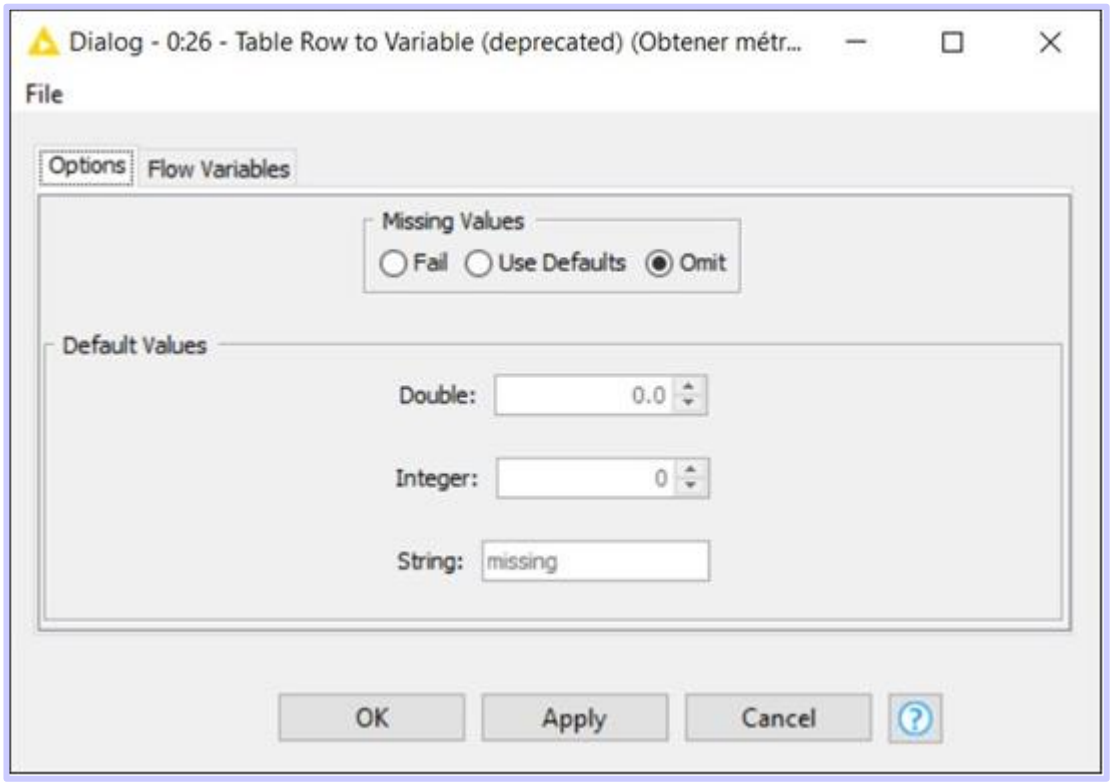
Nota. El Nodo Group Loop Start (Inicio de bucle de grupo) pasamos la columna fold #, este nodo sirve Iniciar el bucle de grupo, cada iteración procesa otro grupo de filas. Aplicamos y aceptamos.

- Configuración: Nodo Scorer



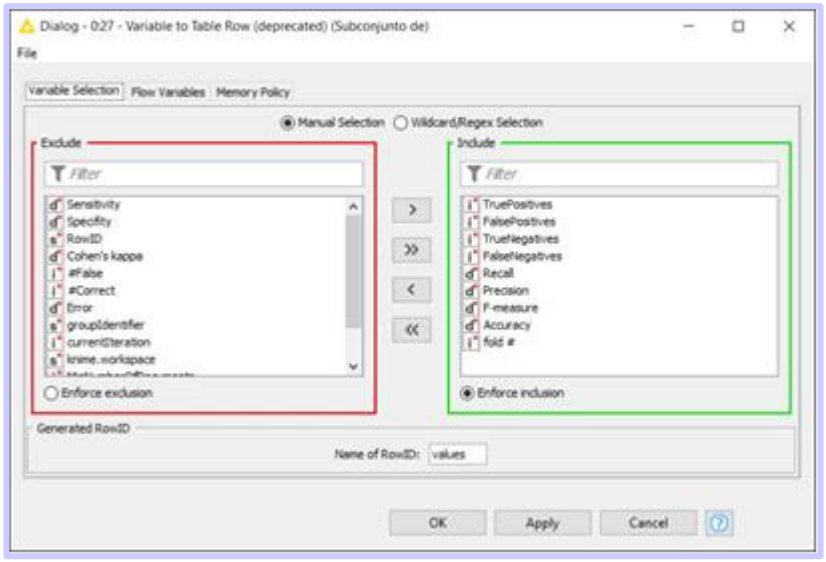
Nota. El Nodo Scorer (Goleador) como resultado muestra una matriz de confusión de la iteración 10. Aplicamos y aceptamos.

- Configuración: Nodo Table Row to Variable



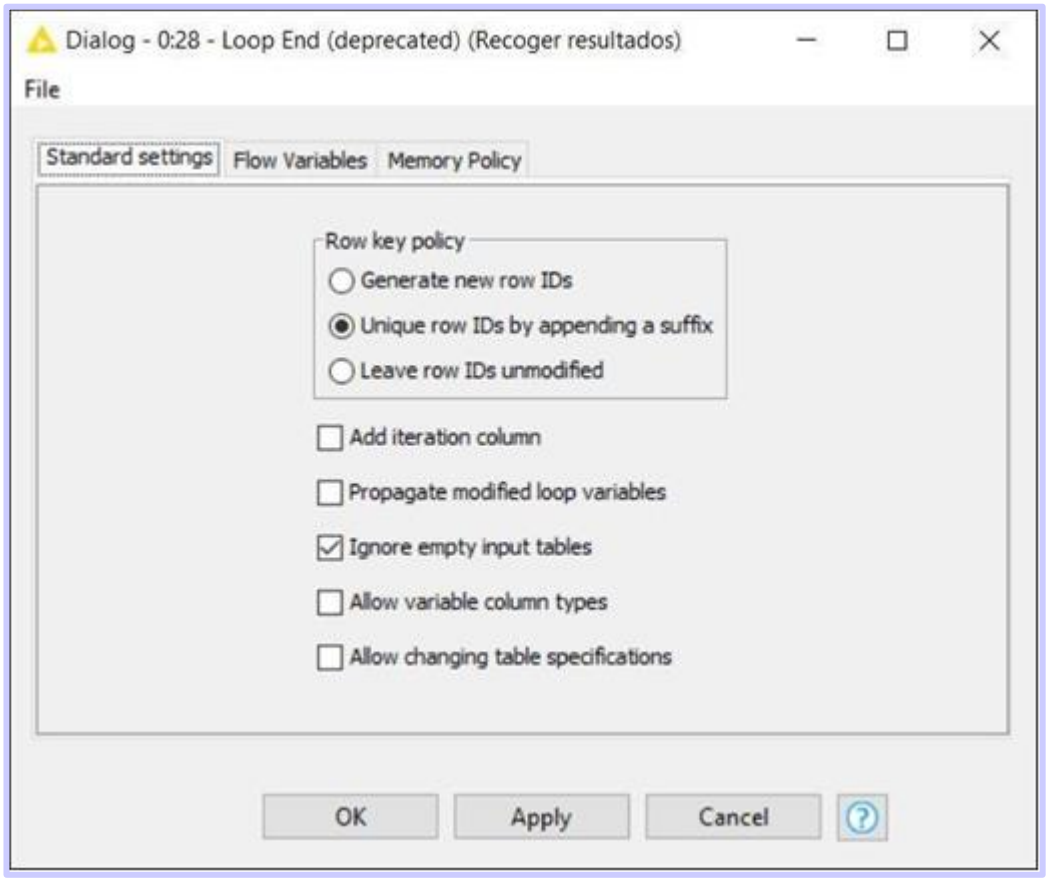
Nota. El Nodo Table Row to Variable (Fila de tabla a variable) Obtiene las métricas para la primera clase. Aplicamos y aceptamos.

- Configuración: Nodo Variable to Table Row



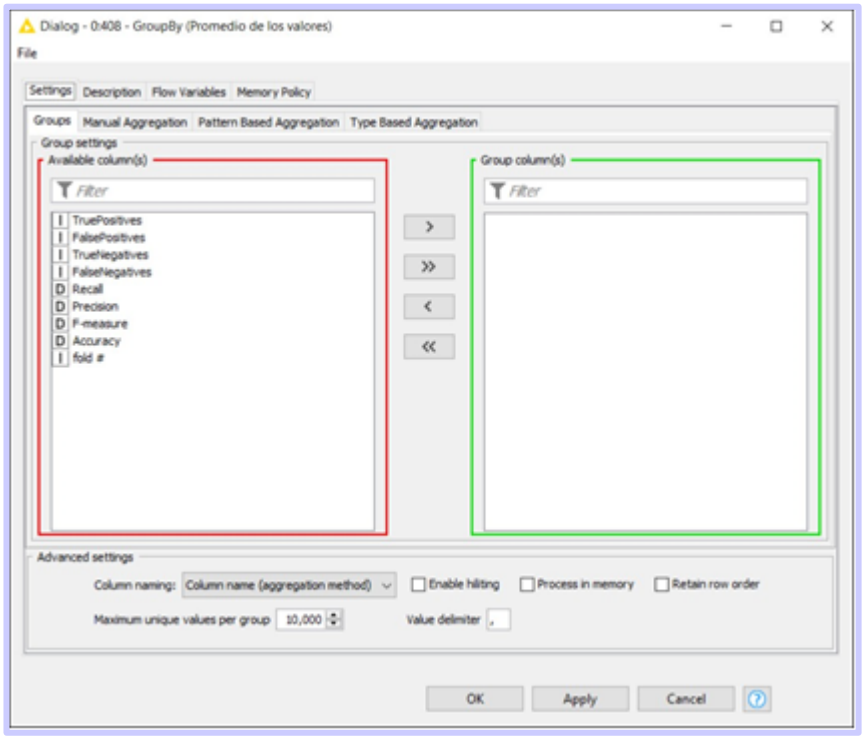
Nota. El Nodo Variable to Table Row (Variable a fila de tabla) Extrae variables y las coloca en una tabla de una sola fila. Extrae un subconjunto de métricas para rastrear. Aplicamos y aceptamos.

- Configuración: Nodo Loop End



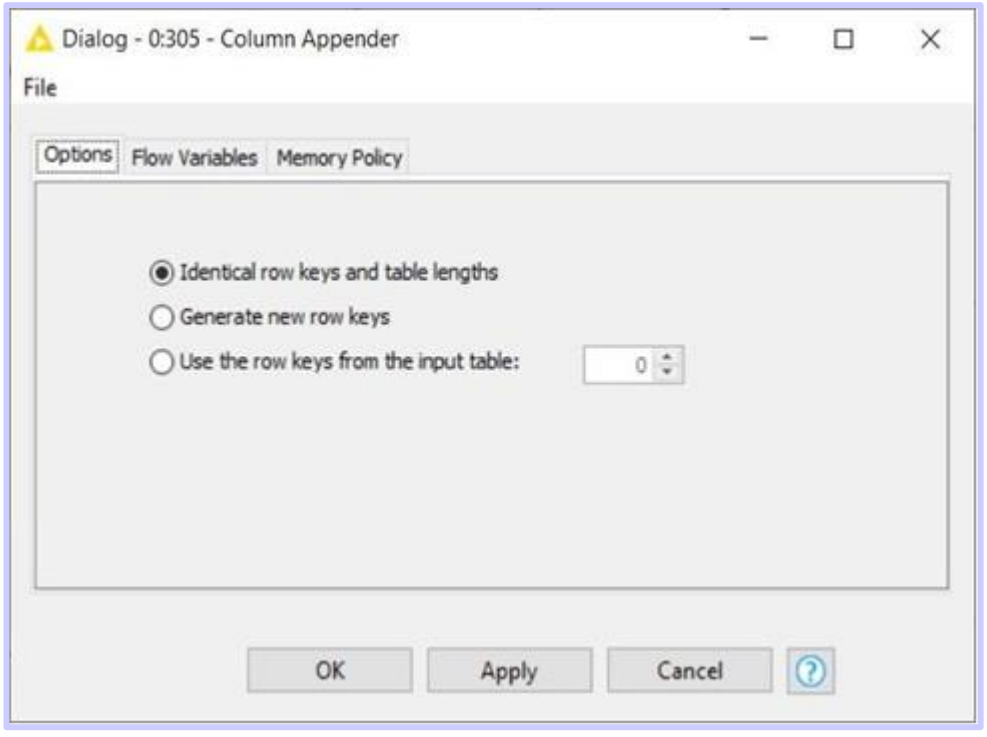
Nota. El Nodo Loop End (Fin de bucle) marcar el final de un ciclo y recoger resultados de cada iteración. Aplicamos y aceptamos.

- Configuración: Nodo GroupBy



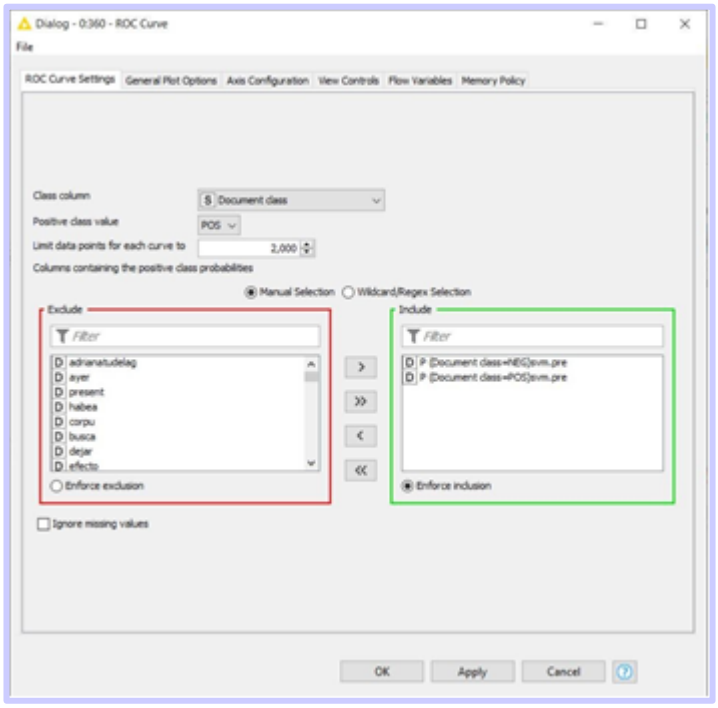
Nota. El Nodo GroupBy realiza el Promedio de los valores de las métricas. Aplicamos y aceptamos.

- Configuración: Nodo Column Appender



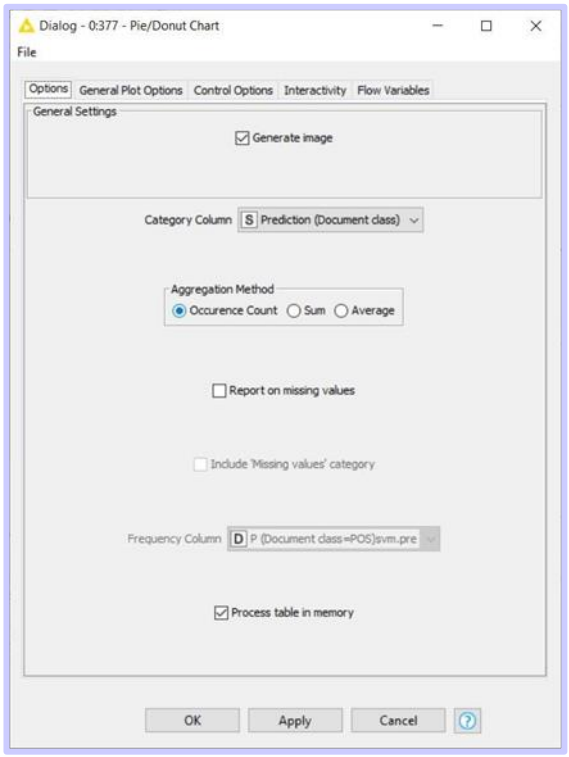
Nota. El Nodo Column Appender (Anexador de columna) toma dos tablas y las combina rápidamente agregando las columnas de la segunda tabla a la primera tabla.

- Configuración: Nodo ROC Curve



Nota. El Nodo ROC Curve (Curva ROC) gráfica curvas ROC para clasificación de dos clases. En este caso seleccionamos para la clase positiva. Aplicamos y aceptamos.

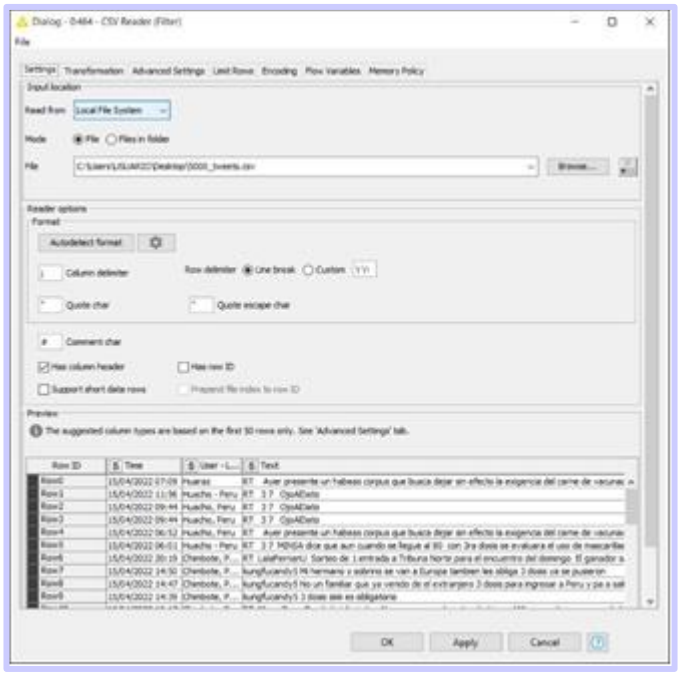
- Configuración: Nodo Pie/Donut Chart



Nota. El Nodo Pie/Donut Chart (Gráfico circular) muestra los resultados de las dos clases positivo y negativo. Aplicamos y aceptamos.

a. Clasificación de documentos: Aplicación del Modelo

- Configuración: Nodo CSV Reader

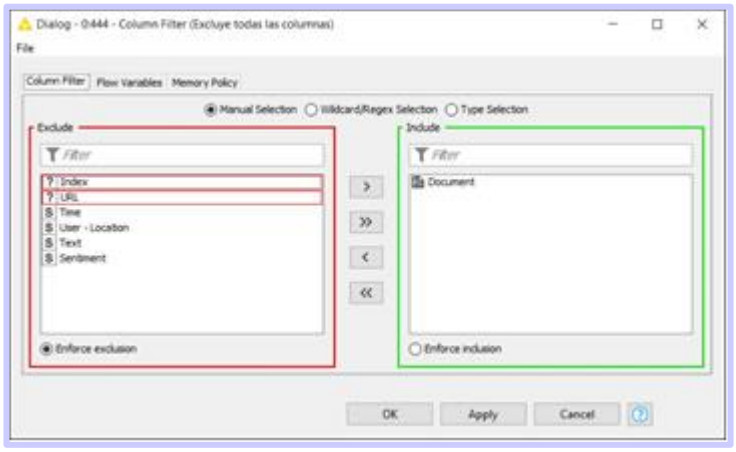


Nota. En el Nodo CSV Reader (Lector CSV) en el cual cargamos el archivo de 5000 tweets aplicamos y ya está listo para visualizar los datos.

- Configuración: Nodo String to Document

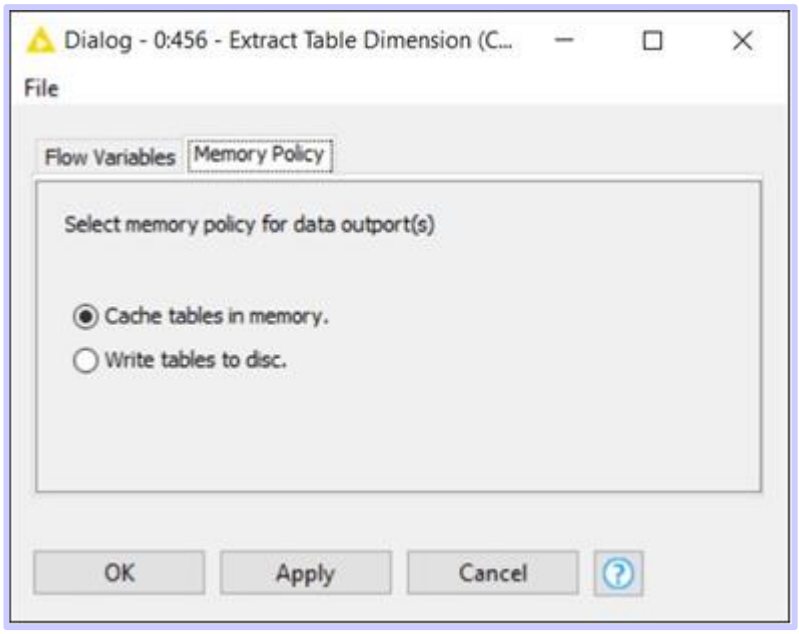
Nota. En el Nodo String to Document (Cadenas para documentar) seleccionamos Text, seleccionamos la categoría sentiment, tokenization en español aplicamos y aceptamos. Este nodo sirve para convertir cadenas en Documentos. Para cada fila se creará un documento.

- Configuración: Nodo Column Filter



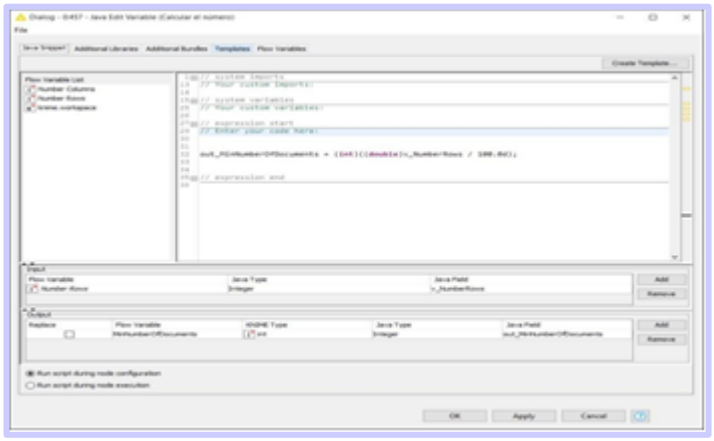
Nota. Este Nodo Column Filter (Filtro de columna) sirve para filtrar columnas, excluye todas las columnas y solo incluye la columna Document aplicamos y aceptamos.

- Configuración: Nodo Extract Table Dimension



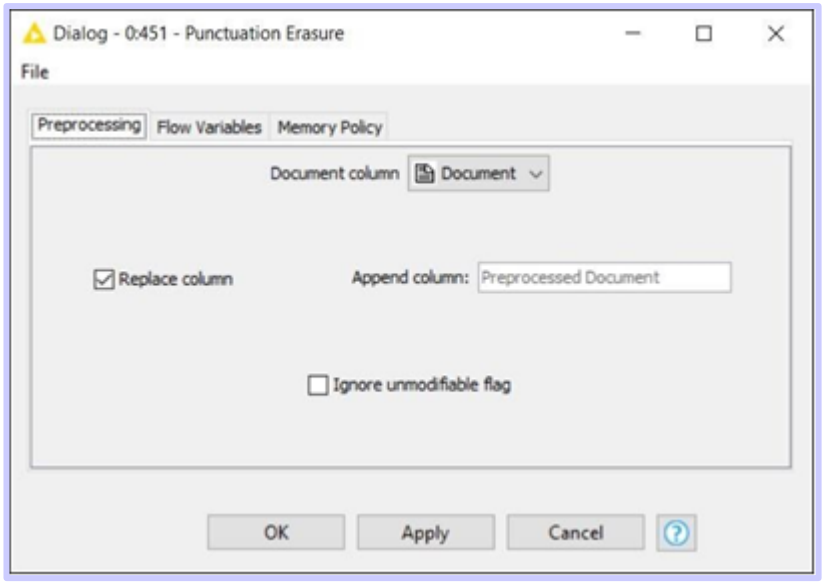
Nota. El Nodo Extract Table Dimension (Extraer dimensión de la tabla) sirve para extraer el número de filas y columnas de una tabla. Aplicamos y aceptamos.

- Configuración: Nodo Java Edit Variable



Nota. El Nodo Java Edit Variable (Variable de edición de Java) realiza el cálculo divide número de filas entre 100 y da un resultado. Aplicamos y aceptamos.

- Configuración: Nodo Punctuation Erasure

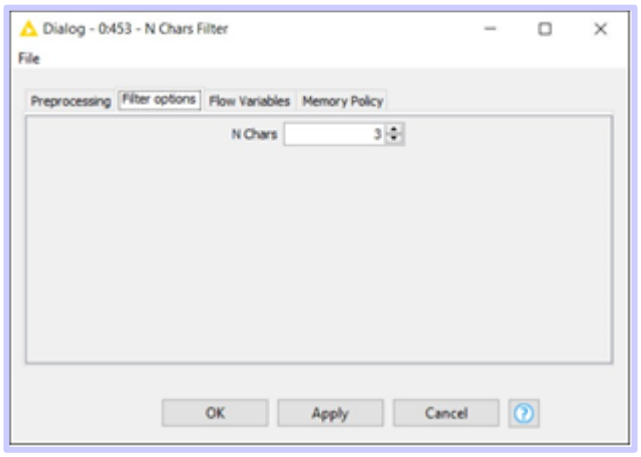


Nota. El Nodo Punctuation Erasure (Borrado de puntuación) elimina todos los caracteres de puntuación de los términos contenidos en los documentos, seleccionamos la columna Documento y reemplazamos la columna con el mismo. Aplicamos y aceptamos.

- Configuración: Nodo Number Filter

Nota. El Nodo Number Filter (Filtro de número) filtran todos los términos contenidos en los documentos que son dígitos "," o "." o "+" o "-". aplicamos y aceptamos.

- Configuración: Nodo N Chars Filter

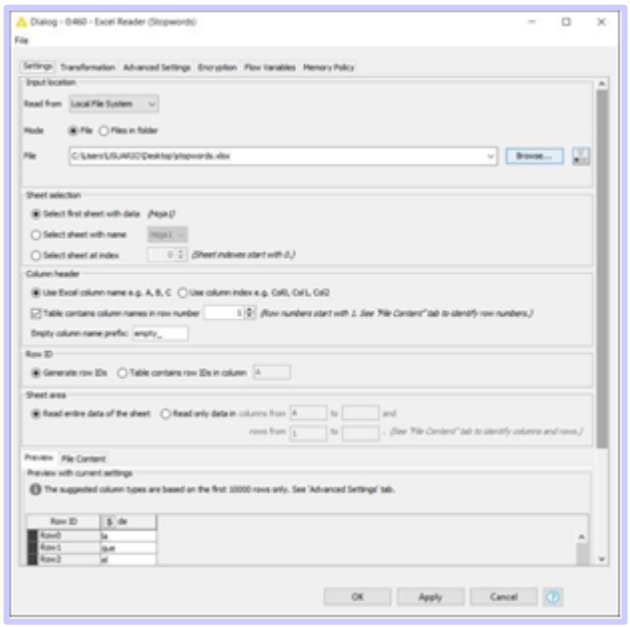


Nota. Nodo N Chars Filter (Filtro de caracteres N) Filtra todos los términos contenidos en los documentos con menos del número 3 caracteres. Aplicamos y aceptamos.

- Configuración: Nodo Stop Word Filter

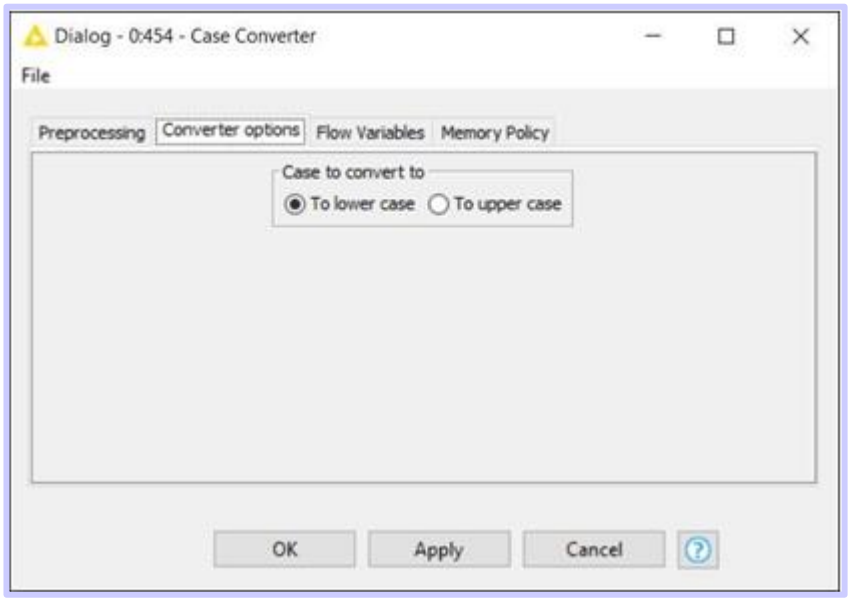
Nota. El Nodo Stop Word Filter (Filtro de palabras de detención) seleccionamos una lista de palabras vacías integradas en español y hacer check y personalizamos otra lista de palabras vacías que es un archivo stopwords.xlsx. Aplicamos y aceptamos.

- Configuración: Nodo Excel Reader



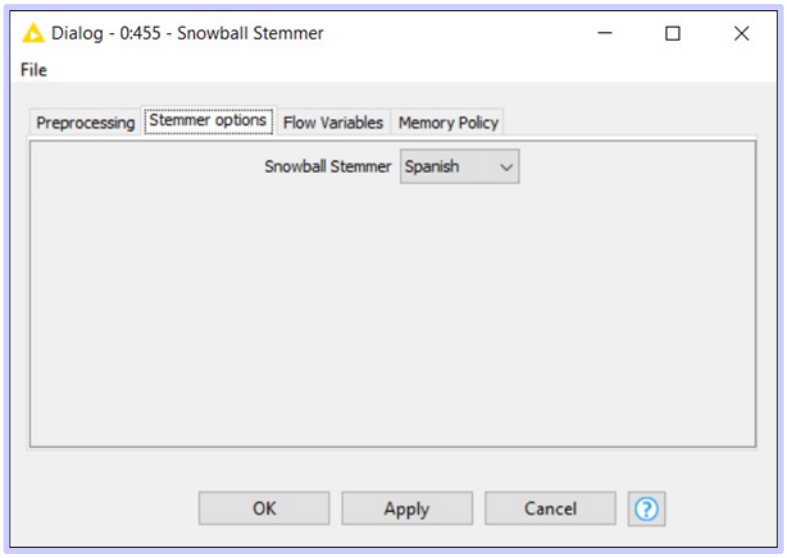
Nota. En el Nodo Excel Reader sirve para leer y cargamos el archivo stopwords.xlsx, estás son palabras como artículos, preposiciones, palabras que no ayudan a la investigación.

- Configuración: Nodo Case Converter



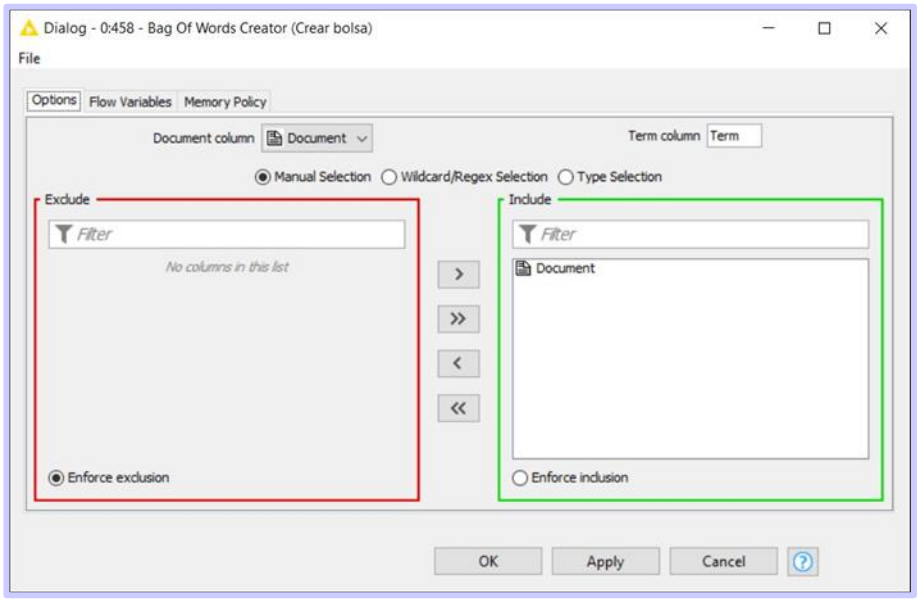
Nota. El Nodo Case Converter (Convertidor de casos) convierte caracteres a minúsculas. Aplicamos y aceptamos.

- Configuración: Nodo Snowball Stemmer



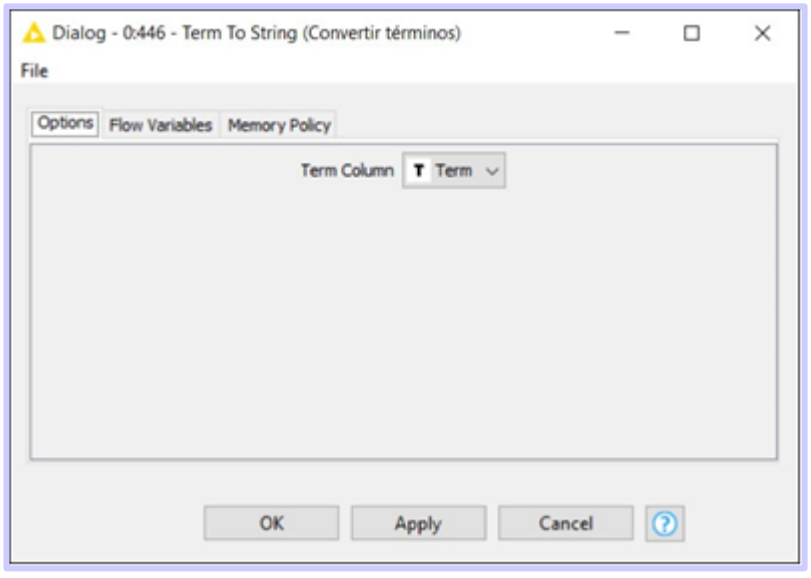
Nota. El Nodo Snowball Stemmer (Talador de bolas de nieve) reduce una palabra a su palabra base o raíz de modo que las palabras similares se encuentren en la raíz común. Seleccionamos el idioma español. Aplicamos y aceptamos.

- Configuración: Nodo Bag of Words Creator



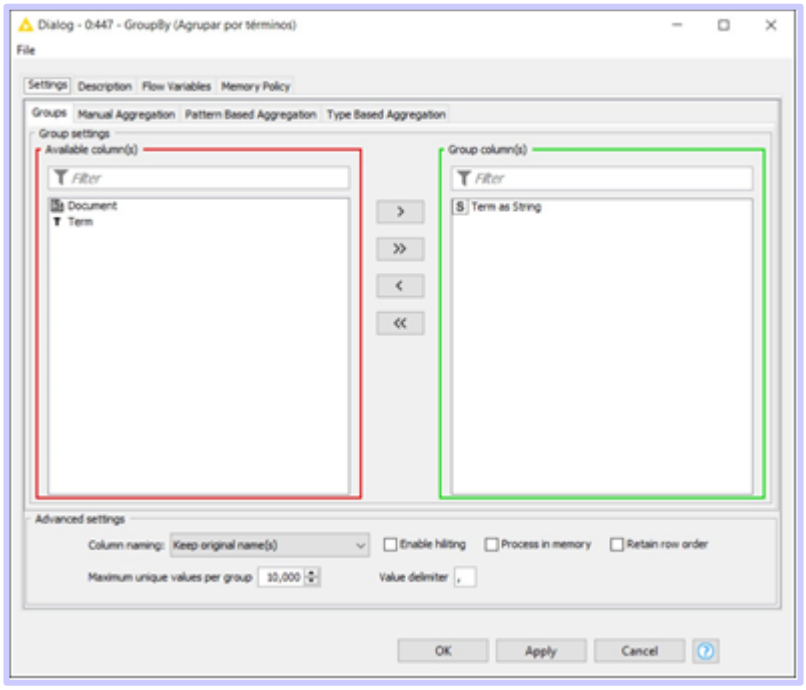
Nota. El Nodo Bag of Words Creator (Creador de Bolsa de Palabras) crea una bolsa de palabras de un conjunto de documentos. Una Bolsa de palabras consta de una columna que contiene los términos que aparecen en el documento. Seleccionamos el Documento aplicamos y aceptamos.

- Configuración: Nodo Term to String



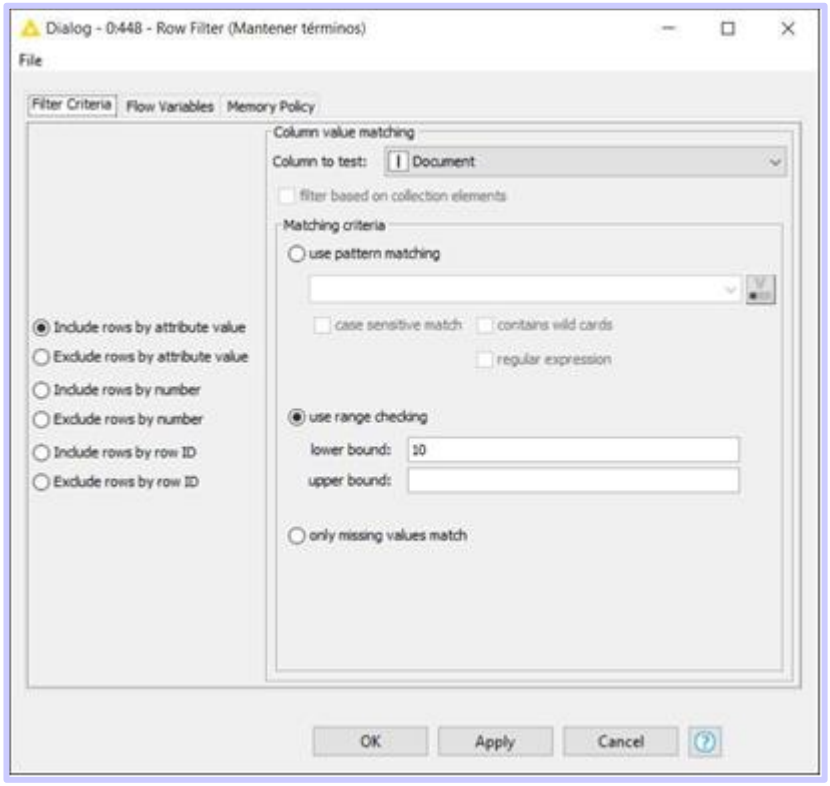
Nota. El Nodo Term to String (Término a cadena) convierte los términos de una columna en cadenas y adjunta una nueva columna con estas cadenas. se pierden las etiquetas de los términos. Aplicamos y aceptamos.

- Configuración: Nodo GroupBy



Nota. El Nodo GroupBy (Agrupar por) Agrupa las filas de una tabla por los valores únicos en columnas del grupo seleccionado. Agrupa por términos. Aplicamos y aceptamos.

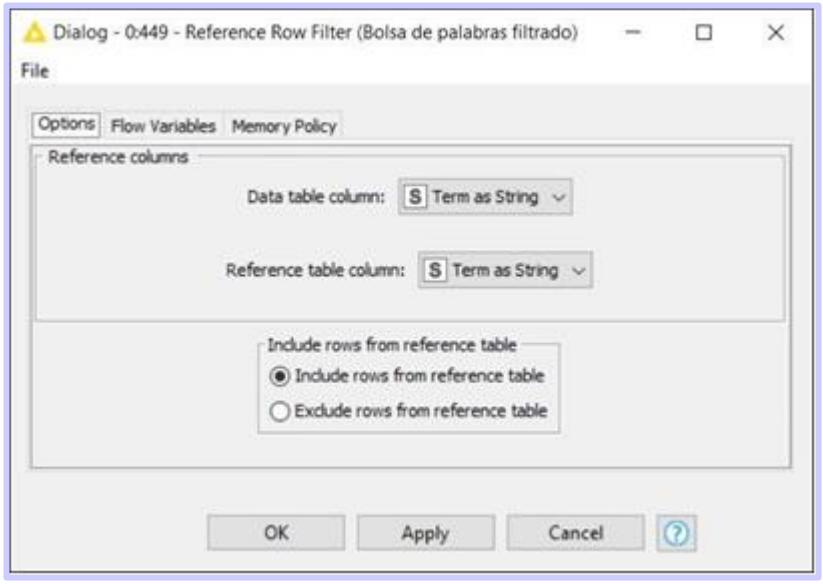
- Configuración: Nodo Row Filter



Nota. El Nodo Row Filter (Filtro de fila) hace filtrado de filas según ciertos criterios.

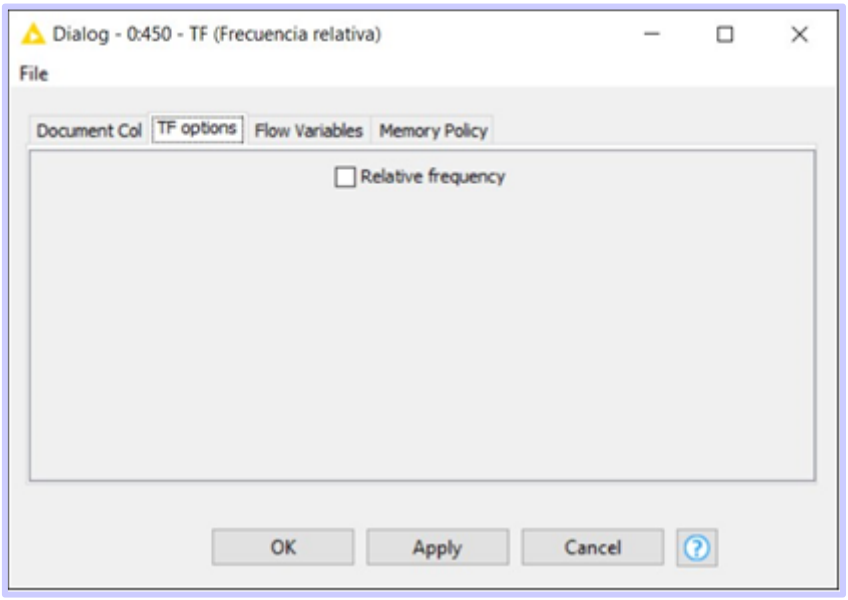
Mantiene términos que aparezcan en un número mínimo de documentos. Aplicar y aceptar.

- Configuración: Nodo Reference Row Filter



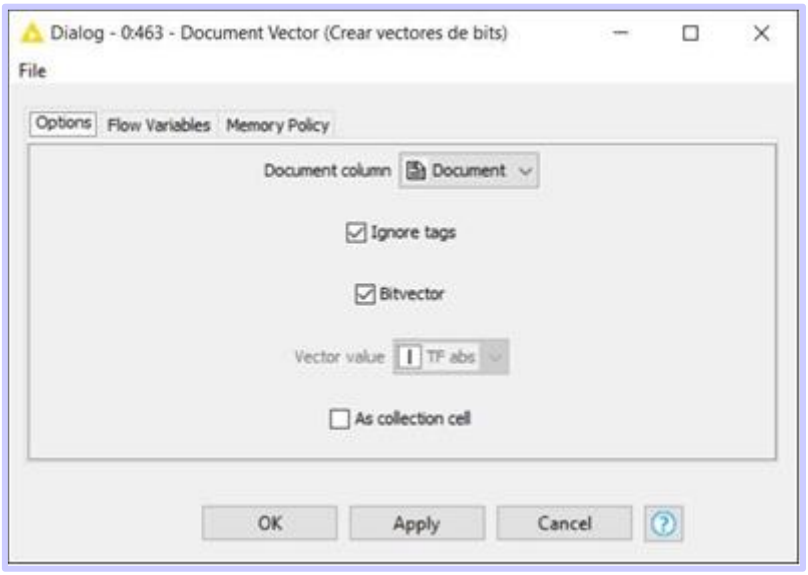
Nota. El Nodo Reference Row Filter (Filtro de fila de referencia) filtra filas de la primera tabla utilizando la segunda tabla como referencia, Bolsa de palabras filtrado. Aplicamos.

- Configuración: Nodo TF



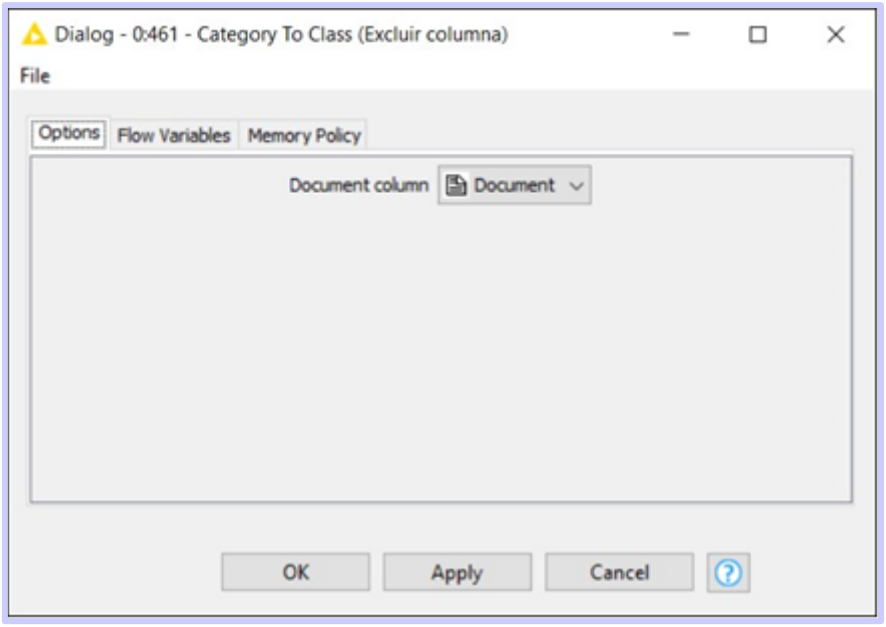
Nota. El Nodo TF Calcula la frecuencia relativa del término (tf) de cada término según cada documento y añade una columna que contiene el valor de tf. aplicamos y aceptamos.

- Configuración: Nodo Document Vector



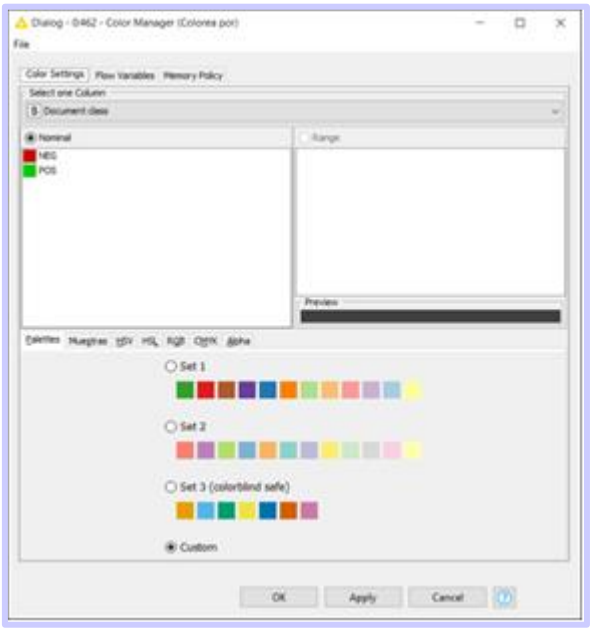
Nota. El Nodo Document Vector (Vector de documento) Crea vectores de bits para documentos. el bit 0 y el bit 1. aplicamos y aceptamos.

- Configuración: Nodo Category to Class



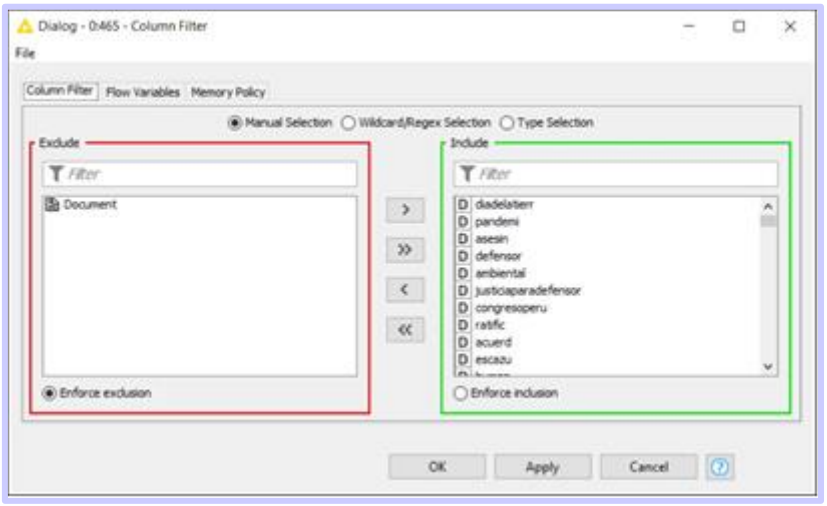
Nota. Nodo Category to Class (Categoría a clase) Excluye la columna sentiment, sirve para agregar una columna de clase (cadena) a cada fila que contiene una celda de documento. El valor de la clase es la categoría del documento como cadena. Aplicamos y aceptamos.

- Configuración: Nodo Color Manager



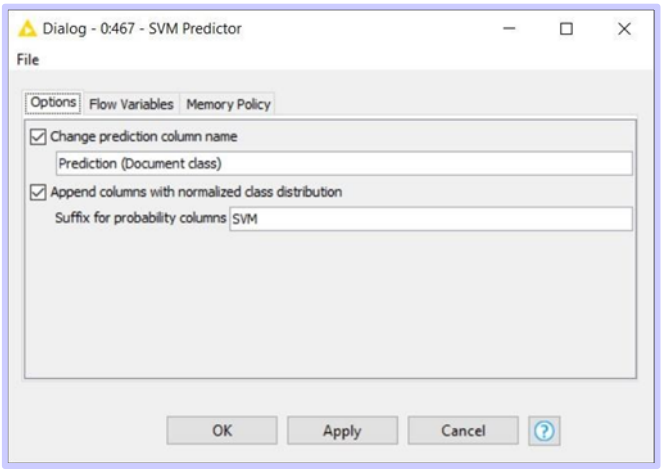
Nota. El Nodo Color Manager (Administrador de color) colorea cada categoría verde para positivo y rojo para negativo se asigna a columnas nominales. Aplicamos y aceptamos.

- Configuración: Nodo Column Filter



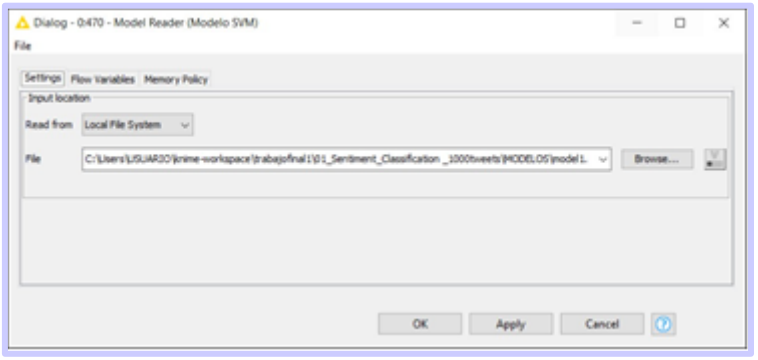
Nota. El Nodo Column Filter (Filtro de columna) filtra las columnas de la tabla mientras que solo las columnas restantes se pasan a la tabla. Aplicamos y aceptamos.

- Configuración: Nodo SVM Predictor



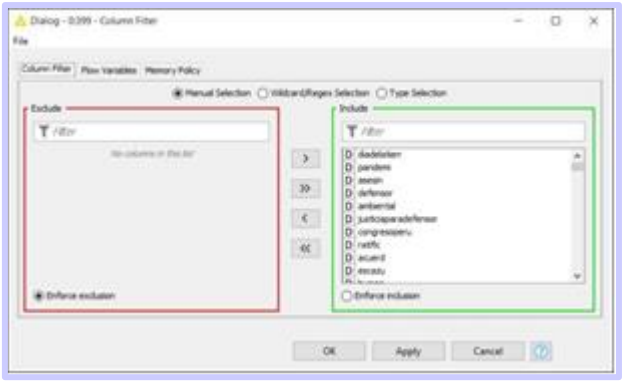
Nota. El Nodo SVM Predictor (Predictor de SVM) emplea un modelo de SVM generado por el nodo de aprendizaje de SVM para predecir el resultado de sentimiento positivo o negativo. Aplicamos y aceptamos.

- Configuración: Nodo Model Reader



Nota. El Nodo Model Reader cargamos un archivo en el cual se escribió y guardo el modelo.

- Configuración: Nodo Column Filter

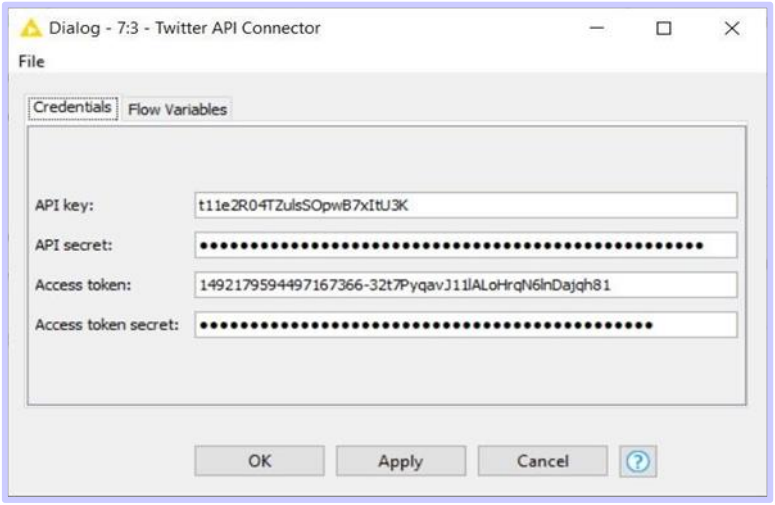


Nota. Nodo Column Filter (Filtro de columna) filtra las columnas de la tabla mientras que solo las columnas restantes se pasan a la tabla. Aplicamos y

aceptamos.

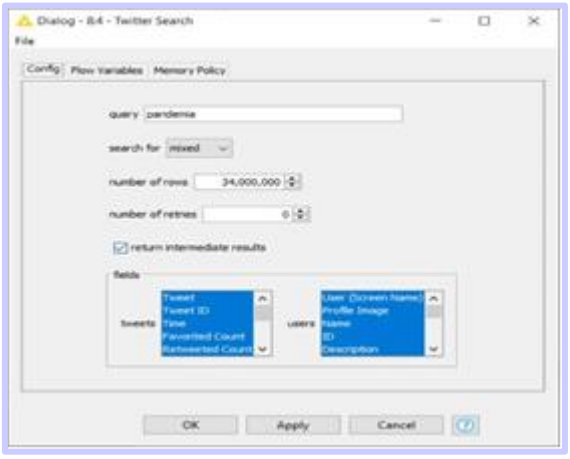
Procedimiento para descargar y limpieza de los tweets

- Configuración: Nodo Twitter API Connector



Nota. El Nodo Twitter API Connector (Conector API de Twitter) Rellenamos los campos de API key, API secret, Access token, Access token secret. Este Nodo Crea una conexión para acceder a la API de Twitter. Deberá iniciar sesión con su cuenta de Twitter en dev.twitter.com y registrar su aplicación en "Mis aplicaciones" para obtener su clave de API y token de acceso. Aplicamos y aceptamos.

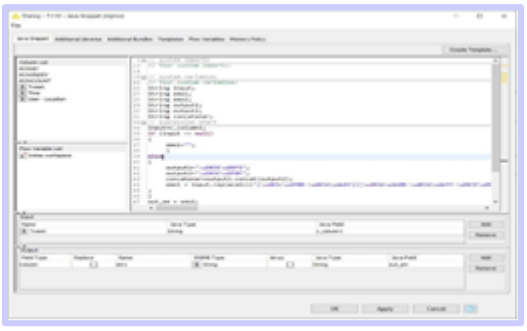
- Configuración: Nodo Twitter Search



Nota. El Nodo Twitter Search (Búsqueda de Twitter) busca por término “pandemia” y seleccionamos todos los campos. Aplicamos y aceptamos.

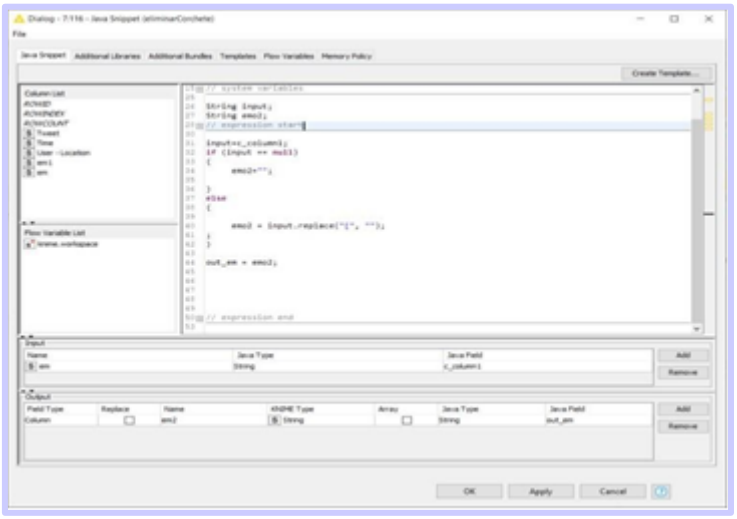
Nota. El Código es para eliminar emojis, eliminamos carácter Unicode. Copiamos y pegamos el código en el Nodo Java Snippet (Fragmento de Java). aplicamos y aceptamos.

- Configuración: Nodo Java Snippet



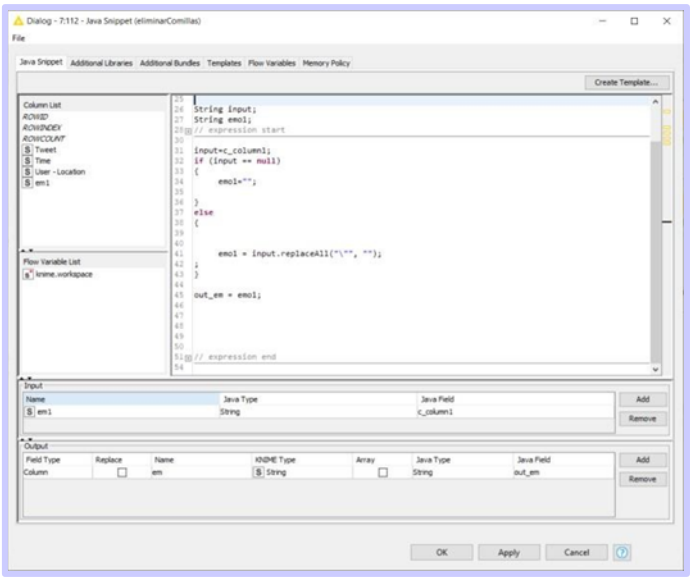
Nota. El Nodo Java Snippet (Fragmento de Java) para eliminar comillas. Aplicamos y aceptamos.

- Configuración: Nodo Java Snippet



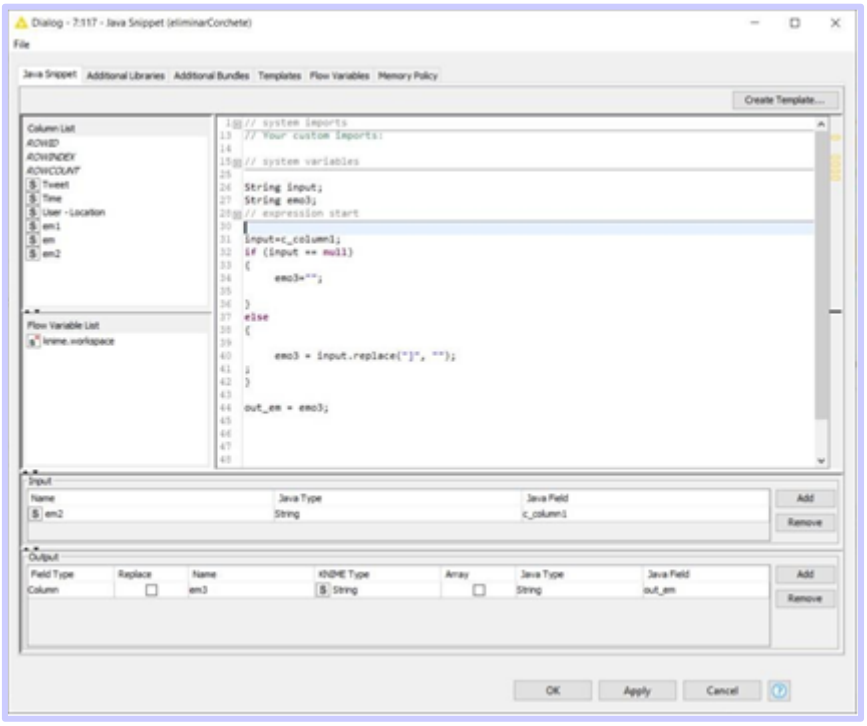
Nota. El Nodo Java Snippet (Fragmento de Java) para eliminar corchetes. Aplicamos y aceptamos.

- Configuración: Nodo Java Snippet



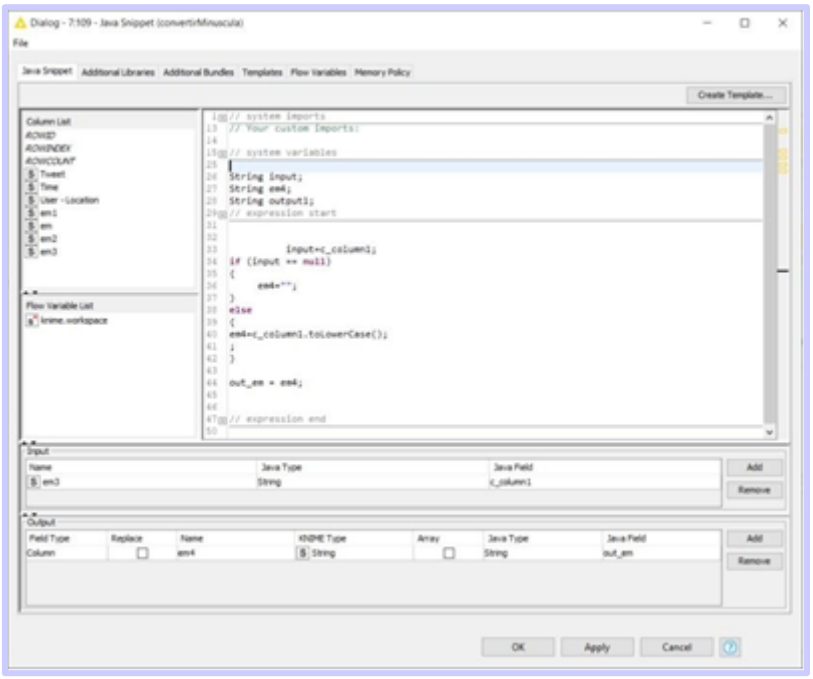
Nota. El Nodo Java Snippet (Fragmento de Java) para eliminar emojis. Aplicamos y aceptamos.

- Configuración: Nodo Java Snippet



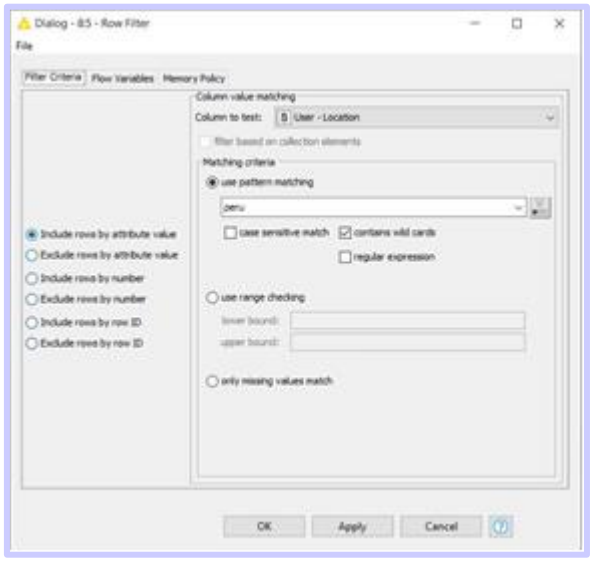
Nota. El Nodo Java Snippet (Fragmento de Java) Eliminar corchetes. Aplicamos y aceptamos.

- Configuración: Nodo Java Snippet



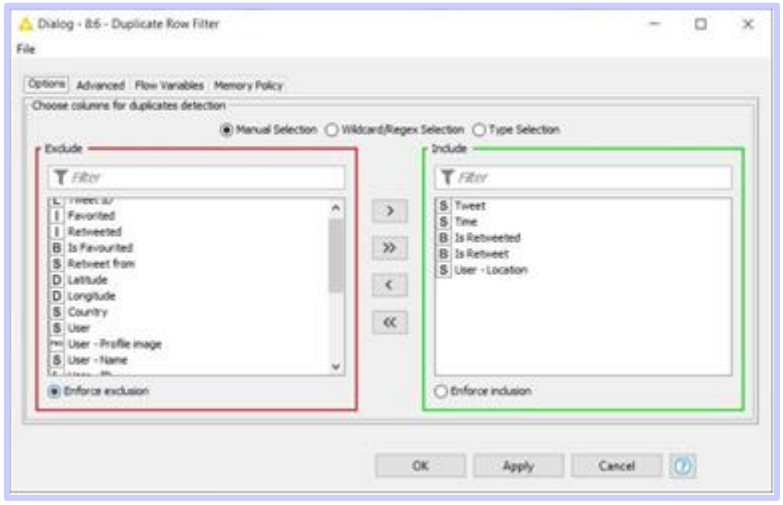
Nota. El Nodo Java Snippet (Fragmento de Java) es para Convertir a minúsculas. Aplicamos y aceptamos.

- Configuración: Nodo Row Filter



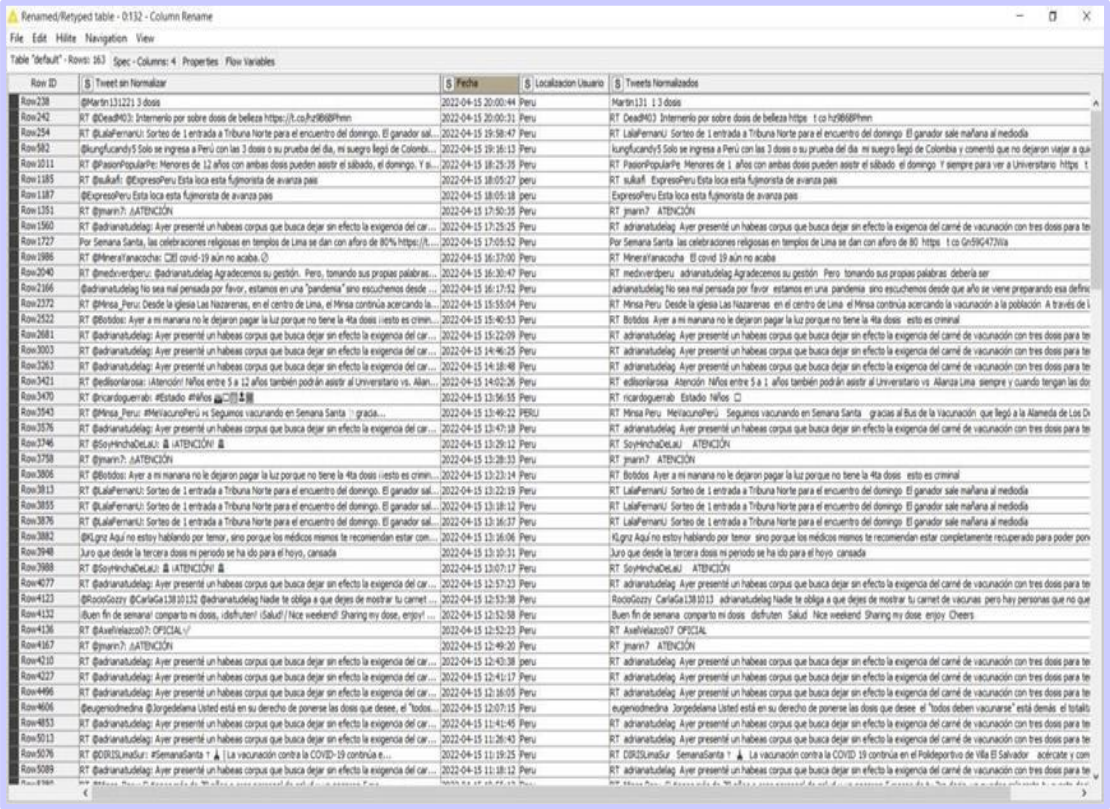
Nota. El Nodo Row Filter (Filtro de fila) es para filtrar la fila por User-Location (localización de usuario), en este caso filtramos por criterio “Perú”. Aplicamos y aceptamos.

- Configuración: Nodo Duplicate Row Filter



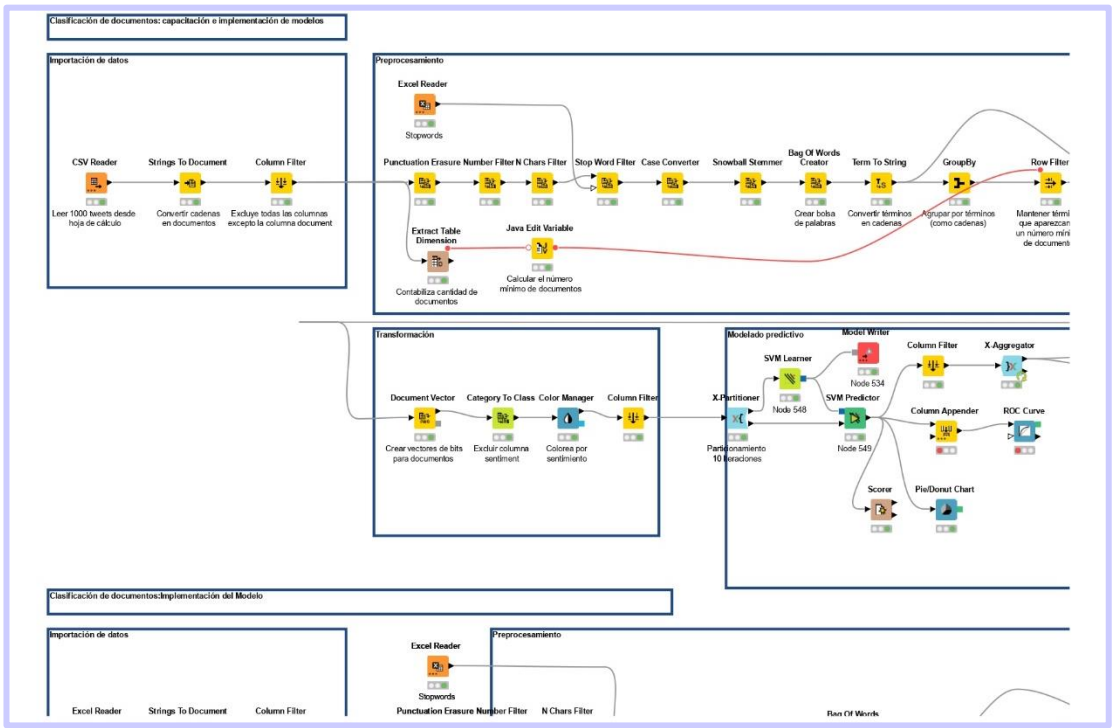
Nota. El Nodo Duplicate Row Filter (Filtro de fila duplicada) Elimina las filas duplicadas. Aplicamos y aceptamos.

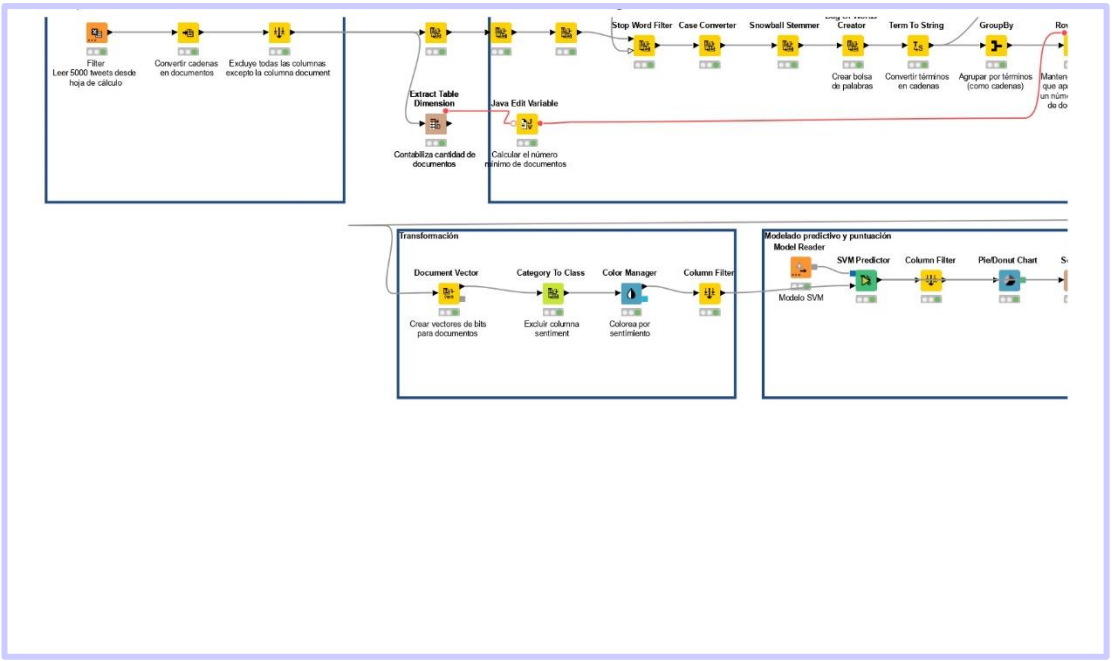
- Resultado de tweets



Nota. Se muestra un ejemplo en la parte izquierda esta los tweets sin limpieza y en la parte derecha están los tweets limpios.

Diseño del algoritmo de predicción Máquinas de Vectores de Soporte



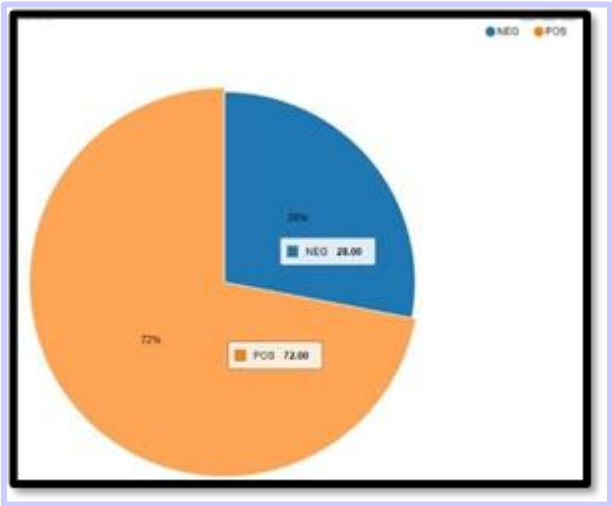

Filter
Leer 5000 tweets desde hoja de cálculo
Convertir cadenas en documentos
Excluye todas las columnas excepto la columna document
Extract Table Dimension
Contabiliza cantidad de documentos
Java Edit Variable
Calcular el número mínimo de documentos
Stop Word Filter
Case Converter
Snowball Stemmer
Creator
Crear bolsa de palabras
Term To String
Convertir términos en cadenas
GroupBy
Agrupar por términos (como cadenas)
Transformación
Document Vector
Crear vectores de bits para documentos
Category To Class
Excluir columna sentiment
Color Manager
Colorea por sentimiento
Column Filter
Modelado predictivo y puntuación
Model Reader
Modelo SVM
SVM Predictor
Column Filter
Pie/Donut Chart

Anexo 3.

Figuras del capítulo IV Resultados

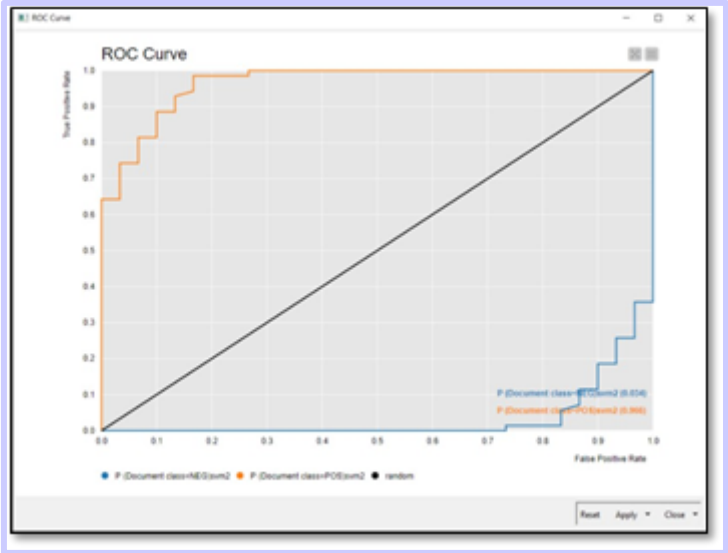
Análisis de los resultados de la Clasificación de documentos: implementación del modelo.

- Resultado Tweets de prueba 100



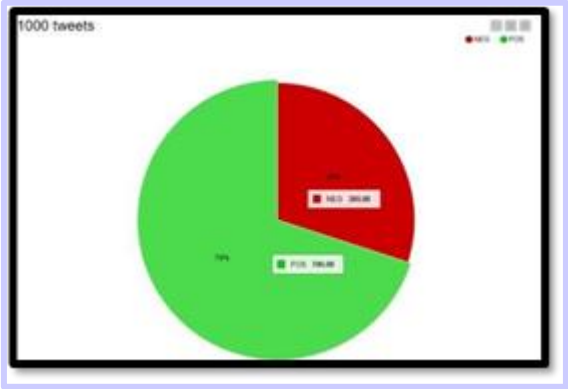
Nota. Figura 14: Tweets de prueba 100, muestra 72% de tweets positivos y 28% de tweets negativos.

- Resultado Curva ROC



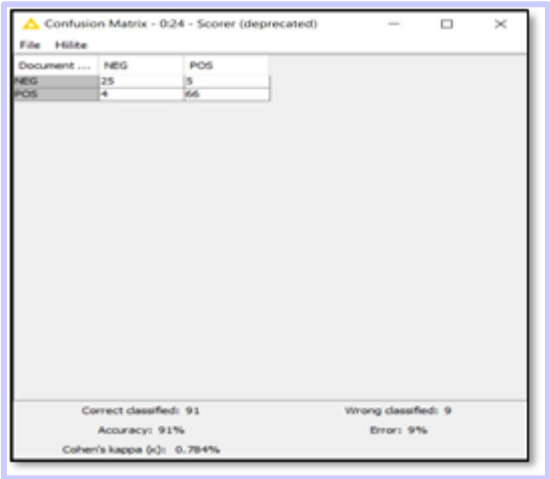
Nota. Figura 15: Curva ROC. Donde el color naranja representa a la curva ROC de la clase positivo y el color azul representa a la curva ROC de la clase negativo.

- Resultado muestra 1000 tweets



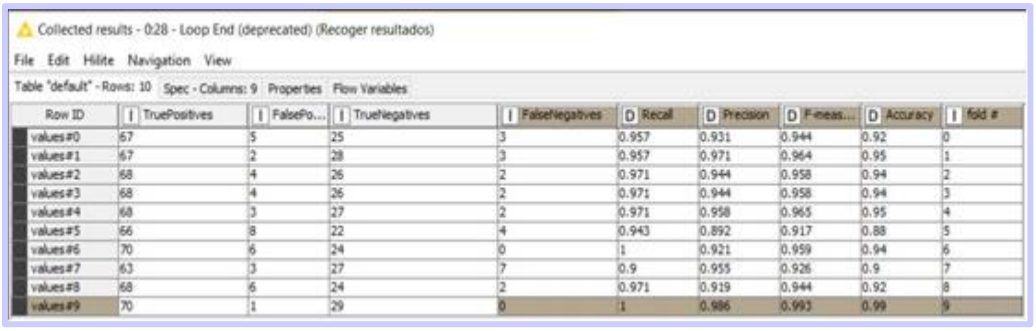
Nota. Figura 35. Posteriormente de la clasificación de los tweets etiquetados como positivo y negativo. Los tweets están listos para aplicar algoritmo de SVM. se muestra 1000 tweets de los cuales el 70% son positivos y 30% son negativos.

- Resultado Matriz de la iteración N° 10



Nota. Figura 36. Matriz de la iteración N° 10, muestra la matriz de confusión de la iteración 9. correctamente clasificados 91 e incorrectamente clasificados 9, error de 9% y la exactitud de 91%.

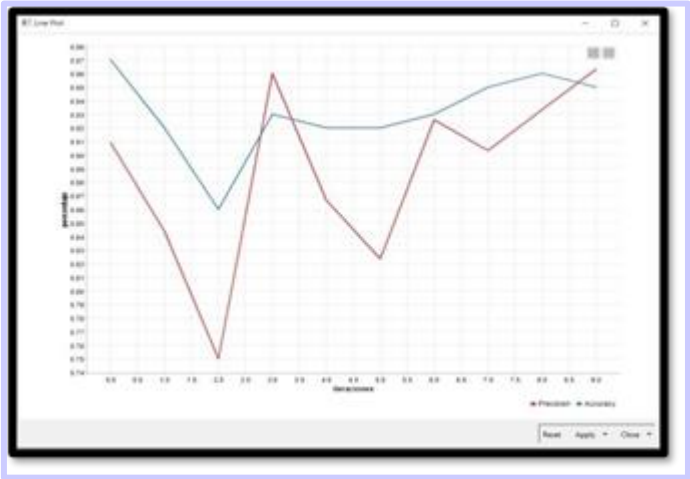
- Resultado *de métricas de evaluación*



Row ID	TruePositives	FalsePo...	TrueNegatives	FalseNegatives	Recall	Precision	F-meas...	Accuracy	fold #
values#0	67	5	25	3	0.957	0.931	0.944	0.92	0
values#1	67	2	28	3	0.957	0.971	0.964	0.95	1
values#2	68	4	26	2	0.971	0.944	0.958	0.94	2
values#3	68	4	26	2	0.971	0.944	0.958	0.94	3
values#4	68	3	27	2	0.971	0.958	0.965	0.95	4
values#5	66	8	22	4	0.943	0.892	0.917	0.88	5
values#6	70	6	24	0	1	0.921	0.959	0.94	6
values#7	63	3	27	7	0.9	0.955	0.926	0.9	7
values#8	68	6	24	2	0.971	0.919	0.944	0.92	8
values#9	70	1	29	0	1	0.986	0.993	0.99	9

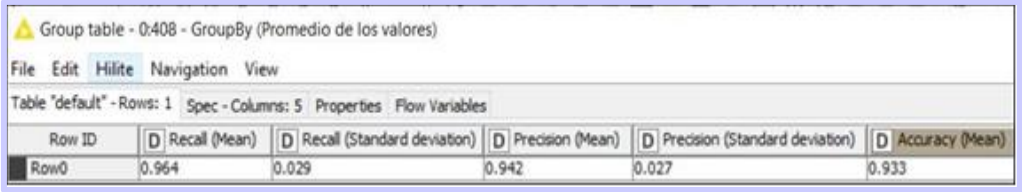
Nota. Figura 37. Determinación de las métricas de evaluación en cada iteración Nos muestra las diez matrices de confusión de las 10 iteraciones.

- Resultado Validación cruzada, exactitud y precisión



Nota. Figura 38. Validación cruzada, exactitud(azul) y precisión(rojo) Nos muestra las 10 iteraciones de Accuracy y las 10 iteraciones de precisión.

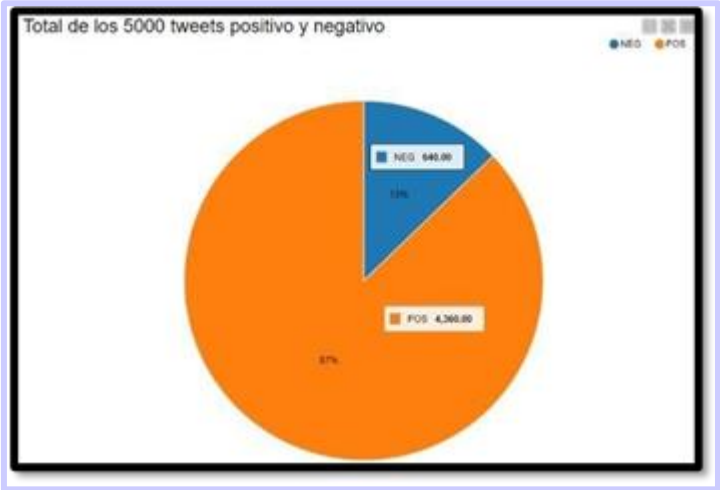
- Resultado de Promedio de valores



Row ID	D Recall (Mean)	D Recall (Standard deviation)	D Precision (Mean)	D Precision (Standard deviation)	D Accuracy (Mean)
Row0	0.964	0.029	0.942	0.027	0.933

Nota. Figura 39. Promedio de valores Nos muestra el promedio de las 10 iteraciones de Accuracy es 0.933 y el promedio de precisión es 0.027.

- Resultado de Aplicación de 5000 tweets



Nota. Figura 40. Aplicación de 5000 tweets Nos muestra el 87% de tweets son negativos y el 13% de tweets son negativos.

- Resultado de Frecuencia de palabras

Nota. Figura 23. Frecuencia de palabras Nos muestra la palabra que mayormente aparece es dosis 635 veces.

- Resultado de Nube de palabras tweets de los usuarios.

*Nota.*Figura 24. Nube de palabras tweets de los usuarios. Nos muestra la palabra “dosis” aparece mayormente por eso se muestra esa palabra grande.

- Resultado de tiempo de extracción de tweets.

Nota. Figura de tiempo de extracción de tweets. Nos muestra el tiempo en minutos, mientras más tweets descargamos más tiempo nos demoramos.

- Resultados número de tweets mes 3 de abril.

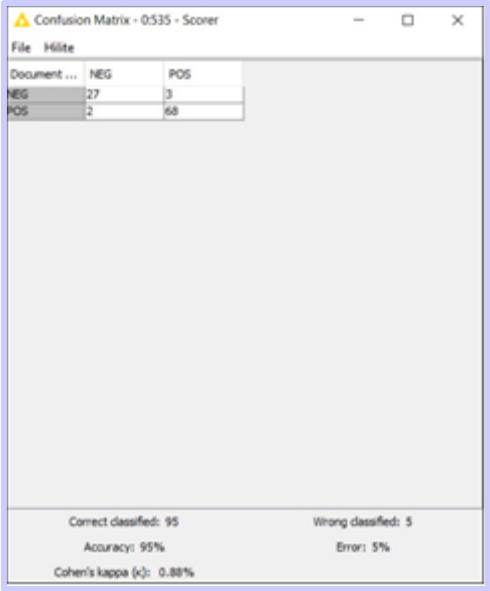
Nota. Figura de número de tweets de fecha 3 de abril. Nos muestra en esta fecha solo se recolecto 9 tweets positivos y 40 tweets negativos.

- Resultado de Sentimiento de los tweets.

Nota. Figura 4. Sentimiento de los tweets. Nos muestra el sentimiento de los tweets 87% positivos y 13% negativos.

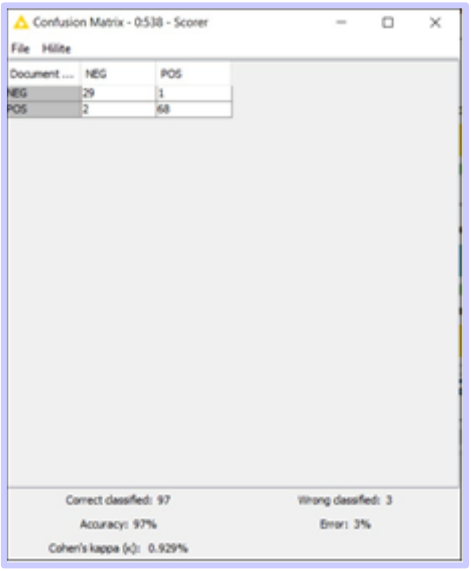
Resultado de la evaluación 1: Análisis de los resultados de los algoritmos clasificadores de Árbol de decisión y Naive Bayes.

- Resultado de Matriz de confusión



Nota. Matriz de confusión Máquinas de vectores de soporte (Support vector machine). nos muestra el 95% de exactitud (Accuracy).

- Resultado de Matriz de confusión Árbol de decisión Decision tree

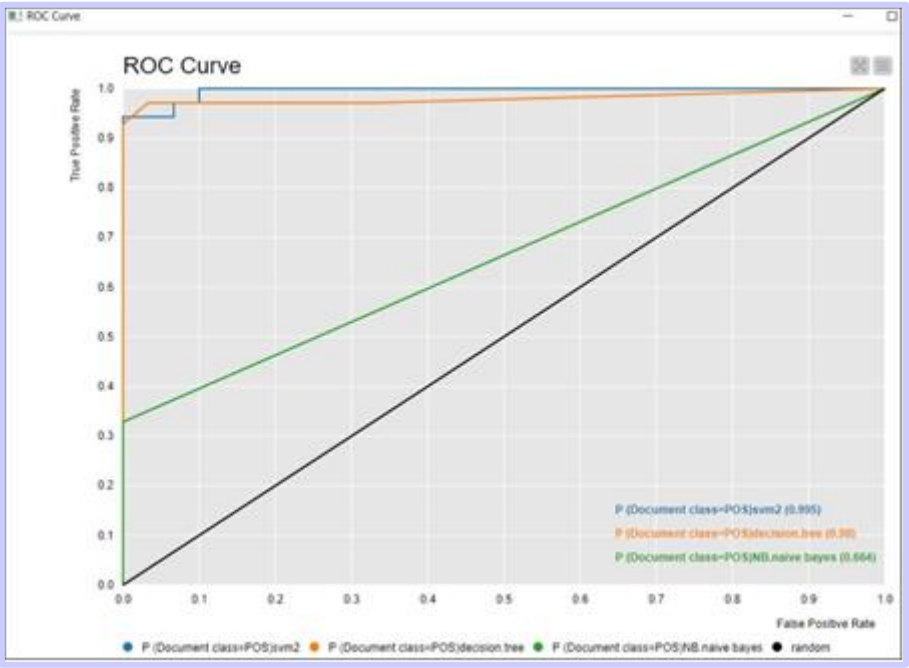


Nota. Matriz de confusión Árbol de decisión (Decision tree). nos muestra el 97% de exactitud (Accuracy).

- Resultado Matriz de confusión Naive Bayes

Nota. Matriz de confusión Bayesiano Ingenuo (Naive Bayes). nos muestra el 30% de exactitud (Accuracy).

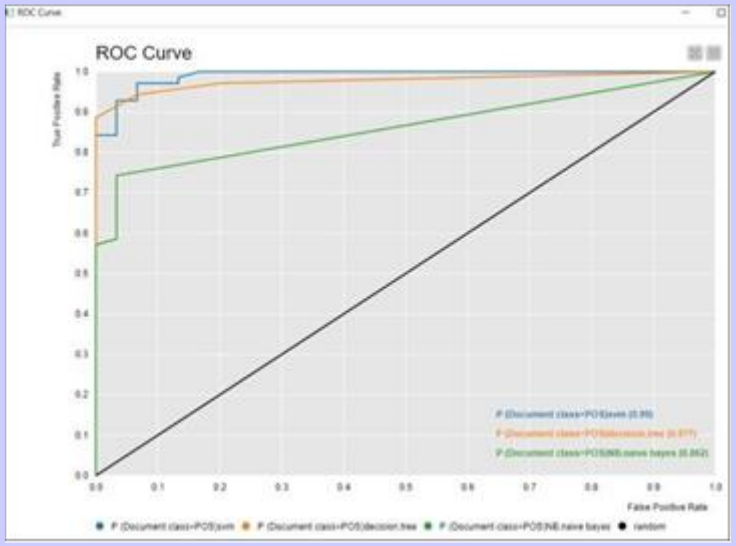
- Resultado de Curva Roc de los 3 clasificadores Support-vector machine, Decision tree y Naive Bayes.



Nota. Realizamos la primera ejecución de Curva Roc de los 3 clasificadores Support-vector machine, Decision tree y Naive Bayes. nos muestra una comparación de las tres curvas ROC, la mejor curva Roc es Support-vector

machine, el área bajo la curva de 0.995 indica que es muy bueno el rendimiento del algoritmo.

- Resultado de comparación de las tres curvas ROC.



Nota. Reseteamos el particionamiento de 10 iteraciones y realizamos por segunda vez la ejecución nuevamente y nos muestra una comparación de las tres curvas ROC, aún sigue estando mejor la curva Roc de Support-vector machine, el área bajo la curva de 0.99 indica que es muy bueno el rendimiento del algoritmo.

Análisis de las métricas de evaluación de los 3 clasificadores Support-vector machine, Decision tree y Naive Bayes.

- Resultado metricas de Máquinas de vectores de soporte

Collected results - 0:28 - Loop End (deprecated) (Recoger resultados)

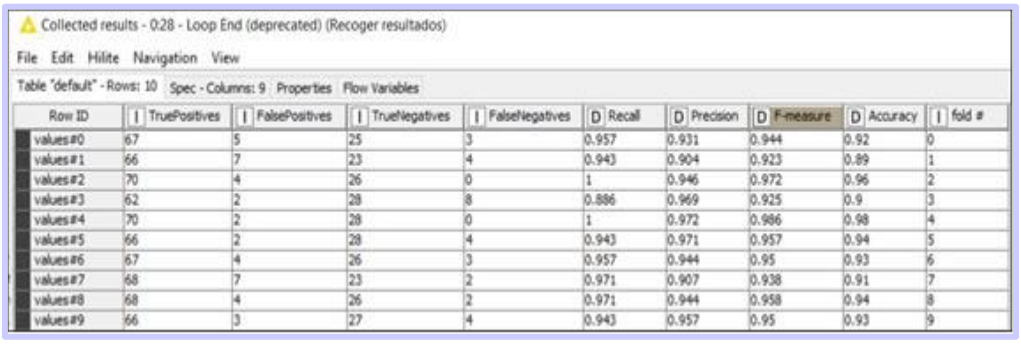
File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 9 Properties Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D F-meas...	D Accuracy	I fold #
values#0	64	3	27	6	0.914	0.955	0.934	0.91	0
values#1	66	2	28	4	0.943	0.971	0.957	0.94	1
values#2	68	3	27	2	0.971	0.958	0.965	0.95	2
values#3	69	6	24	1	0.986	0.92	0.952	0.93	3
values#4	70	3	27	0	1	0.959	0.979	0.97	4
values#5	63	4	26	7	0.9	0.94	0.92	0.89	5
values#6	65	3	27	5	0.929	0.956	0.942	0.92	6
values#7	65	7	23	5	0.929	0.903	0.915	0.88	7
values#8	68	4	26	2	0.971	0.944	0.958	0.94	8
values#9	68	3	27	2	0.971	0.958	0.965	0.95	9

Nota. Máquinas de vectores de soporte (Support vector machine). nos muestra el resultado de las 10 iteraciones o 10 veces corridas en iguales tiempo. La mejor iteración 4 es 0.97 de exactitud (Accurracy).

- Resultado de métrica de Decision tree



Collected results - 0:28 - Loop End (deprecated) (Recoger resultados)

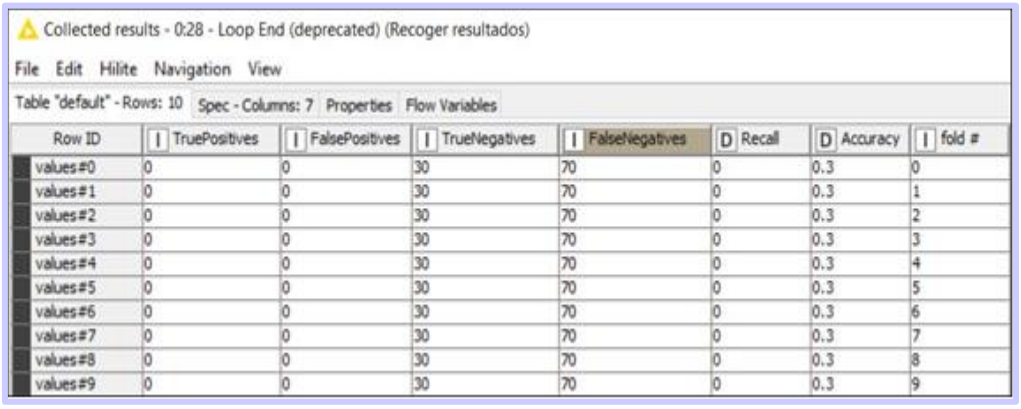
File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 9 Properties Flow Variables

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	F-measure	Accuracy	fold #
values#0	67	5	25	3	0.957	0.931	0.944	0.92	0
values#1	66	7	23	4	0.943	0.904	0.923	0.89	1
values#2	70	4	26	0	1	0.946	0.972	0.96	2
values#3	62	2	28	8	0.886	0.969	0.925	0.9	3
values#4	70	2	28	0	1	0.972	0.986	0.98	4
values#5	66	2	28	4	0.943	0.971	0.957	0.94	5
values#6	67	4	26	3	0.957	0.944	0.95	0.93	6
values#7	68	7	23	2	0.971	0.907	0.938	0.91	7
values#8	68	4	26	2	0.971	0.944	0.958	0.94	8
values#9	66	3	27	4	0.943	0.957	0.95	0.93	9

Nota. Árbol de decisión (Decision tree). nos muestra el resultado de las 10 iteraciones o 10 veces corridas en iguales tiempo. La mejor iteración 2 es 0.96 de exactitud (Accurracy).

- Resultado de métrica de Naive Bayes



Collected results - 0:28 - Loop End (deprecated) (Recoger resultados)

File Edit Hilite Navigation View

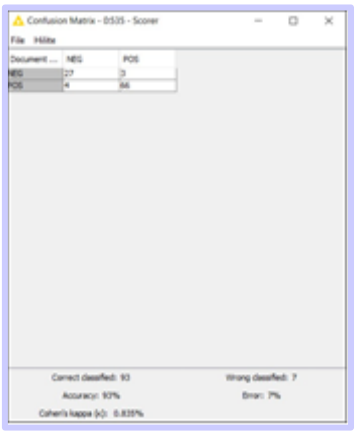
Table "default" - Rows: 10 Spec - Columns: 7 Properties Flow Variables

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Accuracy	fold #
values#0	0	0	30	70	0	0.3	0
values#1	0	0	30	70	0	0.3	1
values#2	0	0	30	70	0	0.3	2
values#3	0	0	30	70	0	0.3	3
values#4	0	0	30	70	0	0.3	4
values#5	0	0	30	70	0	0.3	5
values#6	0	0	30	70	0	0.3	6
values#7	0	0	30	70	0	0.3	7
values#8	0	0	30	70	0	0.3	8
values#9	0	0	30	70	0	0.3	9

Nota. Bayesiano Ingenuo (Naive Bayes). nos muestra el resultado de las 10 iteraciones o 10 veces corridas en iguales tiempo. Todas las iteraciones son iguales el 0.3 de exactitud (Accurracy).

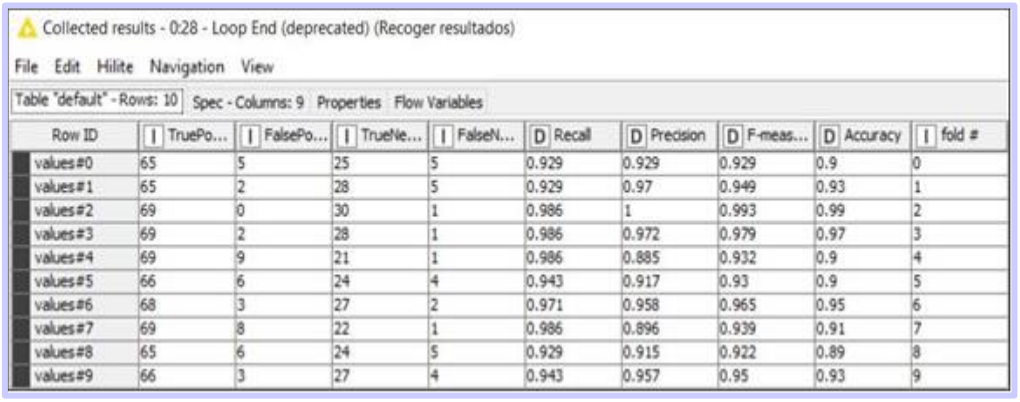
Prueba de Máquinas de vectores de soporte (Support vector machine).

- Resultado Kernel polynomial(polinomial)



Nota. Nos muestra el 93% de exactitud (Accuracy) esto indica que es muy bueno el rendimiento del Algoritmo.

- Resultado de las 10 iteraciones


Collected results - 0:28 - Loop End (deprecated) (Recoger resultados)

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 9 Properties Flow Variables

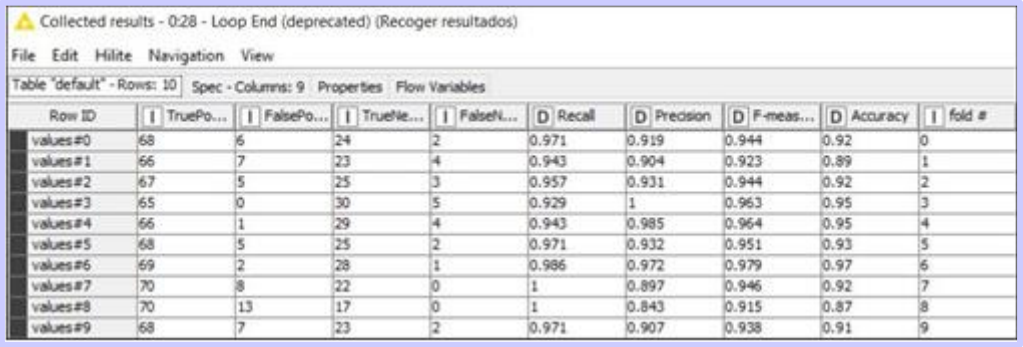
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	F-meas...	Accuracy	fold #
values#0	65	5	25	5	0.929	0.929	0.929	0.9	0
values#1	65	2	28	5	0.929	0.97	0.949	0.93	1
values#2	69	0	30	1	0.986	1	0.993	0.99	2
values#3	69	2	28	1	0.986	0.972	0.979	0.97	3
values#4	69	9	21	1	0.986	0.885	0.932	0.9	4
values#5	66	6	24	4	0.943	0.917	0.93	0.9	5
values#6	68	3	27	2	0.971	0.958	0.965	0.95	6
values#7	69	8	22	1	0.986	0.896	0.939	0.91	7
values#8	65	6	24	5	0.929	0.915	0.922	0.89	8
values#9	66	3	27	4	0.943	0.957	0.95	0.93	9

Nota. Nos muestra el resultado de las 10 iteraciones o 10 veces corridas. La mejor iteración 2 es 0.99 de exactitud (Accurracy).

- Resultado Kernel HyperTangent

Nota. Nos muestra el 91% de exactitud (Accuracy) esto indica que es bueno el rendimiento del algoritmo.

- Resultado de las 10 iteraciones


Collected results - 0:28 - Loop End (deprecated) (Recoger resultados)

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 9 Properties Flow Variables

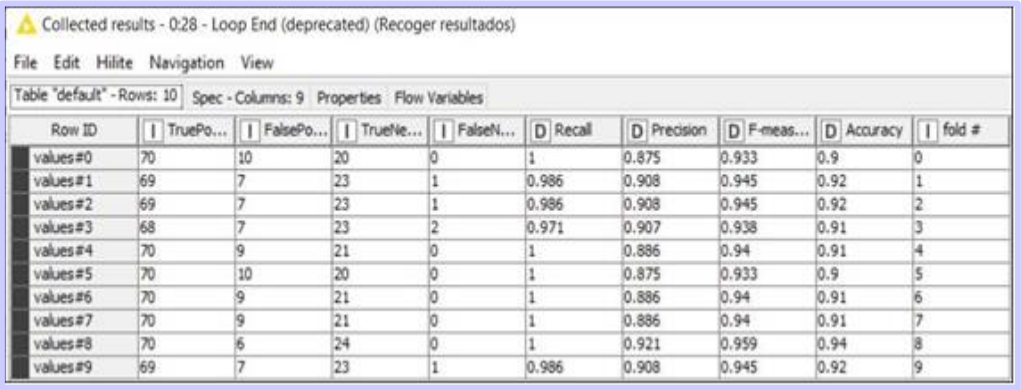
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	F-meas...	Accuracy	fold #
values#0	68	6	24	2	0.971	0.919	0.944	0.92	0
values#1	66	7	23	4	0.943	0.904	0.923	0.89	1
values#2	67	5	25	3	0.957	0.931	0.944	0.92	2
values#3	65	0	30	5	0.929	1	0.963	0.95	3
values#4	66	1	29	4	0.943	0.985	0.964	0.95	4
values#5	68	5	25	2	0.971	0.932	0.951	0.93	5
values#6	69	2	28	1	0.986	0.972	0.979	0.97	6
values#7	70	8	22	0	1	0.897	0.946	0.92	7
values#8	70	13	17	0	1	0.843	0.915	0.87	8
values#9	68	7	23	2	0.971	0.907	0.938	0.91	9

Nota. Nos muestra el resultado de las 10 iteraciones o 10 veces corridas. La mejor iteración 6 es 0.97 de exactitud (Accurracy).

- Resultado de Kernel RBF

Nota. Nos muestra el 92% de exactitud (Accuracy) esto indica que es bueno el rendimiento del algoritmo.

- Resultado de las 10 iteraciones


Collected results - 0:28 - Loop End (deprecated) (Recoger resultados)

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 9 Properties Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	F-meas...	Accuracy	fold #
values#0	70	10	20	0	1	0.875	0.933	0.9	0
values#1	69	7	23	1	0.986	0.908	0.945	0.92	1
values#2	69	7	23	1	0.986	0.908	0.945	0.92	2
values#3	68	7	23	2	0.971	0.907	0.938	0.91	3
values#4	70	9	21	0	1	0.886	0.94	0.91	4
values#5	70	10	20	0	1	0.875	0.933	0.9	5
values#6	70	9	21	0	1	0.886	0.94	0.91	6
values#7	70	9	21	0	1	0.886	0.94	0.91	7
values#8	70	6	24	0	1	0.921	0.959	0.94	8
values#9	69	7	23	1	0.986	0.908	0.945	0.92	9

Nota. Nos muestra el resultado de las 10 iteraciones o 10 veces corridas. La mejor iteración 8 es 0.94 de exactitud (Accurracy).