UNIVERSIDAD DEL AZUAY

FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN

ESCUELA DE INGENIERÍA DE SISTEMAS Y TELEMÁTICA

EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE TEXTOS EN LOS MENSAJES DE LA RED SOCIAL TWITTER.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS Y TELEMÁTICA

AUTOR: IVÁN SEBASTIÁN PAUTE CÁRDENAS

DIRECTOR: ING MARCOS PATRICIO ORELLANA CORDERO

CUENCA, ECUADOR

2018

Dedicatoria

Este trabajo está dedicado a mis padres Iván y Dolores quienes, con su apoyo incondicional, me brindaron el soporte para lograr crecer personal y académicamente.

A mi hermano Joaquín, por sus incentivos diarios.

A mi abuelita Olga que desde el cielo me cuida.

Agradecimientos

Esta tesis agradezco primero a Dios, por darme la calma, y el conocimiento para culminar este proyecto y todas las gracias que me ha brindado a lo largo de estos años

Quiero agradecer infinitamente a mis padres quienes siempre me apoyaron, durante todos estos años de universidad, con sus consejos, su paciencia, y me enseñaron los mejores valores de vida.

A mi familia que me ha brindado toda su ayuda, principalmente a mi tía Katty y mi prima Maoly, por su ayuda sin importar las horas.

A mi director de tesis Ing. Marcos Orellana, que por sus conocimientos y tiempo me ayudaron a desarrollar esta tesis.

A la Ing. Belén Arias, por la ayuda, apoyo y conocimientos compartidos para el desarrollo de este proyecto.

Índice de contenidos

Dedicatoria ii

Agradecimientos iii

Índice de contenidos iv

Índice de tablas x

Índice de figuras xii

Índice de ecuaciones xiii

Resumen xiv

Abstract xv

CAPÍTULO I: INTRODUCCIÓN 16

1.1 Objetivos 16
1.1.1 Objetivo general 16
1.1.2 Objetivos específicos 16
1.2 Justificación 17
1.3 Alcance y resultados esperados 17
CAPÍTULO II: ESTADO DEL ARTE 19

Introducción 19
2.1 Red social Twitter 19
2.1.1 Características de Twitter 20
2.1.2 Hechos históricos 20
2.1.3 Estadísticas financieras 21
2.1.4 Estadísticas de usuarios 21
2.2 Minería de texto 21
2.2.1 Fases de la minería de texto 23
2.2.1.1 Tareas de pre-procesamiento 24
2.2.1.2 Operaciones de minería 24
2.2.1.3 Presentación y navegación 24
2.2.1.4 Técnicas de refinamiento 24
2.2.1.5 Técnicas de pre-procesamiento y de refinamiento 25
2.3 Software de minería de texto 25

2.4 Análisis de sentimiento 25
2.5 Cuadrante mágico de Gartner 26
2.5.1 Escenarios de casos de uso 27
2.5.1.1 Refinamiento de la producción 27
2.5.1.2 Exploración comercial 27
2.5.1.3 Prototipos avanzados 27
2.5.2 Puntos críticos de caso de uso 27
2.5.2.1 Acceso a los datos 28
2.5.2.2 Preparación de datos 28
2.5.2.3 Exploración y visualización de datos 28
2.5.2.4 Automatización 28
2.5.2.5 Interfaz de usuario 28
2.5.2.6 Aprendizaje automático 29
2.5.2.7 Otros análisis avanzados 29
2.5.2.8 Flexibilidad, extensibilidad y apertura 29
2.5.2.9 Rendimiento y escalabilidad 29
2.5.2.10 Entrega 29
2.5.2.11 Plataforma y gestión de proyectos 30
2.5.2.12 Gestión de modelos 30
2.5.2.13 Soluciones pre canalizadas 30
2.5.2.14 Colaboración 30
2.5.2.15 Coherencia 31
2.5.3 Métricas 32
2.5.3.1 Integridad de la visión 32
2.5.3.2 Habilidad para hacer 33
2.5.4 Descripción de cuadrantes 34
2.5.4.1 Leaders (líderes) 34
2.5.4.2 Challengers (Desafiantes) 35
2.5.4.3. Visionaries (Visionarios) 35
2.5.4.4 Niche Players (Proveedores especializados) 36
2.6 Forrester Wave 36
2.7 Herramientas de minería de texto 37
2.7.1 Alteryx 38
2.7.1.1 Descripción 38

2.7.1.2 Licenciamiento 38
2.7.1.3 Soporte de minería con Twitter 39
2.7.2 SAS 39
2.7.2.1 Descripción 39
2.7.2.2 Licenciamiento 39
2.7.2.3 Soporte de minería con Twitter 39
2.7.3 SAP 40
2.7.3.1 Descripción 40
2.7.3.2 Licenciamiento 40
2.7.3.3 Soporte de minería con Twitter 41
2.7.4 Knime 41
2.7.4.1 Descripción 41
2.7.4.2 Licenciamiento 41
2.7.4.3 Soporte de minería con Twitter 42
2.7.5 Rapidminer 42
2.7.5.1 Descripción 42
2.7.5.2 Licenciamiento 42
2.7.5.3 Soporte de minería con Twitter 43
2.8Algoritmos 43
2.8.1 Support Vector Machine 43
2.8.2 k-nearest neighbor classification 44
2.8.2.1 Problemas 45
2.8.3 Naive Bayes 45
2.9 Evaluación 46
2.9.1 Modelos de calidad 46
2.9.1.1 Modelo fijo de McCall 46
2.9.1.1.1 Revisión de producto 47
2.9.1.1.2 Transición del producto 47
2.9.1.1.3 Operación del producto 47
2.9.1.2 Modelo de Boehm 47
2.9.1.2.1 Características de alto nivel 48
2.9.1.2.2 Características de Nivel intermedio 48
2.9.1.2.3 Características primitivas 48
2.9.2 Modelo de Evaluación de calidad de software 48

2.9.2.1 Pesos de Evaluación de calidad....49
2.9.3 Modelo de Evaluación de funcionabilidad de software....51
2.9.3.1 Pesos de evaluación de funcionabilidad....51
2.9.4 Modelo de evaluación de métricas estadísticas, para determinar la exactitud del modelo....52
2.9.4.1 Índices de concordancia de Kappa....53
2.10 Calificación....54
2.10.1 Metodología de calificación de herramientas de minería de texto....54
2.11 Conclusión....56
CAPÍTULO III: RECOPILACIÓN Y GENERACIÓN DE DATOS....57
3.1. Introducción....57
3.2. Preparación de datos....60
3.3. Modelado....62
3.4 Limpieza de datos....62
3.5 Kernel....63
3.6 Proceso general....64
3.7 Conclusión....65
CAPÍTULO IV: EXPERIMENTACIÓN....66
4.1 Experimentación....66
4.2 Criterios de evaluación....66
4.3 Ejecución de pruebas....66
4.3.1 Rapidminer....66
4.3.1.1 Modelado....66
4.3.1.2 Resultados....68
4.3.1.2.1 Matriz de confusión....69
4.3.1.3 Gráfico de resultados de Rapidminer....70
4.3.1.4 Modelo de Evaluación de calidad de software de Rapidminer....71
4.3.1.5 Modelo de Evaluación de funcionabilidad de software de Rapidminer 71
4.3.2 Knime....72
4.3.2.1 Modelado....72
4.3.2.2 Resultados....74
4.3.2.2.1Matriz de confusión....74
4.3.2.3 Gráfico de resultados de Knime....75
4.3.2.4 Modelo de Evaluación de calidad de software de Knime....76

4.3.2.5 Modelo de Evaluación de funcionabilidad de software de Knime 76
4.3.3 Alteryx. 77
4.3.3.1 Modelado 77
4.3.3.2 Resultados 78
4.3.3.2.1Matriz de confusión 78
4.3.3.3 Gráfico de resultados de Alteryx 79
4.3.3.4 Modelo de Evaluación de calidad de software 80
4.3.3.5 Modelo de Evaluación de funcionabilidad de software 81
4.3.4 SAP 81
4.3.4.1 Modelado 81
4.3.4.2 Resultados 82
4.3.4.3 Modelo de Evaluación de calidad de software de SAP 82
4.3.4.4 Modelo de Evaluación de funcionabilidad de software de SAP 83
4.3.5 SAS 84
4.3.5.1 Modelado 84
4.3.5.1.1 Administrar Datos 85
4.2.5.1.2Preparar datos 86
4.2.5.1.3 Explorar y visualizar datos 88
4.2.5.1.4 Construir Modelos 88
4.3.5.2 Resultados 89
4.3.5.2.1 Matriz de confusión de SAS 89
4.3.5.3 Gráfico de resultados de SAS 90
4.3.5.4 Modelo de Evaluación de calidad de software de SAS 91
4.3.5.5 Modelo de Evaluación de funcionabilidad de software de SAS 92
4.4 Conclusión 93
CAPÍTULO V: ANÁLISIS DE RESULTADOS 94
5.1 Evaluación de calidad de software 94
5.2 Evaluación de la funcionabilidad de software 95
5.3 Evaluación de métricas estadísticas para determinar la exactitud del modelo 95
5.4 Conclusión 96
CAPÍTULO VI: VENTAJAS Y DESVENTAJAS DE HERRAMIENTAS 98
6.1 Tabla general de ventajas 98
6.2 Tabla general de desventajas 98
6.3 Descripción detallada de ventas y desventajas 99

6.3.1 Rapidminer.....99
6.3.2 Knime.....100
6.3.3 Alteryx.....100
6.3.4 SAP.....101
6.3.5 SAS.....101
CAPÍTULO VII: ANÁLISIS DE SENSIBILIDAD CON LA MEJOR HERRAMIENTA 103
CAPÍTULO VIII: CONCLUSIONES.....104
Referencias.....106

Índice de tablas

Tabla 1. Modelo de evaluación de calidad de software 49
Tabla 2. Modelo de evaluación de calidad de software 50
Tabla 3. Modelo de Evaluación de funcionabilidad de software 51
Tabla 4.Modelo de Evaluación de funcionabilidad de software 52
Tabla 5. Modelo de Evaluación de métricas estadísticas para determinar la exactitud del modelo.....52
Tabla 6 Índices de concordancia de Kappa.....53
Tabla 7. Fragmento de normalización de tweets obtenida por LIDI.....61
Tabla 8. Fragmentos de tweets para ser procesados por el modelo predictivo obtenidos por LIDI 61
Tabla 9. Matriz de confusión de Rapidminer.....69
Tabla 10. Resultados Recall, Precision, F-Measure, Kappa de Rapidminer.....70
Tabla 11. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Rapidminer 71
Tabla 12. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Rapidminer 72
Tabla 13. Matriz de confusión de Knime.....74
Tabla 14. Resultados Recall, Precision, F-Measure, Kappa de Knime.....75
Tabla 15. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de knime 76
Tabla 16. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de knime 77
Tabla 17. Matriz de confusión de Alteryx 78
Tabla 18. Resultados Recall, Precision, F-Measure, Kappa de Alteryx 79
Tabla 19. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Alteryx.....80
Tabla 20. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Alteryx.....81
Tabla 21. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAP.....83

Tabla 22. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAP 84
Tabla 23. Matriz de confusión de SAS 89
Tabla 24, Resultados Recall, Precision, F-Measure, Kappa de SAS 90
Tabla 25. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAS 92
Tabla 26. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAS 93
Tabla 27. Tabla de puntuación en las herramientas en calidad de software 94
Tabla 28. Puntuación en las herramientas en funcionabilidad de software 95
Tabla 29. Tabla de métricas estadísticas para determinar la exactitud del modelo 95
Tabla 30. Ventajas de herramientas 98
Tabla 31. Desventajas de herramientas 99

Índice de figuras

Figura 1. Cuadrante Mágico de Gartner 32
Figura 2. Forrester Wave ™: Predictive Analytics y Machine Learning Solutions 37
Figura 3. Captura de pantalla de página de desarrolladores de Twitter 57
Figura 4. Captura de pantalla de campos para nueva aplicación 58
Figura 5. Captura de pantalla de los datos de conexión 59
Figura 6. Captura de pantalla del software R 60
Figura 7. Proceso general 64
Figura 8. Modelo de Rapidminer 67
Figura 9: Pre procesamiento de datos Rapidminer 68
Figura 10. Cross Validation Rapidminer 68
Figura 11. Resultados de predicción de Rapidminer 69
Figura 12. Representación gráfica de los datos en Rapidminer 70
Figura 13. Modelado de Knime 73
Figura 14. Modelado de Knime pre procesamiento 74
Figura 15. Representación gráfica de los datos en Knime 75
Figura 16. Modelado de Alteryx 78
Figura 17. Representación gráfica de los datos en Alteryx 79
Figura 18. Consola SAP Hana 82
Figura 19. Menú principal de SAS Viya 85
Figura 20. Administrar datos de SAS Viya 86
Figura 21. Convertir en minúsculas de SAS Viya 86
Figura 22. Stopword de SAS Viya 87
Figura 23. Error de desbordamiento de SAS Viya 87
Figura 24. Configuración del algoritmo SVM de SAS Viya 88
Figura 25. Modelo de solución de SAS Viya 89
Figura 26. Gráfico de resultados de SAS Viya 91
Figura 27. Sentimiento con respecto a la congestión vehicular 103

Índice de ecuaciones

Ecuación 1. Kappa 53
Ecuación 2. Observed accuracy 53
Ecuación 3. Expected accuracy 53
Ecuación 4. Evaluación de puntuación 55
Ecuación 5. Kernel polinómico 63
Ecuación 6. Kernel polinómico 63
Ecuación 7. Kernel polinómico 63
Ecuación 8. Kernel Laplace 63
Ecuación 9. Kernel Laplace 64
Ecuación 10. Kernel Laplace 64

Resumen

El constante crecimiento de las ciudades ha generado el aumento de su parque automotor, lo que ha traído una mayor circulación de vehículos, por consiguiente molestos atascos en las calles. El proyecto busca comparar diferentes herramientas de minería de textos, con el fin de encontrar la mejor opción y mostrarlas en un cuadro comparativo; de esta manera se pretende seleccionar la herramienta adecuada para ser aplicada en un análisis de sensibilidad, acerca del tráfico que se vive en grandes ciudades. Se experimentará con los datos generados por la red social Twitter, catalogados como datos no estructurados. A continuación, se procesarán los datos a través de la extracción, almacenamiento y depuración de los elementos del texto. Se realizarán, a su vez, pruebas de algoritmos en herramientas de minería de texto, sustentadas en los datos sobre el tráfico, previamente obtenidos, así como una selección según los resultados obtenidos.

ABSTRACT

The constant growth of the cities generated the increase of its automotive fleet. This caused a greater circulation of vehicles and annoying traffic jams in the streets. This study sought to compare different text mining tools and show them in a comparative table to find the best option. It was intended to select the appropriate tool to use in a sensitivity traffic analysis of large cities. The data generated by the social network Twitter were cataloged as unstructured data and served for experimentation. The data was processed through the extraction, storage and debugging of text elements. Algorithm tests were performed on text mining tools that were supported in the traffic data previously obtained. Finally, a selection was made.

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by

Ing. Paul Arpi

CAPÍTULO I: INTRODUCCIÓN

El constante crecimiento de las grandes ciudades produce un aumento en la cantidad de vehículos que circulan por sus calles, lo que genera altos índices de tráfico y embotellamientos; ante esta problemática los conductores expresan su descontento a través de la red social Twitter, lo que la convierte en una gran fuente de información, la misma que podría ser explotada de modo que permita encontrar patrones sobre comportamientos, antecedentes, situación actual y problemáticas alrededor de los atascos vehiculares, información que podría ser procesada por distintas herramientas del mercado.

En la siguiente investigación se experimentará con el algoritmo *Support Vector Machine* (SVM) y con una base de datos depurada y basada en tweets relacionados al tránsito de la ciudad de Cuenca. La experimentación será realizada empleando diferentes herramientas de minería de texto, esto con la finalidad de encontrar la mejor herramienta para procesar datos y análisis de sentimiento.

1.1 Objetivos

1.1.1 Objetivo general

Desarrollar un modelo de calidad a partir de las distintas herramientas que existen en el mercado para minería de texto y recomendar la idónea para el análisis de sensibilidad sobre la congestión vehicular en las grandes ciudades.

1.1.2 Objetivos específicos

1. Realizar una investigación bibliografía sobre las diferentes herramientas de minería de texto.

2. Probar las herramientas en un entorno real sobre la red social Twitter y encontrar las mejores para la investigación.

3. Realizar un análisis comparativo de las herramientas de minería de texto a fin de seleccionar la mejor alternativa y generar un cuadro analítico de herramientas que muestre sus ventajas y desventajas.

4. Recomendar la herramienta idónea para el análisis de sensibilidad sobre la congestión vehicular.

1.2 Justificación

Este análisis surge de la gran cantidad de datos generados en la red social Twitter, y de la gran variedad de herramientas para minar textos, de ahí que es necesario encontrar una que procese dicha información de manera ágil y que presente resultados significativos sobre lo que se sondee en la red social, de modo que a partir de los resultados obtenidos se pueda realizar un análisis de sentimientos confiable y eficiente.

Con los datos obtenidos sobre el tráfico y el sentimiento de los usuarios de las vías, se contribuirá en el futuro a una investigación más exhaustiva acerca de la movilidad, la que aplicando software adecuado podría dar a conocer de forma automática el tráfico dentro de una ciudad.

1.3 Alcance y resultados esperados

- Se busca establecer conceptos claros y precisos sobre minería de texto, al tiempo que se pretende indagar sobre las distintas herramientas de minería, obteniendo las más adecuadas para la investigación propuesta y tomando en cuenta aspectos como: la facilidad de uso, el posicionamiento en el mercado, la veracidad de los resultados y las licencias de uso.

- Obtener un procedimiento claro sobre las herramientas de minería de texto seleccionadas, realizando pruebas con un mismo algoritmo, para luego ser comparadas y, a partir de ello, determinar la mejor opción en minería de texto.
- Documentar los resultados de análisis de sensibilidad sobre una base de datos previamente extraída de la red social Twitter, estos análisis se realizarán sobre las herramientas seleccionadas para la investigación.
- Se elaborará, al término de esta investigación, un modelo de calidad, que puntuará a cada herramienta escogida, para la futura aplicación de la misma sobre datos de Twitter en tiempo real.

CAPÍTULO II: ESTADO DEL ARTE

Introducción

En el presente capítulo se ampliará el concepto sobre la red social Twitter, minería de texto y software para minar texto. Consecutivamente, se explicarán los conceptos y la utilidad de los análisis de sentimiento. A continuación, se detallarán las principales herramientas del mercado. Posteriormente, se explicará el método en que las herramientas van a ser evaluadas. Finalmente, se puntualizarán las herramientas de minería de texto seleccionadas y el por qué son las más adecuadas para la investigación.

2.1 Red social Twitter

En el medio actual, las redes sociales son significativamente importantes para la investigación social computacional que busca patrones; investiga preguntas o quiere saber el comportamiento de la humanidad (Batrinca & Treleaven, 2014). Actualmente Twitter cuenta con más de 320 millones de usuarios al mes, produciendo una significativa cantidad de datos, de los que se puede explotar y extraer ideas, pensamientos y sentimientos, los mismos son aplicables a una gran área de estudio (Byrd, Mansurov, & Baysal, 2016).

Twitter es una plataforma web donde los subscritores pueden compartir mensajes de 280 caracteres, 240 destinados al texto y 20 reservados para el nombre del usuario, estos mensajes cortos se denominan tweets. En los últimos años la red social se ha popularizado, convirtiéndose en la más popular de *microblogging* y donde los usuarios pueden publicar frases, enlaces, pensamientos e imágenes.

En esta red social los usuarios reportan eventos en vivo, las empresas dan a conocer sus servicios y productos, los artistas promueven sus actividades y nuevos lanzamientos. Twitter es usado con varios propósitos; se publican 500 millones de tweets por día, esto origina una gran cantidad de datos, que podrían ser analizados (Bonzanini, 2016).

2.1.1 Características de Twitter

Una de las características más significativas de Twitter es su portabilidad. Estar en la web y en dispositivos móviles, permite a los usuarios de esta red, publicar en cualquier lugar y momento. La facilidad de agregar información a la red hace que sea un repositorio de datos que revela eventos del mundo real. Muchos eventos sucedidos fueron señalados por los usuarios de esta red, al mismo tiempo o antes de que los medios tradicionales de comunicación como la televisión o la radio pudieran darlos a conocer (Parikh, 2013). Otra característica de Twitter, que marca una diferencia frente a otras redes sociales, es el tamaño limitado de los mensajes, con un máximo de 140 caracteres, hasta noviembre de 2017, en la actualidad la red social acepta mensajes de 280 caracteres, lo que obliga a los usuarios a expresarse en palabras o frases clave; volviendo a esta información más significativa para su posterior tratamiento. Otra de las principales ventajas de Twitter es que lo utilizan todas los estratos sociales, ampliando la visión de lo que está pasando en todo el mundo (De Groot, 2012).

2.1.2 Hechos históricos

Jack Dorsey, Evan Williams, Biz Stone y Noah Glass crearon Twitter en marzo de 2006; su lanzamiento oficial fue en julio del mismo año.

El 21 de marzo de 2006 se publicó el primer tweet, este fue subido por Jack Dorsey. En el año 2007 se presentó el primer *#hashtag*, el cual fue una propuesta de un suscriptor de la red social (Smith, 2016).

2.1.3 Estadísticas financieras

Durante los primeros tres meses del año 2016 Twitter produjo $595 millones de dólares, la mayor parte de esta suma fue generada por la venta de publicidad. Durante el 2015 pasó a ser una compañía rentable reportando un ingreso neto de $7 millones de dólares en los últimos meses del 2015. La empresa de Jack Dorsey, tuvo una nómina de personal de 3.900 empleados hasta el 2016 (Smith, 2016).

2.1.4 Estadísticas de usuarios

Hoy en día existen 310 millones de usuarios de Twitter activos al mes, la red social consta de 1,3 mil millones de cuentas creadas, pero solo el 66% de estas han registrado enviar un tweet. El 80% de usuarios activos accede desde un teléfono móvil. Estados Unidos tiene el 29,2% de los usuarios de Twitter. El 24,6% de cuentas activas y verificadas pertenece a periodistas, así como el 83% de líderes mundiales tiene una cuenta activa en la plataforma. La empresa estima que, del total de sus cuentas, 23 millones de usuarios son *bots* (Smith, 2016).

En un solo día se suben a Twitter 500 millones de tweets, esto demuestra que 6.000 mensajes son enviados por segundo; como resultado, la cantidad de información generada en esta plataforma es muy extensa (Smith, 2016).

2.2 Minería de texto

La minería de texto o (*text mining*) se engloba dentro de las técnicas y modelos de minería de datos; por esta razón, se debe revisar y entender previamente qué es y en qué consiste la minería de datos.

En la actualidad la automatización de sistemas, ha hecho que la información sea digitalizada; siendo fácil de: capturar, almacenar, compartir, trasmitir y procesar. Hoy en día el avance de gestión de base de datos, no se restringe solo a números y texto, esto hace

posible almacenar todo tipo de datos, generando repositorios que pueden ser de tipo: imagen, texto, video y numéricos. La minería de datos busca encontrar información útil, entre grandes repositorios de datos sin procesar (José C. Riquelme, 2015).

La minería de datos se puede definir como: el análisis matemático para deducir tendencias y patrones de comportamiento que se encuentran en los datos, donde dichos comportamientos no se pueden detectar con una exploración normal de los datos, ya que la información se encuentra en grandes volúmenes de datos. Al recopilar estos comportamientos y patrones de datos, se puede definir un modelo de minería de datos. Deduciendo que Data Mining se llama al conjunto de métodos estadísticos que proporcionan información relevante, cuando se tiene un gran número de datos (Rochina, 2017).

Debido al incremento en el uso de redes sociales, *Text mining* se ha convertido en una tecnología ascendente, la cual pretende extraer información valiosa de datos textuales no estructurados (He, Zha, & Li, 2013). Según (He, Zha, & Li, 2013), alrededor del 80% de la información de una institución, está comprometida en documentos de texto; tales como: informes, órdenes de pedido de los clientes, facturas, correos electrónicos y memorandos. Según el informe de (Gandomi & Haider, 2015), los análisis de texto permiten a las empresas convertir grandes cantidades de textos generado por sus empleados, clientes y competencia, en resúmenes significativos, los cuales respaldan la toma de decisiones con un fundamento en evidencias.

Para obtener información valiosa, de un gran banco de información textual con eficacia, es indispensable el uso de métodos informáticos automatizados (He, Zha, & Li, 2013); siendo el objetivo principal de la minería de texto hallar: tendencias, guías, patrones ocultos, normas o modelos provechosos; todo esto extraído de datos textuales no estructurados, como por ejemplo: correos, redes sociales, correos electrónicos y archivos web (He, Zha, & Li, 2013).

La minería de texto también se define como un grupo de normas que, por medio de datos específicos, relacionados a cierto tema de investigación y proporcionados por usuarios, produce información relevante, obteniendo tendencias más citadas y permitiendo analizar y agrupar a los usuarios que aportaron la información.

La minería de textos aplicada a redes sociales puede entregar interesantes resultados que van desde el comportamiento humano hasta la interacción entre las personas (He, Zha, & Li, 2013).

A la extracción, análisis y representación de información proveniente de la interacción, entre los usuarios en redes sociales, se le denomina minería de medios sociales. Varias técnicas de minería de datos son aplicadas en la minería de texto, siendo esta última una prolongación de la primera técnica. Minar texto, en conclusión, se fundamenta en extraer información relevante de textos no estructurados, sin importar el tamaño de estos, con el fin de obtener modelos que proporcionen información relevante y posterior conocimiento (Arias, 2016).

2.2.1 Fases de la minería de texto

En el análisis que se realizará es importante contar con datos estandarizados para garantizar la imparcialidad de los mismos al momento de evaluar las herramientas de minería de texto.

La minería de texto consta de 5 fases principales: Tareas de pre-procesamiento, Operaciones de minería, Presentación y navegación, Técnicas de refinamiento, Técnicas de pre-procesamiento y de refinamiento; las cuales se detallan a continuación, estas garantizan la obtención de datos destacados y útiles.

2.2.1.1 Tareas de pre-procesamiento

Antes de procesar la información y extraerla es necesario preparar la misma, otorgando a los datos un formato único. Para facilitar el análisis de la información en esta fase de clasificación y pre-procesamiento, se incorporan distintos métodos, prácticas y procesos para acondicionar los datos no estructurados o datos sin procesar, antes que estos sean minados (Arias, 2016).

2.2.1.2 Operaciones de minería

La operación de minería es la fase más importante de *text mining*. Las distintas operaciones aplicadas sobre los datos nos dan como resultado: algoritmos de extracción de conocimiento, análisis del comportamiento de tendencias y descubrimiento de patrones, siendo los más empleados: comparaciones entre niveles de inclinación, asociación y conceptos del cercano más reiterado; de igual forma, realiza patrones de distribución y proporciones (Arias, 2016).

2.2.1.3 Presentación y navegación

En esta fase se incorporan herramientas de navegación y visualización, las cuales contribuyen al entendimiento de los datos; incluyendo caracteres o gráficos para generar agrupaciones conceptuales, patrones, perfiles particulares en conceptos o modelos (Arias, 2016).

2.2.1.4 Técnicas de refinamiento

Las técnicas de refinamiento pertenecen a la etapa de pre-procesamiento; en esta fase la información es filtrada mediante procedimientos de supresión, poda, ordenamiento y agrupación. Así también, se agrupan, ordenan, resumen y generalizan datos redundantes, con el fin de optimizar la información obtenida (Arias, 2016).

2.2.1.5 Técnicas de pre-procesamiento y de refinamiento

Esta última fase es crucial en la minería de texto, en este punto los datos son organizados, para sobre estos, aplicar algoritmos; y de la aplicación de los resultados de estos algoritmos obtener información, la cual finalmente será depurada (Arias, 2016).

2.3 Software de minería de texto

En la actualidad existe una extensa variedad de herramientas para realizar minería de textos, como: Alteryx, Rapidminer, Knime, SAS, SAP, Leximancer, Nvivo 9, SPSS Modeler (anteriormente Clementine). Estas herramientas usan complejos modelos de programación como: árboles de decisión, agrupación, programación lógica y algoritmos estadísticos para encontrar patrones en datos textuales no estructurados (He, Zha, & Li, 2013).

2.4 Análisis de sentimiento

Según (He, Wu, Yan, Akula, & Shen, 2015), el análisis del sentimiento se refiere a la extracción automática de opiniones positivas o negativas de los textos, ya que normalmente los textos contienen una mezcla de sentimientos positivos, negativos o neutros. El análisis de sentimientos se utiliza para determinar la actitud de los clientes y usuarios en temas específicos, tales como: revisiones de productos (por ejemplo, electrónica, software, ropa), revisiones de servicios, finanzas y el estado de ánimo en general de la población.

El análisis de sentimiento ofrece la opinión de los usuarios de una red social, esta puede variar entre positiva, negativa o neutra, denota el comportamiento humano e influye en la toma de decisiones; por ejemplo, una empresa al realizar un análisis de sentimiento sobre uno de sus productos, y al obtener los resultados del mismo puede analizar cuáles son las falencias de su producto, a qué mercado y target está llegando, para mejorar el producto o emprender una nueva campaña de mercadeo. (Bannister, 2015)

El análisis de sentimiento puede ser realizado en diferentes campos de la investigación, por ejemplo, se puede conocer el nivel de aceptación de un nuevo producto, la popularidad de una canción, la ideología política de una persona, analizar los motivos por lo que ciertos productos son más vendidos que otros (Bannister, 2015).

Los pensamientos de opinión expresados por un usuario se denominan datos no estructurados, adquieren este nombre en razón de no estar guardados en una base de datos tradicional (jerárquica, de red o estructura relacional); los datos pueden provenir de sensores de ciudad, edificaciones inteligentes, dispositivos móviles, redes sociales, entre otros (Bannister, 2015).

El crecimiento de la interacción de las personas en medios sociales, como Twitter, proporciona una fuente de minería, la cual es respaldada por técnicas adecuadas para generar análisis de sentimiento; dará como resultado la opinión de los usuarios sobre un tema en específico, y para este estudio, específicamente, el sentir de las personas sobre el tráfico vehicular (Arias, 2016).

2.5 Cuadrante mágico de Gartner

Cada año el cuadrante mágico de Gartner presenta un gráfico sobre las distintas herramientas para minería de datos, que se utilizan para generar soluciones de aprendizaje automático. Gartner llama a las herramientas plataformas de ciencia de datos, definiendo a estas plataformas como: “Una aplicación de software coherente que ofrece una combinación de bloques de construcción básicos para crear todo tipo de soluciones de ciencia de datos, e incorporar esas soluciones en procesos de negocios, infraestructura y productos circundantes”. La metodología para la investigación de Gartner no evalúa plataformas puras de código abierto como Python y R (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.1 Escenarios de casos de uso

Todas las herramientas del cuadrante mágico de Gartner son evaluadas bajo los siguientes escenarios de casos de uso, que representan el correcto funcionamiento de la plataforma o el más habitual: refinamiento de la producción, exploración comercial y prototipos avanzados, estos escenarios son la secuencia de pasos que la herramienta debe cumplir, los mismos que se detallan a continuación:

2.5.1.1 Refinamiento de la producción

En este escenario se centra el mayor tiempo de trabajo de los equipos de ciencia de datos; se considera la implementación de soluciones para facilitar y agilizar el tiempo de este proceso (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.1.2 Exploración comercial

La exploración comercial se enfoca en el descubrimiento de lo desconocido; analiza la capacidad de preparación, exploración, visualización de datos existentes y nuevas fuentes de información (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.1.3 Prototipos avanzados

Este escenario abarca soluciones novedosas para el aprendizaje automático de información y se basa en la mejora de enfoques tradicionales. Los enfoques tradicionales pueden incluir: minería de datos, soluciones exactas, heurística y el juicio humano (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2 Puntos críticos de caso de uso

La empresa que genera el cuadrante mágico de Gartner, utiliza quince puntos críticos de evaluación en los 3 escenarios antes descritos, para puntuar las herramientas de minería de texto se consideran los siguientes aspectos:

2.5.2.1 Acceso a los datos

Hace referencia a la manera en que admite el programa de minería, el acceso, datos e integración; independientemente de la fuente de que procedan (locales o en la nube), sin importar el tipo, ya sea: textual, transaccional, imagen, audio, series de tiempo, datos de ubicación, entre otros (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.2 Preparación de datos

Revisa si el software tiene una variedad significativa de codificación o no codificación, como: la transformación de datos y el filtrado, para alistar los elementos de evaluación para el modelado (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.3 Exploración y visualización de datos

Verifica que el software permita una variedad de instrucciones de visualización, incluyendo la visualización interactiva (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.4 Automatización

En este aspecto se revisa que el software facilite la automatización en la creación de características y ajustes de parámetros (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.5 Interfaz de usuario

Se evalúa si el software tiene un aspecto coherente y proporciona una interfaz gráfica intuitiva al usuario (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.6 Aprendizaje automático

Analiza qué tan amplios son los aspectos, en cuanto al aprendizaje automático, así como el soporte para enfoques modernos de aprendizaje automático, como: técnicas de conjunto (refuerzo, embolsado y bosques aleatorios) y aprendizaje profundo (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.7 Otros análisis avanzados

Se consideran análisis avanzados a otros métodos que ocupen herramientas como: procesamiento de texto, análisis de imagen, estadística, optimización, simulación, y que estén integrados en el entorno gráfico del software (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.8 Flexibilidad, extensibilidad y apertura

Este caso de uso revisa la forma en que se pueden integrar las bibliotecas de código abierto en el software; cómo los usuarios pueden crear sus propias funciones y analiza el modo en que el programa se comporta y gestiona los recursos de notebooks (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.9 Rendimiento y escalabilidad

Examina la forma en que se emplean las configuraciones multinúcleo y multinodo, también el modo de implementación, ya sea de escritorio, servidor o nube (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.10 Entrega

Inspecciona qué tan adecuadamente soporta la capacidad de crear API o contenedores (como código, *Predictive Model Markup Language* [PMML] y aplicaciones

empaquetadas), que se aplican en escenarios de negociación de una forma más rápida (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.11 Plataforma y gestión de proyectos

Estudia la capacidad de gestión del programa, si éste brinda seguridad, la gestión de recursos e información, la reutilización de versiones de anteriores de proyectos; así como la categoría de auditoría y reproducibilidad (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.12 Gestión de modelos

Revisa qué disposición tiene la plataforma para supervisar y recalibrar cientos o miles de modelos. Se compone de capacidad de prueba de modelos como: validación cruzada, entrenamiento, validación y divisiones de prueba, AUC, ROC, matrices de pérdida y comprobación de modelos, uno al lado del otro (por ejemplo, prueba de campeón / retador [A / B]) (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.13 Soluciones pre canalizadas

Analiza si el software propone soluciones «pre canalizadas» como, por ejemplo: detección de anomalías, predicción de fallas, predicción de compra, sistemas de recomendación, detección de fraude, análisis de redes sociales y soluciones para ventas cruzadas, que pueden adjuntarse por medio de galerías, bibliotecas y mercados (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.14 Colaboración

Valora la forma en que trabajan los usuarios, con diferentes habilidades en los mismos proyectos y flujos de trabajo; también cómo se pueden reutilizar, comentar y archivar los proyectos (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.2.15 Coherencia

Principalmente considera cuán consistente, integrado e intuitivo es el software; para soportar una gran cantidad de datos que serán analizados posteriormente. La plataforma de análisis tiene que soportar metadatos e integración para los 14 casos de uso anteriores. Esta meta capabilidad asegura que los formatos de entrada / salida de datos estén estandarizados, de modo que los componentes tengan un aspecto y tacto consistente y la terminología sea unificada a través de la plataforma (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

En la figura 1, se muestra que el proceso identificó 16 proveedores de software, a los cuales los clasifica en 4 cuadrantes: a) *leaders* donde están las plataformas: Rapidminer, Knime, SAS, IBM; b) *challengers* en este cuadrante están: Alteryx, Quest, Angoss, MathWorks; c) *niche players* que contiene las herramientas: SAP, FICO, Teradata; d) *visionaries* este último cuadrante contiene: Microsoft, H2O.ai, Dataiku, Domino Data Lab y Alpine Data.

Figura 1. Cuadrante Mágico de Gartner *(Linden, Krensky, Hare, Idoine, & Sicular, 2017)*

2.5.3 Métricas

El cuadrante mágico de Gartner toma en cuenta 15 métricas de evaluación, centrándose en dos categorías: integridad de la visón y la habilidad para hacer; de igual forma para calificar estas métricas usa una escala de "alta", "baja" o "media", que en algunos casos, puede tener un peso "sin calificación" esto debido a que tiene poca importancia para el análisis, o no proporciona una diferenciación suficiente. (David Black, 2016)

2.5.3.1 Integridad de la visión

En esta categoría se encuentran estipulados los siguientes criterios:

- **Comprensión del mercado:** la capacidad de un proveedor para comprender las necesidades de los usuarios y traducir estas necesidades en productos y servicios (David Black, 2016).

- **Estrategia de mercadeo:** Un grupo claro y diferenciado de mensajes que se comunican constantemente en toda la organización y se publicitan a través de: presencia en línea, publicidad, programas de clientes, eventos y declaraciones de posicionamiento (David Black, 2016).
- **Estrategia de ventas:** es una habilidad para vender productos o servicios que utilizan la red de ventas apropiada, marketing, servicios y comunicación; para ampliar el alcance del mercado de un proveedor, habilidades, experiencia, tecnologías, servicios y base de clientes (David Black, 2016).
- **Estrategia de oferta (producto):** hace referencia al enfoque de un proveedor para el desarrollo de productos y la prestación de servicios que enfatiza en la diferenciación, las funciones, la metodología y el conjunto de características en relación con los requisitos actuales y futuros (David Black, 2016).
- **Modelo de negocio:** la validez y la lógica de la propuesta comercial de un proveedor en este mercado (David Black, 2016).
- **Estrategia vertical / industrial:** destreza de un proveedor para dirigir recursos, habilidades y ofertas para satisfacer las necesidades de clientes individuales, incluidas las industrias verticales (David Black, 2016).
- **Innovación:** combinación de recursos, experiencia o capital, para obtener ventajas competitivas (David Black, 2016).
- **Estrategia geográfica:** pericia de un proveedor de software para dirigir recursos, habilidades y ofertas para satisfacer las necesidades de las regiones más allá de su mercado habitual (David Black, 2016).

2.5.3.2 Habilidad para hacer

Esta categoría incluye los siguientes aspectos:

- **Productos / servicios:** productos principales y servicios ofrecidos por el proveedor que compiten y sirven a los clientes (David Black, 2016).
- **Viabilidad general:** incluye una evaluación de la situación financiera general del proveedor (David Black, 2016).
- **Ejecución de ventas / fijación de precios:** las capacidades del vendedor en actividades de preventas y ventas. Este criterio también incluye: administración de

acuerdos, fijación de precios y negociación, soporte de preventas y efectividad general del canal de ventas (David Black, 2016).

- **Capacidad de respuesta del mercado y trayectoria:** la destreza del proveedor para responder, cambiar de dirección, ser flexible y lograr el éxito competitivo sobre sus rivales (David Black, 2016).
- **Ejecución de marketing:** La claridad, calidad, creatividad y eficacia de la ejecución de los programas de marketing, diseñados para entregar el mensaje del vendedor para influenciar el mercado, promover su marca y negocio, aumentar la conciencia de sus productos y servicios, y establecer una identificación positiva con el producto (David Black, 2016).
- **Experiencia del cliente:** relaciones, productos, servicios y programas que permiten a los clientes tener éxito con los productos que se evalúan. Este criterio incluye las formas en que los clientes reciben soporte técnico (David Black, 2016).
- **Operaciones:** la capacidad del desarrollador, para cumplir sus objetivos y compromisos, que permiten al usuario operar de manera efectiva y eficiente (David Black, 2016).

2.5.4 Descripción de cuadrantes

2.5.4.1 Leaders (líderes)

Los líderes tienen una repetitiva y significativa aparición en el mercado. Disponen de recursos especializados en sus herramientas. Los líderes indican robustez en profundidad y amplitud, por medio de un completo proceso de desarrollo e implementación del modelo. Cuentan con una gran cantidad de clientes que están afiliados a dichas herramientas por un largo tiempo. Los programas de este cuadrante se adaptan a las condiciones del cambiante mercado (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

Los líderes están bien posicionados para influir en el crecimiento y dirección del mercado. Las herramientas involucradas en esta posición abordan todas las industrias, geografías, casos de uso y dominio de datos. Esto les otorga una gran ventaja sobre la compresión y

estrategia para el mercado de la ciencia de datos, lo que les permite desarrollar ideas innovadoras y líderes, frente a otras herramientas de minería (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.4.2 Challengers (Desafiantes)

Los desafiantes o retadores tienen una comparecencia establecida, viabilidad, credibilidad y sólidas capacidades de producto. No obstante, es probable que no denoten innovación y liderazgo intelectual, como las herramientas en el cuadrante de líderes (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

Existen dos clases de herramientas desafiantes:

1. Las empresas desarrolladoras que están en el mercado desde hace tiempo, las mismas que tienen éxito por su estabilidad, relación a largo plazo con los clientes y previsibilidad. Estos programas requieren cambiar su enfoque y mantenerse al tanto de la evolución del mercado (Linden, Krensky, Hare, Idoine, & Sicular, 2017).
2. Proveedores muy bien establecidos, en mercados similares de la ciencia de datos que, con soluciones moderadas, ingresan al mercado y que son tomados en cuenta por los clientes; debido a que estos desarrolladores son altamente influyentes en el mercado, son propensos a convertirse en líderes (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.4.3. Visionaries (Visionarios)

Los visionarios son los nuevos o pequeños proveedores que influyen en nuevas tendencias que forman, o en un futuro serán parte del mercado. Sin embargo, la fiabilidad de estos proveedores decrece por dudas de si son capaces de seguir ejecutándose eficazmente. Otro factor que impide que estos crezcan es la falta de conocimiento de su existencia, por parte

de los clientes, generando muy poco impulso para estos nuevos programas (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.5.4.4 Niche Players (Proveedores especializados)

Los proveedores de esta sección denotan robustez en una industria en particular, o están asociados a un solo campo de la tecnología. Algunos proveedores especializados demuestran un alto grado de enfoque en otros campos, lo cual los conduce a ser visionarios (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.6 Forrester Wave

Para el primer trimestre de 2017, el analista Mike Gualtieri lanzó su Forrester Wave™, un informe sobre: *Predictive Analytics y Machine Learning Solutions* (Piatetsky, 2017). El documento evalúa y examina 14 empresas, en 3 ejes principales: presencia en el mercado, oferta actual y estrategia; los resultados obtenidos se observan en la figura2.

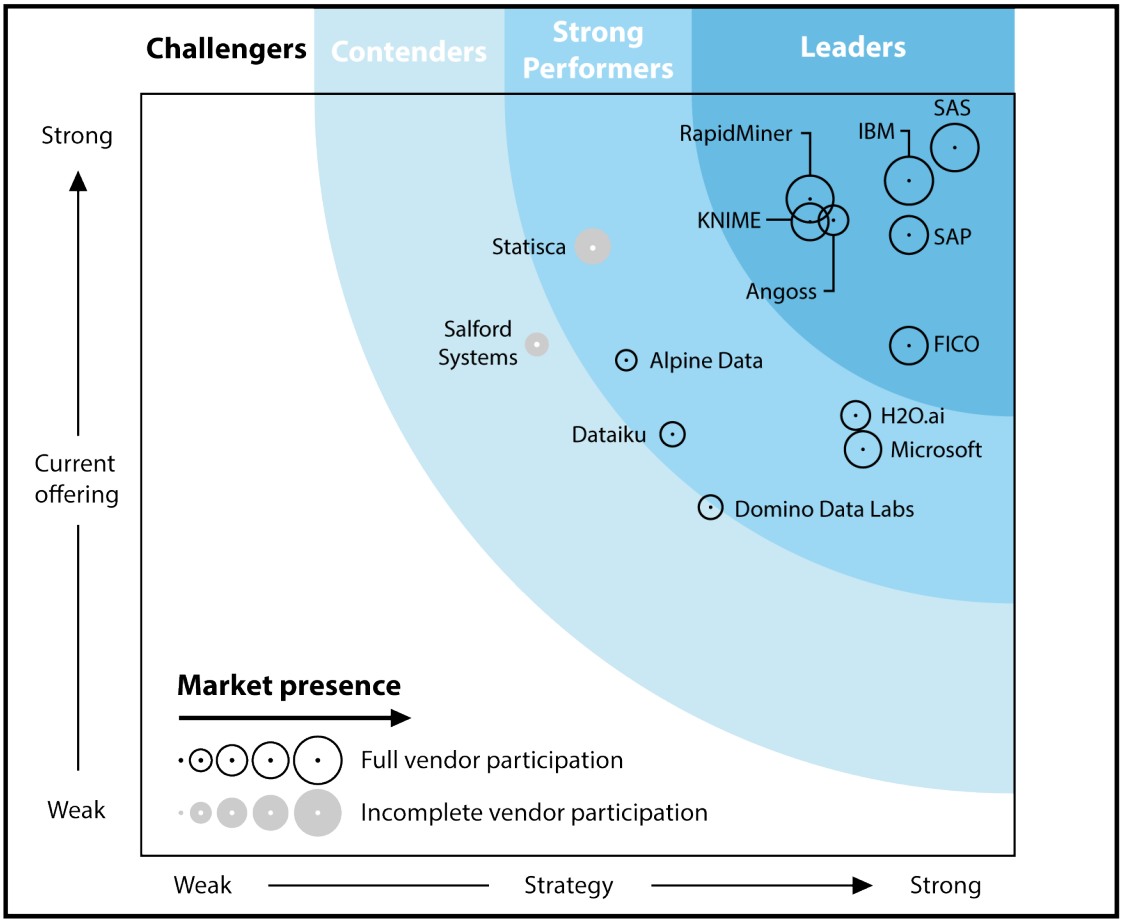

Figura 2. Forrester Wave ™: Predictive Analytics y Machine Learning Solutions *(Piatetsky, 2017)*

2.7 Herramientas de minería de texto

Para el desarrollo de esta investigación, se ha considerado ocupar, los análisis de Forrester Wave y el cuadrante mágico de Gatner, por el prestigio que han adquirido y la cantidad y calidad de información que año a año presentan en análisis de software; por lo tanto las herramientas que se detallan a continuación, se han seleccionado en base a los análisis mencionados (cuadrante mágico de Gartner (2017) y Forrester Wave™ 2017). Las plataformas de minería de texto escogidas para el análisis, corresponden a los principales softwares, dados a conocer por el análisis Forrester Wave: SAS, SAP, Rapidminer y Knime. La razón de escoger estas herramientas, es porque se encuentran de igual forma, como las mejores en el cuadrante de líderes del análisis de Gartner. Se aprecia en la figura 1 y 2, que la plataforma IBM se encuentra también como una de las principales, esta no se analizará por motivos de licenciamiento, no obstante la herramienta Alteryx será

analizada, aunque se encuentra en el cuadrante de Challengers (desafiantes), ya que es la más cercana al cuadrante superior de líderes.

2.7.1 Alteryx

2.7.1.1 Descripción

Alteryx es desarrollado en Irvine, California, EE.UU. Ofrece a sus usuarios la capacidad única de preparar, combinar y analizar fácilmente todos los datos, utilizando un flujo de trabajo repetible, para luego implementar y compartir análisis a escala que permitan obtener conocimientos más profundos en horas y no en semanas (Alteryx, 2017).

El software también oferta la capacidad de unir datos de fuentes internas y externas, para luego procesarlos usando herramientas preceptivas y prescriptivas, empleando para ello la misma interfaz gráfica en un solo flujo de trabajo. Además, tiene integración con R, que permite a los usuarios ampliar las funciones del software, creando y ejecutando líneas de código de la plataforma R. Alteryx también ofrece una galería analítica, basada en la nube para colaboración, intercambio y control de flujos de trabajo (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

La plataforma Alteryx Analytics puede conectarse y limpiar datos del repositorio de datos, aplicaciones en la nube, hojas de cálculo y otras fuentes, unir fácilmente estos datos y luego realizar análisis (predictivos, estadísticos y espaciales), utilizando la misma interfaz de usuario intuitiva (Alteryx, 2017).

2.7.1.2 Licenciamiento

Esta plataforma ofrece a los usuarios tres tipos de licencia:

- Diseñador Alteryx.
- Servidor Alteryx.

- Galería Alteryx Analytics.

Estas licencias son de pago, sin embargo el software ofrece una prueba de 30 días gratis (Smith, 2016).

2.7.1.3 Soporte de minería con Twitter

El software ya no tiene soporte directo con la API de Twitter. Para procesar datos de la red social; es necesario ocupar otra herramienta de extracción como R y pasar estos datos a un archivo de texto *comma-separated values* (csv).

2.7.2 SAS

2.7.2.1 Descripción

SAS procede de Carolina del norte, EE.UU. Oferta una gran gama de soluciones software para el análisis y ciencia de datos (Linden, Krensky, Hare, Idoine, & Sicular, 2017). La plataforma racionaliza todo el proceso de minería, desde el acceso a los datos hasta la generación de un modelo de valoración. Todas las tareas son realizadas a través de una única solución integrada, logrando la máxima flexibilidad y asegurando un eficaz trabajo en equipo (SAS, 2018).

2.7.2.2 Licenciamiento

El software consta de varias licencias de pago, para sus distintos productos, lo cual afecta su capacidad de crecimiento en el mercado; sin embargo, esta herramienta consta de una versión de prueba, la cual permite procesar solo hasta 100 datos (Linden, Krensky, Hare, Idoine, & Sicular, 2017).

2.7.2.3 Soporte de minería con Twitter

SAS *Event Stream Processing* (ESP) no solo puede procesar eventos de transmisión estructurados (una colección de campos) en tiempo real, sino que también posee

características muy avanzadas con respecto a la recopilación y el análisis de eventos no estructurados. Twitter es una de las aplicaciones de redes sociales más conocidas, y probablemente, la primera que se viene a la mente cuando se piensa en una fuente de transmisión de datos. Por otro lado, SAS tiene potentes soluciones para analizar datos no estructurados con SAS Text Analytics. Esta aplicación de SAS recopila datos no estructurados, provenientes de Twitter, y realizar un procesamiento de análisis de texto en tweets (extracción contextual, categorización de contenido y análisis de sentimiento) (SAS, 2018).

2.7.3 SAP

2.7.3.1 Descripción

La plataforma SAP tiene cede en Wakdorf, Alemania. El software consta de una gran variedad de ofertas de análisis (Linden, Krensky, Hare, Idoine, & Sicular, 2017). SAP HANA, transforma la inteligencia analítica. Utiliza el procesamiento de datos avanzado para datos comerciales, de texto, espaciales, gráficos y series en un sistema, para obtener una visión sin precedentes. Brinda, a su vez, conocimientos más profundos, empleando potentes capacidades de aprendizaje automático y análisis predictivo (SAP, 2018).

2.7.3.2 Licenciamiento

La plataforma consta de diferentes tipos de claves de licencia HANA; la base de datos de SAP HANA admite dos tipos de claves de licencia:

1. Clave de licencia temporal: Las claves de licencia temporales, se instalan automáticamente con una nueva instalación de base de datos de SAP HANA. Es válido por 90 días a partir de la fecha de instalación. Durante este período, se debería solicitar en el mercado del servicio y aplicar una clave de licencia permanente.
2. Clave de licencia permanente: Las claves de licencia permanentes son válidas hasta la fecha de vencimiento predefinida. Debe solicitarse en SAP Service

Marketplace en Claves y solicitudes, y aplicarse a la base de datos SAP HANA individual. Además, especifican la cantidad de memoria con licencia para la instalación de destino de SAP HANA (SAP, 2018).

2.7.3.3 Soporte de minería con Twitter

SAP HANA mantiene el soporte de conexión con Twitter, el que se explica a detalle en la página web de la plataforma (SAP, 2018).

2.7.4 Knime

2.7.4.1 Descripción

Konstanz Information Miner (KNIME) tiene su sede en Zurich, Suiza. Es una plataforma completamente funcional y escalable gracias a su código abierto, KNIME Analytics Platform (Linden, Krensky, Hare, Idoine, & Sicular, 2017). KNIME es un entorno totalmente gratuito para el desarrollo y ejecución de técnicas de minería de datos, desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa continúa su desarrollo, al tiempo que presta servicios de formación y consultoría (KMINE, 2018).

KNIME está desarrollado sobre la plataforma Eclipse y programado, esencialmente, en Java. Su uso se basa en el diseño de un flujo de ejecución que plasme las distintas etapas de un proyecto de minería de datos (KMINE, 2018).

2.7.4.2 Licenciamiento

La Licencia Pública General de GNU es una licencia *copyleft* gratuita para software y otros tipos de trabajos.

Las licencias para la mayoría del software y otras obras prácticas están diseñadas para quitarle la libertad de compartir y cambiar las obras. Por el contrario, la Licencia Pública General de GNU, está destinada a garantizar su libertad para compartir y cambiar todas las versiones de un programa, para asegurarse de que siga siendo un software libre para todos sus usuarios (KMINE, 2018).

2.7.4.3 Soporte de minería con Twitter

La plataforma KNIME, explicada a detalle para el soporte de conexión con Twitter, se encuentra en la documentación de la página oficial (KMINE, 2018).

2.7.5 Rapidminer

2.7.5.1 Descripción

Rapidminer ofrece sus datos basados en la interfaz gráfica del usuario. Rapidminer, es una herramienta adecuada para crear modelos y, por consiguiente, para la realización de análisis predictivos de grandes volúmenes de datos (Rapidminer, 2018). Es una solución que facilita el autoservicio de análisis predictivo, permitiendo una avanzada analítica y empleando solamente *drag and drop*; además cuenta con una herramienta para la generación de código. Es utilizada para realizar análisis de minería de datos en aplicaciones empresariales, para el gobierno y para el mundo académico (Rapidminer, 2018).

2.7.5.2 Licenciamiento

Rapidminer tiene el compromiso de proporcionar a sus usuarios un núcleo de código abierto y gratuito; sin embargo, la plataforma también ofrece productos bajo licencias comerciales.

2.7.5.3 Soporte de minería con Twitter

Rapidminer, ofrece extensa documentación certificada por la compañía para la conexión con Twitter disponible en su página oficial donde cuenta con ejemplos y manuales para una correcta y fácil conexión con la red social.

2.8Algoritmos

Los tres mejores algoritmos clasificadores de minería de datos, según (Xindong Wu, 2007) son: *Support vector machines, k-nearest neighbor classification* (KNN), *Naive Bayes*. Estos algoritmos, frecuentemente, son utilizados para construir soluciones de clasificación, dichos sistemas necesitan una entrada de datos de aprendizaje, cada entrada se describe por sus valores, para un conjunto fijo de atributos y generan un clasificador, que permite predecir con precisión la clase a la que pertenece una nueva entrada de datos.

2.8.1 Support Vector Machine

SVM (Support Vector Machine) es un algoritmo que, en términos matemáticos, se puede definir como el intento de encontrar una holgura n-dimensional, que permita dividir los patrones de entrenamiento positivo de los negativos, posibilitando denotar el margen más extenso posible (Xindong Wu, 2007).

El principal objetivo, que pretende este algoritmo, es hallar el hiperplano óptimo que magnifique la distancia entre los polos, positivo y negativo (Ortíz & Martín, 2015).

Entre los algoritmos conocidos, SVM ofrece uno de los métodos más sólidos y precisos, además cuenta con una sólida base teórica, este algoritmo requiere un grupo de datos de aprendizaje, puede procesar grandes cantidades de información; el objetivo de este algoritmo es descubrir la mejor opción de clasificación, que pueda diferenciar dos clases en el grupo de datos de entrenamiento. La métrica para el significado de la “mejor” función de clasificación, se puede realizar geométricamente, una función de clasificación lineal pertenece a un hiperplano de separación f (x) que pasa por el medio de las dos clases, separando los dos tipos de clases encontradas a partir de los datos de entrenamiento. Una

vez que se determina la mejor función, la nueva instancia de datos xn se puede clasificar, simplemente, probando el signo de la función f (xn); xn pertenece a la clase positiva si f (xn)> 0 (Xindong Wu, 2007).

Debido a que existen varios hiperplanos lineales, el algoritmo SVM, garantiza que la mejor función se encuentra maximizando el margen entre las dos polaridades, el margen es la cantidad de espacio o separación entre las dos clases según lo definido por el hiperplano. Para la experimentación y evaluación de las herramientas de minería de texto, se propone ocupar el algoritmo SVM, el mejor para resolver problemas de clasificación según (Xindong Wu, 2007), es el segundo en la lista de los 10 mejores algoritmos de minería de datos, y el primero en diferenciar polaridades de un grupo de datos, además que está presente en la mayor parte de estudios similares de análisis de sensibilidad, como en "*Real-Time Detection of Traffic From Twitter Stream Analysis*" de (Eleonora D'Andrea, 2015), que menciona a SVM como un clasificador binario de eventos positivos y eventos negativos (terremotos y tifones); de igual modo el documento de (Mustafa Sofean, 2012) con el título "*A Real-Time Architecture for Detection of Diseases using Social Networks: Design, Implementation and Evaluation*", utiliza el algoritmo SVM para resolver problemas de clasificación. Por lo cual, se ha determinado el uso de dicho algoritmo, sobre las 5 herramientas en evaluación: Rapidminer, Knime, SAS, SAP y Alteryx.

2.8.2 k-nearest neighbor classification

KNN, o vecino más cercano, encuentra un grupo de k objetos, que se encuentra en un conjunto de entrenamiento, el cual está más cerca del objeto de prueba, y basa la asignación de una clase en la mayoría de objetos encontrados en un grupo o barrio; este método ocupa tres elementos claves para su desarrollo: un conjunto de objetos previamente seleccionados asignados a una clase, una medida de distancia o comparación para calcular la distancia entre objeto y objeto, y el valor de k, el número de vecinos más cercanos. Para clasificar un objeto sin clase, se calcula la distancia de este nuevo objeto a los objetos ya clasificados, se identifica sus k vecinos más cercanos y la clase de estos

vecinos más cercanos, se usan para determinar la etiqueta de clase del objeto (Xindong Wu, 2007).

2.8.2.1 Problemas

El método tiene problemas clave que afectan el rendimiento de KNN. Uno es la selección de k; si k es muy bajo, entonces los resultados pueden ser afectos por punto de ruido. De igual manera si k es muy alto, entonces el vecindario puede incluir puntos de otras clases. Otro problema es el enfoque, que se le da a las clases para etiquetar, la forma más simple es utilizar una mayoría de aciertos, pero genera problemas si los vecinos más cercanos varían, ampliamente, en su distancia y estos vecinos indican la clase de un objeto. (Xindong Wu, 2007)

2.8.3 Naive Bayes

El algoritmo Naive Bayes, es un simple clasificador probabilístico, dado un grupo de objetos, cada uno de los elementos de este grupo pertenece a una clase, y cada uno tiene un vector de variables; el objetivo del algoritmo, es construir una regla que permita clasificar objetos futuros a una clase, ingresando solo las variables del vector del objeto nuevo. Las características importantes del método de Naive Bayes son: fácil de construir, no necesita un esquema de estimación de parámetros complicado y es fácil de interpretar; esto permite al algoritmo procesar grandes grupos de datos (Xindong Wu, 2007).

El principio básico de su funcionamiento asume que solo existe dos tipos de clases 0 y 1, el objetivo es usar un grupo inicial de objetos previamente clasificados en la clase 0 o 1; a esto se denomina conjunto de entrenamiento, con estos datos se construye una puntuación, de tal forma que los valores más altos están asociados, por ejemplo: a la clase 1 y los valores más bajos están dentro de la clase 0. La clasificación se logra luego de comparar estos valores con un modelo, t (Xindong Wu, 2007).

Este método es usado por su simplicidad, facilidad de interpretación y robustez. Es el algoritmo de clasificación formal más antiguo, que a pesar de su simplicidad, es bastante efectivo. Se usa en el área de minería de datos, clasificación de texto, aprendizaje de texto,

reconocimiento de patrones y filtrado de correo no deseado. Con el tiempo se ha desarrollado para hacerlo más flexible, pero las distintas modificaciones del algoritmo son complicadas, quitándole valor a su principio básico de sencillez. (Xindong Wu, 2007)

2.9 Evaluación

Para evaluar las herramientas seleccionadas: Rapidminer, Knime, Alteryx, SAS, SAP; se tomarán en cuenta tres ejes principales: modelo de calidad de software, modelo funcionabilidad de software y modelo de confiabilidad; según estos ejes y la calificación que cada herramienta obtenga en la evaluación se determinará la mejor.

2.9.1 Modelos de calidad

Se puede definir un modelo de calidad, como un objeto especialmente planteado y construido para soportar la evaluación y selección de componentes de programas computacionales. Este proporciona la descripción de requerimientos, la especificación estructurada de criterios de evaluación, la identificación de discontinuidades de manera metódica; facilitando el proceso de evaluación y selección de software (Javier Saldarini, 2017).

Para calificar la calidad de software, se puede adoptar dos tipos de modelos: a) ocupar un modelo fijo, que considere a todos los factores de calidad relevantes, como subconjunto de un modelo base; y b) desarrollar un modelo de calidad mixto propio, que considere varios atributos, pero también adopte lo impuesto por modelos fijos (Fillottrani, 2007).

2.9.1.1 Modelo fijo de McCall

El modelo de calidad de McCall, fue desarrollado por US Air Force y DoD en 1977, identifica atributos clave desde la perspectiva del usuario; estos atributos identificados por el modelo, se denominan factores de calidad, estos factores de evaluación son medidos por medio de criterios de calidad (Fillottrani, 2007).

McCall, propone tres perspectivas para asociar los factores de calidad, los cuales son: a) revisión del producto, capacidad para ser cambiado; b) transición del producto, adaptabilidad al nuevo ambiente; y c) operación del producto, características de operación (Fillottrani, 2007).

2.9.1.1.1 Revisión de producto

Esta perspectiva contiene factores de calidad, los mismos son: 1) mantenibilidad, esfuerzo requerido para localizar y corregir fallas; 2) flexibilidad, facilidad de realizar cambios; y 3) testeabilidad, facilidad para realizar el *testing*, para verificar que el producto no tenga errores y cumpla con las especificaciones requeridas (Fillottrani, 2007).

2.9.1.1.2 Transición del producto

La perspectiva transición del producto, incluye los siguientes factores de calidad: 1) portabilidad esfuerzo requerido, para transferir la plataforma a distintos ambientes de operación; 2) reusabilidad, facilidad de reusar el software en diferentes escenarios; y 3) interoperabilidad, esfuerzo requerido para acoplar el producto con otros sistemas (Fillottrani, 2007).

2.9.1.1.3 Operación del producto

La perspectiva de operación del producto, tiene los siguientes factores de calidad: 1) correctitud, grado en que el producto cumple con su especificación; 2) confiabilidad, capacidad del producto de responder ante escenarios no esperados; 3) eficiencia, uso de los recursos, tales como: tiempo de ejecución y memoria de ejecución; 4) integridad, protección del programa y sus datos, de accesos maliciosos 5) usabilidad, facilidad de operación del producto por parte del usuario (Fillottrani, 2007).

2.9.1.2 Modelo de Boehm

El modelo de Boehm, se presentó por primera vez en 1978, desarrollado por Barry Boehm; este modelo de calidad incorpora características de: alto nivel, nivel intermedio y básicas; todas estas características contribuyen al nivel general de calidad (Fillottrani, 2007).

2.9.1.2.1 Características de alto nivel

Las características de alto nivel, hacen referencia a los requerimientos generales de uso, las cuales pueden ser: 1) utilidad per-se cuan (usable, confiable, eficiente), es el producto; 2) mantenibilidad, facilidad para ser modificado, entenderlo y volver a probarlo; 3) utilidad general, se puede utilizar al cambiar el ambiente (Fillottrani, 2007).

2.9.1.2.2 Características de Nivel intermedio

Las características de nivel intermedio indican los factores de calidad de Boehm, siendo: 1) portabilidad (utilidad general), 2) confiabilidad (utilidad per-se), 3) eficiencia (utilidad per-se), 4) usabilidad (utilidad per-se), 5) testeabilidad (mantenibilidad), 6) facilidad de entendimiento (mantenibilidad), y 7) modificabilidad o flexibilidad (mantenibilidad).

2.9.1.2.3 Características primitivas

Las características primitivas, son el nivel más bajo, corresponde a características directamente asociadas a una o dos métricas de calidad; las cuales pueden ser: 1) de portabilidad: independencia de dispositivos y auto-contención; 2) de confiabilidad: auto-contención, exactitud, completitud, consistencia y robustez/integridad; 3) de eficiencia: accesibilidad y eficiencia de uso de dispositivos; 4) de usabilidad: robustez/integridad, accesibilidad y comunicación; 5) de testeabilidad: comunicación, auto descripción y estructuración de calidad; 6) de entendibilidad: consistencia, estructuración, concisidad y legibilidad; y 7) de modificabilidad: estructuración y aumentabilidad.

2.9.2 Modelo de Evaluación de calidad de software

Para la ejecución del estudio se propone un modelo de evaluación de calidad mixto, tomando como referencia los factores del modelo de evaluación de McCall; porque este modelo se ajusta más a las necesidades de la investigación por su sencillez y facilidad de modificación, se puntúa la calidad de software. El nuevo modelo de evaluación consta de 14 factores de comparación, que se aprecian en la tabla 1, con los cuales se evaluarán las herramientas de minería de texto, a cada factor se le asigna un peso, que se detalla en la tabla 2, dependiendo de la importancia del factor para la evaluación, en este modelo la

calificación de cada factor evaluado irá de 1 a 10, esto dictaminará las ventajas y desventajas de las plataformas sometidas al modelo de calidad.

Tabla 1. Modelo de evaluación de calidad de software

Factores	Descripción	Peso
Flexibilidad	Permite modificaciones.	8
Portabilidad	Permite usarlo en diferentes máquinas independientemente del sistema operativo.	2
Interoperabilidad	Permite la comunicación con otros sistemas.	10
Manual técnico	Posee documentación técnica.	5
Manual de usuario	Posee documentación para el usuario.	7
Ayuda en línea	Posee ayuda del desarrollador en la página oficial (foros, ejemplos).	15
Fácil de instalar	Instalación guiada.	4
Fácil de configurar	Configuración guiada.	5
Amigable	Entorno intuitivo y fácil uso para el usuario.	20
Íconos con ayuda	Posee ayuda de fácil acceso.	5
Seguridad	Posee mecanismos que controlen o protejan los programas o los datos.	2
Actualizaciones	Facilidad de agregar actualizaciones o complementos.	10
Soporte técnico	Ayuda del desarrollador.	2
Independencia de hardware	Independencia de componentes de Hardware, requerimientos mínimos	5

2.9.2.1 Pesos de Evaluación de calidad

En la tabla 2, se describe la relevancia dada a los factores de evaluación ordenados de mayor a menor, sumando 100 en total, estos factores fueron puntuados en base a la utilidad y criterio de importancia para el desarrollo de esta comparación de plataformas; estos valores son especificados, con la finalidad de encontrar una herramienta de fácil manejo y compresión para el usuario.

Tabla 2. Modelo de evaluación de calidad de software

Factores	Peso	Descripción
Amigable	20	Alta relevancia, para esta comparación es de vital importancia la interacción del software con el usuario que tan intuitivo es, ya que son herramientas de las cuales no se tiene la experiencia de haberlas usado antes.
Ayuda en línea	15	Alta relevancia, al experimentar con herramientas de minería de datos nuevas, es necesario ayuda en línea con ejemplos, explicación de cómo funciona sus componentes.
Interoperabilidad	10	Muy relevante, la conexión con Twitter es muy importante para el desarrollo de esta evaluación
Actualizaciones	10	Muy relevante, es necesario contar con complementos que ayuden a la herramienta en la minería de texto
Flexibilidad	8	Relevante, es importante que el programa permita interactuar con sus herramientas y modificar datos que van a ser procesados
Manual de usuario	7	Relevante, e de mucha ayuda para entender un programa nuevo, que la plataforma tenga descripción sobre sus componentes
Íconos con ayuda	5	Media relevancia, cuando un programa es nuevo es importante tener ayuda sobre el mismo de fácil acceso
Fácil de configurar	5	Media relevancia, la configuración guiada del programa es importante cuando no se conoce aún la plataforma
Independencia de hardware	5	Media relevancia, saber si el programa es compatible con la computadora que se usará es importante
Manual técnico	5	Media relevancia, es de importancia tener la descripción de las herramientas del software
Fácil de instalar	4	Baja relevancia, la instalación se asume es guiada paso a paso
Soporte técnico	2	No es de relevancia, por el motivo de tiempo de respuesta del proveedor
Portabilidad	2	No es de relevancia , las pruebas se harán en una sola máquina, con un solo sistema operativo
Seguridad	2	No es de relevancia, ya que los datos usados en la herramienta son para pruebas

2.9.3 Modelo de Evaluación de funcionabilidad de software

La evaluación de la funcionabilidad de cada software estará medida por 5 factores importantes en la minería de texto descritos en la tabla 3; a los cuales se asigna un peso, que se detalla en la tabla 4; el cual considera la importancia de cada factor; la calificación que reciban los factores a ser evaluados de las plataformas de minería, será de 1 a 5. Cabe resaltar un factor muy importante: las herramientas serán probadas con un mismo algoritmo (SVM), con la misma limpieza de datos, los mismos datos de aprendizaje, los mismos datos para el análisis de sentimiento, el mismo gráfico de barras y se tomará muy en cuenta el porcentaje de validez.

Tabla 3. Modelo de Evaluación de funcionabilidad de software

Factores	Descripción	Peso
Procesamiento de Datos	Facilidad para limpiar y preparar los datos	15
Modelos De análisis de sentimiento	Facilidad de uso de logaritmos	20
Métodos de aprendizaje de datos	Facilidad y correcto funcionamiento de aprendizaje	20
Validación del Modelo	Veracidad del modelo	30
Graficación	Facilidad para graficar los resultados	15

2.9.3.1 Pesos de evaluación de funcionabilidad

La descripción de cada uno de los pesos de la evaluación de funcionabilidad, se detalla en la tabla 4; estos valores fueron asignados a criterio de la necesidad de los resultados obtenidos por las herramientas. Luego de procesar los mismos datos, los pesos están ordenados de mayor a menor importancia; estos valores se calificarán luego de la experimentación con cada una de las herramientas de minería de texto seleccionadas.

Tabla 4.Modelo de Evaluación de funcionabilidad de software

Factores	Peso	Descripción
Validación del Modelo	30	Alta relevancia, es muy importante que los resultados sean correctos con respecto a su polaridad.
Modelos De análisis de sentimiento	20	Muy relevante, es importante que la plataforma permita configurar el algoritmo propuesto para clasificar los datos.
Métodos de aprendizaje de datos	20	Muy relevante, es de gran valor que la herramienta permita un fácil procesamiento de los datos de aprendizaje del algoritmo.
Procesamiento de Datos	15	Relevante, es necesario que la herramienta permita pre procesar los datos antes de ser aplicados en el algoritmo
Graficación	15	Relevante, la herramienta debe permitir visualizar gráficos de los resultados, haciendo más fácil la compresión de los mismos

2.9.4 Modelo de evaluación de métricas estadísticas, para determinar la exactitud del modelo

Para este modelo no se tomará en cuenta el factor *accuracy*; ya que este valor se usa para datos balanceados y en esta experimentación se realizará sobre datos no balanceados; en la tabla 5, no se coloca un peso, puesto que la mejor herramienta, será aquella que de un valor de f-measure más alto. Los factores de este modelo son:

Tabla 5. Modelo de Evaluación de métricas estadísticas para determinar la exactitud del modelo

Factores	Descripción
Recall	Cobertura
f-measure	Media armónica entre recall y precisión
precision	Cuantos son realmente positivos
Kappa	Fuerza de concordancia entre el análisis humano y el análisis automático

2.9.4.1 Índices de concordancia de Kappa

Kappa se obtiene con el empleo de la ecuación 1:

$$k = \frac{P_o - P_e}{1 - P_e}$$

Ecuación 1. Kappa (Hospital Universitario Ramón y Cajal, 2018)

Siendo:

- P_0 (*observed accuracy*) la proporción de verdaderos encontrados, por medio de la ecuación 2:

$$P_o = (TP + TN) / (TP + FP + FN + TN)$$

Ecuación 2. Observed accuracy (Hospital Universitario Ramón y Cajal, 2018)

- P_e (*expected accuracy*) la proporción de verdaderos esperados de los observados, en la ecuación 3.

$$P_e = (TP + TN) / (TP + FP + FN + TN)^2$$

Ecuación 3. Expected accuracy (Hospital Universitario Ramón y Cajal, 2018)

Donde:

- TP = (*True positive)* verdadero positivo
- TN = (*True negative)* verdadero negativo
- FP = (*False positive)* falso positivo
- FN= (*False negative)* falso negativo

(Hospital Universitario Ramón y Cajal, 2018)

Los índices de grado concordancia se detallan en la tabla 6.

Tabla 6 Índices de concordancia de Kappa. (Hospital Universitario Ramón y Cajal, 2018)

Kappa	Grado de acuerdo
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

2.10 Calificación

Para la calificación de las herramientas, se tomará en cuenta los factores de los modelos de evaluación anteriormente descritos; esta calificación se genera mediante la experimentación directa sobre las herramientas, para lo cual cada plataforma será descargada, instalada, configurada y probada, esto como parte del modelo de calidad; de igual forma, será calificada la facilidad de obtención de los resultados evaluados por los factores del modelo de funcionabilidad.

Con el documento "A Methodology for Evaluating and Selecting Data Mining Software" de (Collier, Carey, Sautter, y Curt (2015), se usará para calificar las plataformas de minería de texto, ya que el documento menciona una metodología basada en experimentación real en minería de datos, utilizando conjuntos de datos comerciales de una variedad de empresas.

El costo de seleccionar una herramienta de minería inapropiada, para un uso en particular, implica gastos de recursos de personal, tiempo perdido y el riesgo de actuar sobre resultados falsos o erróneos. (Collier, Carey, Sautter, y Curt (2015)

2.10.1 Metodología de calificación de herramientas de minería de texto

Según (Collier, Carey, Sautter, y Curt (2015), la elección de una herramienta puede estar dada por los siguientes pasos:

1. Preselección de herramientas: Este primer paso tiene como objetivo principal reducir el conjunto de herramientas a una cantidad menor, con el fin de evaluarlas de mejor manera.
2. Identificar criterios de selección adicionales: Para este caso es importante definir para qué se va a usar la herramienta. El objetivo es identificar criterios específicos y necesarios para la realización del proyecto y la obtención de resultados útiles.
3. Criterios de selección de peso: Dentro de este paso se establecerán pesos a los factores de evaluación, con equivalencia total a 1.00 o 100%; esta asignación se realizará con respecto al uso previsto de la plataforma.

4. Puntaje de la herramienta: Los criterios de evaluación se contrastarán con el desenvolvimiento de la herramienta frente al factor de evaluación. Como ejemplo: si el software no es portable o es poco portable, recibirá una calificación de 2; mientras que si, el software puede ser ocupado en cualquier sistema operativo la, calificación será de 5. Los valores de calificación van de 1 a 5 y corresponden a la siguiente escala:
    1 Pésimo.
    2 Malo.
    3 Regular.
    4 Bueno.
    5 Muy bueno.

Los valores que van de 1 a 10 corresponden a la siente escala:

1 No existe.
2 Muy malo.
3 Malo.
4 Deficientes.
5 Regular.
6 Bueno.
7 Aceptable.
8 Bien.
9 Muy bien.
10 Excelente.

5. Evaluación de puntuación: La metodología para obtener la puntuación, será el total de la sumatoria de la multiplicación entre la calificación, dependiendo del factor por el peso del factor, por medio de la ecuación 4:

$$Total = (Calif_1 * Peso_1) + (Calif_2 * Peso_2) \dots + (Calif_n * Peso_n)$$

Ecuación 4. Evaluación de puntuación

6. Selección de herramientas: La selección de la mejor herramienta estará dictaminada por el mayor valor obtenido de la sumatoria de sus puntajes.

2.11 Conclusión

La selección de una herramienta para minería de texto, está fuertemente ligada al tipo de análisis que se pretende realizar; puesto que las plataformas de minería, ofrecen una gran variedad de soluciones para *text mining*, por lo cual es preciso identificar qué resultados se esperan, ya sean de predicción, clasificación o análisis de sentimiento.

Las herramientas elegidas para este proyecto corresponden a los análisis de Forrester Wave y al cuadrante mágico de Gartner, los que están al día con las ofertas del mercado en *text mining*. Estos softwares, además de ser los primeros en los análisis mencionados, también aparecen en documentos afines a este proyecto, lo cual ratifica su elección.

CAPÍTULO III: RECOPILACIÓN Y GENERACIÓN DE DATOS

3.1. Introducción

A partir de la problemática del proyecto se he determinado obtener mensajes en español en la red social Twitter, relacionados al tráfico de todo el mundo. Los tweets se extraen directamente con la API de Twitter y la plataforma R, en la cual se implementó "Search Twitter" bajo la cadena de búsqueda: "(congestion OR congestión OR circulación OR circulación OR transito OR tránsito) AND vehicular", la cual devolvió 3.949 tweets.
Para la obtención de los archivos *.csv*, como primer paso se crea un aplicativo, que tendrá la utilidad de conectar Twitter con R; este aplicativo se genera por medio de un usuario de la API para desarrolladores.

Se ingresa a la página web de desarrolladores de Twitter en la dirección: https://apps.twitter.com/

En la figura 3, se indica la página principal de twitter apps:

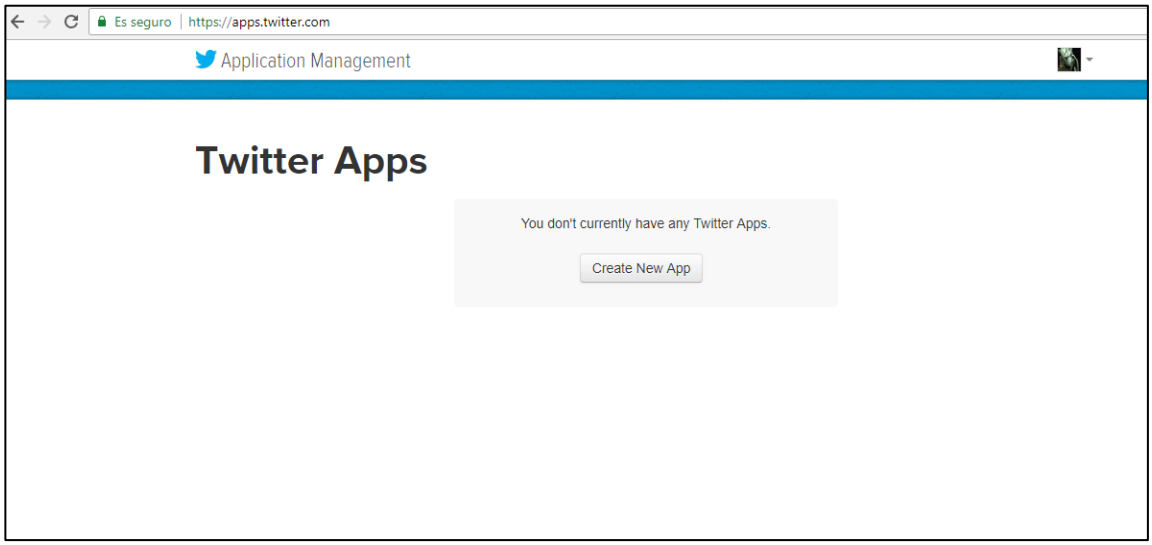

Figura 3. Captura de pantalla de página de desarrolladores de Twitter (https://apps.twitter.com/)

Se crea una nueva aplicación, donde se llenarán los parámetros indicados en la figura 4:

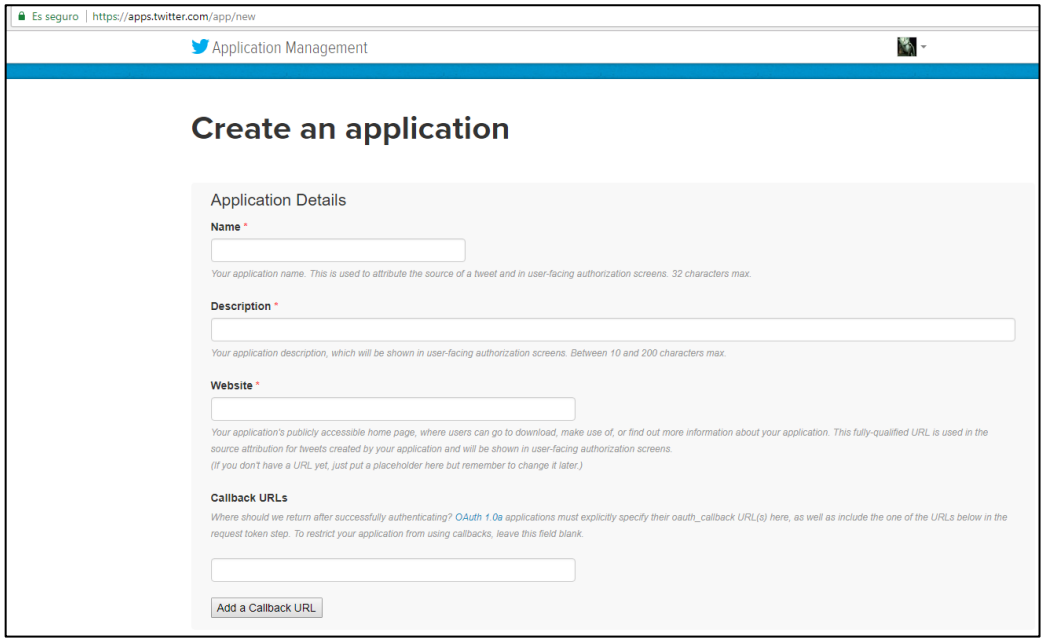

Figura 4. Captura de pantalla de campos para nueva aplicación (https://apps.twitter.com/)

Name: nombre de la aplicación.

Description: describe en pocas líneas la aplicación.

WebSite: Este campo es la página de inicio de acceso público, la nueva aplicación que se está realizando; donde los usuarios pueden descargar, usar o encontrar más información sobre la aplicación generada. Como no se tiene una página de descripción para el desarrollo de esta evaluación, se utilizará la dirección http://127.0.0.1, ya que funciona correctamente para pruebas.

CallBack URL: este campo se deja vacío, porque no se va a direccionar a ningún sitio.

Se aceptan los términos y condiciones como paso final.

Una vez creada la aplicación se obtiene: *consumer key*, *consumer secret*, *access token* y *access secret*; estos datos son los que necesita R para realizar la conexión por medio de la librería “twitter”.

- *Consumer key*: Es la clave API, asociada con la aplicación (Twitter). Esta clave identifica al usuario de la red social, un cliente es un sitio web / servicio que intenta acceder a los recursos de un usuario final. (Sapir, 2016)
- *Consumer secret*: Es la contraseña que usa el cliente para autentificarse con el servidor de Twitter (Sapir, 2016)
- *Access token*: El *token* define los privilegios del cliente (a qué datos puede y no puede acceder el cliente), se emite al usuario una vez se autentica correctamente (*Consumer key* y *Consumer secret*). (Sapir, 2016)
- *Access secret*: Cada ocasión que el usuario desea acceder a los datos del usuario final, se envía con un token de acceso como contraseña (similar al Consumer secret). (Sapir, 2016)

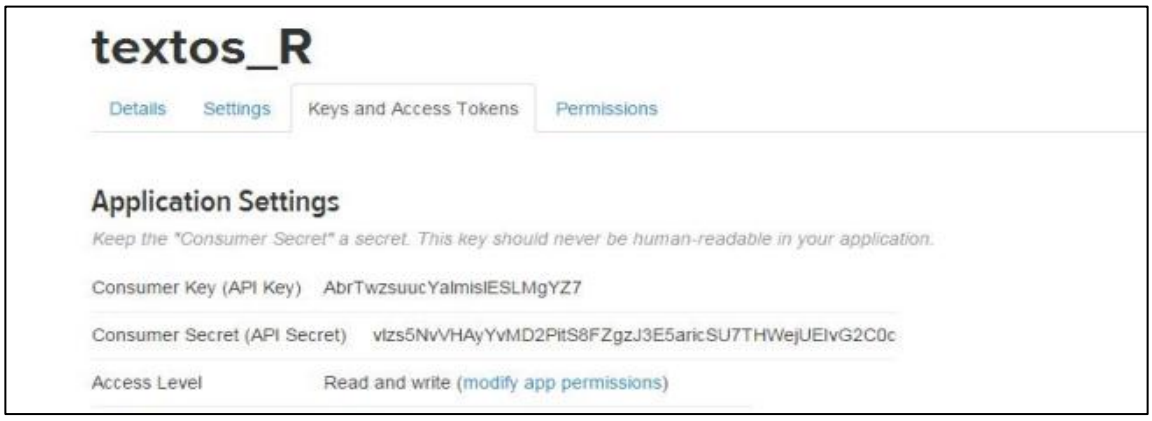

Figura 5. Captura de pantalla de los datos de conexión (https://apps.twitter.com/)

Finalmente se extraen los datos desde R, utilizando las líneas de código indicadas en la figura 6, a continuación:

```
library(twitteR)

#iniciar sesion en twitter para extraer información

consumer_key <-'zovysySilMJyfCyWFENVs4Um1'

consumer_secret <-'GQA6Dy9SJjCExELMGmNDG2JRFFTeObqhuFnPX6hiIjbeVfJIZP'

access_token <- '403195288-xWEMPxgQ61DX1JAi7oW3ArHoQ5ivQuhmenSpv2cr'

access_secret <-'j2pIc92QP416QMDlCWgrFOnQ8GVlab8bWaFqkvWfxvBjY'

setup_twitter_oauth(consumer_key, consumer_secret, access_token,
access_secret)
```

Figura 6. Captura de pantalla del software R

El proceso final consiste en transformar la información conseguida, en una colección de datos y posteriormente se obtiene solo el texto de los mismos.

3.2. Preparación de datos

En esta fase se pasan los datos en bruto, obtenidos anteriormente, a una hoja de cálculo donde serán clasificados a mano en tweets positivos y negativos, como se muestra en la tabla 7; posteriormente se transforma esta hoja en un archivo *.csv*. Todo esto es necesario para la parte de aprendizaje del modelo de análisis de sentimiento.

Para generar la matriz de entrenamiento, desarrollada por el laboratorio de investigación y desarrollo de informática (LIDI) de la Universidad del Azuay, se etiqueta uno por uno los *tweets* por parte de dos miembros del departamento.

Tabla 7. Fragmento de normalización de tweets obtenida por LIDI

Id	Polaridad	Texto
1	negativo	fuerte congestión vehicular por accidente sobre la avda. pasoancho desde la cra 89 hasta la cra 102, sentido norte- sur. #tráficocali @elpaiscali @twiteroscali y
2	negativo	río de los remedios, presenta carga vehicular moderada
3	positivo	hermilio mena y río de los remedios, presenta buen avance vehicular
4	positivo	tenayuca ambas direcciones presenta ligera carga vehicular
5	positivo	calzada vallejo ambas direcciones, presenta ligera carga vehicular
6	positivo	mario colín a la altura del tren suburbano, presenta ligera carga vehicular
7	negativo	07.11 el distribuidor miguel hidalgo presenta congestión vehicular sobre 61uechu 61uech mateos (oriente-poniente) y en paseo vicente guerrero
8	negativo	fanb #gnb 8:50 a.m #agma sentido caracas fluye el tránsito, retraso moderado para ingresar al túnel por la alta afluencia vehicular
9	negativo	vía túnel san 61uechurab: tránsito lento en salida sector av. El cerro y túnel dirección providencia por alto flujo vehicular uoct_rm radioc…
10	negativo	vía túnel san 61uechurab: tránsito lento en salida el salto dirección 61uechuraba por alto flujo vehicular uoct_rm biobio radiocarab

Como paso final, se genera un archivo .csv, que se usará en el proceso de pruebas y análisis de sentimiento. Este archivo contiene 325 tweets de texto plano, sobre la ciudad de Cuenca-Ecuador, relacionados al tráfico vehicular, como se indica en la tabla 8.

Tabla 8. Fragmentos de tweets para ser procesados por el modelo predictivo obtenidos por LIDI

Id	Texto
1	#ecu911reporta 24 de mayo y max uhle llamada indica accidente de tránsito unidades en el lugar, al momento circulación vehicular habilitada. F
2	#tránsitoecu911, se visualiza circulación vehicular normal, panamericana sur y camino a rayoloma #cuencac
3	#tránsitoecu911, se visualiza circulación vehicular baja, av.américas y héroes de verdeloma #cuenca5
4	#tránsitoecu911, se visualiza circulación vehicular normal, max uhle y pumapungo #cuenca5
5	#tránsitoecu911, se visualiza circulación vehicular baja, #latroncalw

3.3. Modelado

En esta sección se presenta la configuración del algoritmo que se seguirá, para la experimentación en las herramientas de minería de texto, además se muestra el proceso básico que se utilizará en las distintas plataformas previamente seleccionadas para minar texto.

Una vez obtenido el archivo de aprendizaje y de prueba de análisis de sentimiento, el siguiente paso es determinar el proceso general, que se seguirá en la experimentación, repitiéndose la misma configuración y los mismos pasos de limpieza de texto en cada una de las herramientas. Es importante señalar que se debe ocupar la misma configuración del kernel del algoritmo, para todas las herramientas, garantizando de esta manera que la evaluación y comparación va a ser imparcial, dejando a los softwares en igualdad de condiciones.

3.4 Limpieza de datos

En la limpieza de datos se realizará 3 pasos importantes para la realización de la experimentación: convertir todo a minúsculas, *Stopword* y *tokenizer*. El pre procesamiento de los datos es una parte clave en el proceso de minería de *tweets*.

1. Convertir todo a minúsculas. - Este primer paso convierte todo el texto de los *tweets* a minúsculas, con el fin de empezar la normalización del texto que va a ser procesado.
2. *Stopword*. - Este paso elimina los conectores de texto, los cuales no tienen relevancia para la minería como: los, las, y, en, de, ahí, ahora, etc. Es importante señalar que este proceso se realiza con un documento generado por el LIDI, con los conectores a eliminar, el cual se cargará a las plataformas, por lo que comúnmente el documento pre cargado en las herramientas contiene conectores en inglés.
3. *Tokenizer*. - Es el proceso de dividir una cadena de texto en palabras significativas. El objetivo de este paso es el análisis palabra por palabra del texto previamente procesado.

3.5 Kernel

Se denomina kernel al conjunto de funciones matemáticas que usa el algoritmo SVM; la función del kernel es tomar los datos de entrada y transformarlos en la forma que el usuario especifique. (Data Flair, 2017)

Tipos de Kernel del algoritmo:

- Polinómico: este núcleo se usa comúnmente en procesamiento de imágenes, está definido por la ecuación 5:

$$k(x_i, x_j) = (x_i \,.\, x_j + 1)^d$$

Ecuación 5. Kernel polinómico (Data Flair, 2017)

Donde d es el grado del polinomio. (Data Flair, 2017).

- Gaussiano: este kernel es de propósito general; comúnmente usado cuando no hay conocimiento previo de los datos, está definido por la ecuación 6:

$$k(x, y) = \exp(\frac{\|x - y\|^2}{2\sigma^2})$$

Ecuación 6. Kernel polinómico (Data Flair, 2017)

- Radial: Es un núcleo de propósito general, está definido por la ecuación 7:

$$k(x_i, x_j) = \exp(-\gamma \|(x_i - x_j)\|^2)$$

Ecuación 7. Kernel polinómico (Data Flair, 2017)

Para: $\gamma > 0$

- Laplace: un kernel comúnmente usado cuando no hay conocimiento previo de los datos, está definido por la ecuación 8:

$$k(x, y) = \exp(-\frac{\|x - y\|}{\sigma})$$

Ecuación 8. Kernel Laplace (Data Flair, 2017)

- Tangente hiperbólica: núcleo usado en redes neuronales, está definido por la ecuación 9:

$$k(x_i, x_j) = \tanh(kx_i \,.\, x_j + c)$$

Ecuación 9. Kernel Laplace (Data Flair, 2017)

- Anova: kernel usado en problemas de regresión, está definido por la ecuación 10:

$$k(x_i, x_j) = \sum_{k=1}^{n} \exp(-\sigma(x^k - y^k)^2)^d$$

Ecuación 10. Kernel Laplace (Data Flair, 2017)

Para resultados más imparciales, es necesario que el kernel sea el mismo en todas las herramientas, al revisar la configuración de los algoritmos pertenecientes a las plataformas. La plataforma Knime no posee todos los tipos de configuración, por esta razón el kernel seleccionado para el algoritmo SVM, es el de radio; este es elegido porque dicha configuración está presente en todas las herramientas.

3.6 Proceso general

El proceso general se indica en la figura 7, que a continuación se puede observar:

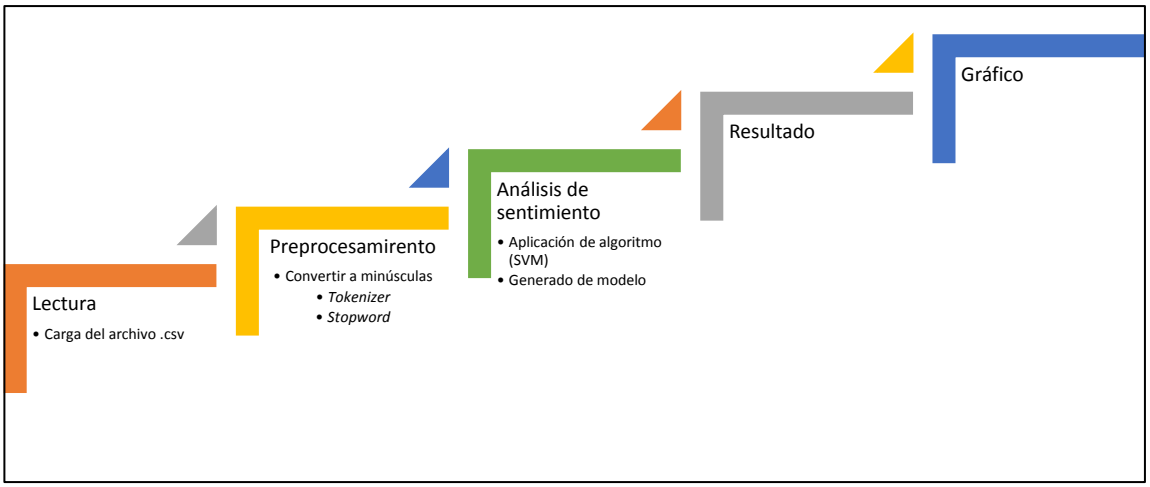

Figura 7. Proceso general

Es indispensable que el proceso de limpieza, uso de diccionario en español, configuración de Kernel del algoritmo y la graficación en barras sea el mismo en todas las herramientas; independientemente de las opciones que las plataformas ofrezcan, se requiere que el proceso sea el mismo para que la comparación final sea válida. Todas las pruebas se

realizarán en un mismo equipo, asegurando que sea el mismo escenario para las herramientas.

Características del equipo:

- Procesador: Intel Core i7 7500u 2.7-2.9 Ghz.
- Memoria Ram (Random Access Memory): 8GB.
- Sistema Operativo: Windows 10 Pro.
- GPU: NVIDIA GeForce 940MX 2GB.
- Disco duro: 500TB a 54000 RPM.

3.7 Conclusión

Es importante realizar los mismos pasos básicos en la fase de pruebas; esto determinará el éxito o fracaso del análisis de herramientas. Si una herramienta tiene diferente configuración puede beneficiar o afectar los resultados finales del análisis de sentimiento.

CAPÍTULO IV: EXPERIMENTACIÓN

4.1 Experimentación

Para la experimentación, las herramientas son evaluadas siguiendo el protocolo dispuesto con anterioridad y los resultados que estas herramientas generen, serán evaluados para determinar qué plataforma es la más recomendable para *text mining*.

4.2 Criterios de evaluación

Para determinar la mejor herramienta se ha dispuesto de tres ejes de evaluación: a) modelo de calidad de software; donde se evaluará el proceso de instalación, configuración, apariencia y respuesta del programa con el usuario, b) modelo funcionabilidad de software; en el cual las herramientas serán calificadas por los resultados obtenidos, el proceso para obtener dichos resultados, y el último eje, c) modelo de confiabilidad; este último hace referencia al porcentaje de confiabilidad que tiene los resultados de la matriz cruzada.

4.3 Ejecución de pruebas

A continuación, se presenta una explicación de cómo fue probada cada herramienta, teniendo en cuenta: los mismos datos de aprendizaje, los mismos datos de prueba para el análisis de sentimiento, la misma configuración de kernel del algoritmo, la misma limpieza de datos y el mismo diccionario de aprendizaje en español.

4.3.1 Rapidminer

4.3.1.1 Modelado

Una gran ventaja de esta plataforma son sus soluciones pre canalizadas; las cuales reducen el tiempo de modelado, colocando automáticamente los bloques básicos; en este caso para

un análisis de sentimiento, como se muestra en la figura 8, Rapidminer autogeneró este modelo; el cual como primera parte lee el archivo .csv de entrenamiento, a continuación se selecciona la columna por la cual los *tweetts* van a ser clasificados, en el bloque de proceso *Process Documents from* se realiza la limpieza de datos (conversión a minúsculas, *stopword, tokenizer),* los cuales pasan a *Cross Validation*, donde el algoritmo es entrenado con los datos ingresados y donde se construye el modelo que será utilizado para procesar los datos de prueba. Como segunda parte se ingresa en el bloque *Read CSV*, los datos de prueba, los cuales pasan hacia el proceso de limpieza, para posteriormente ser evaluados por el modelo construido anteriormente. Finalmente, se escriben los resultados en un archivo tipo xlsx y posterior a ello se muestran los resultados.

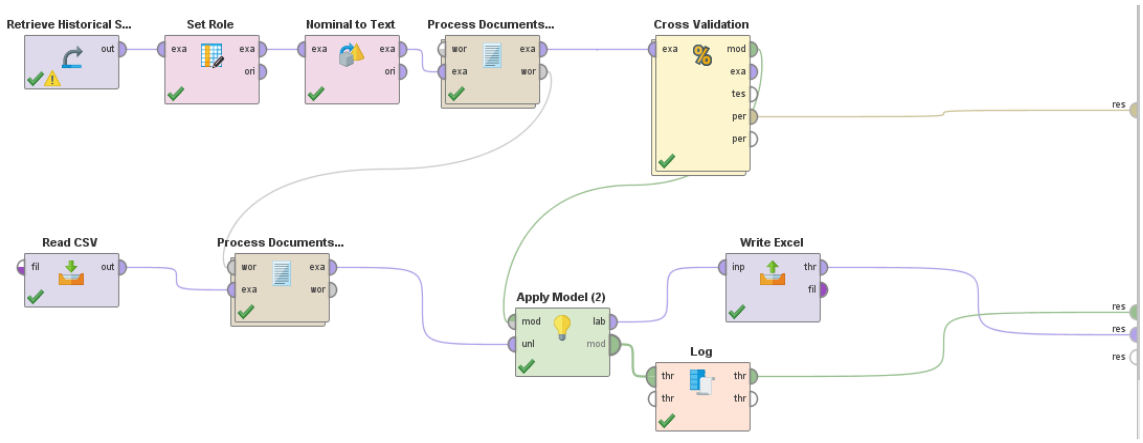

Figura 8. Modelo de Rapidminer

Para cumplir el modelo general establecido, el pre procesamiento de los datos a evaluar se cumple en el bloque *Process Documents from Data*, como se describe en la figura 9, para lo cual: en primera instancia está el proceso de *tokenize*, donde el texto es separado en palabras; como siguiente paso se cambian todas las mayúsculas por minúsculas y finalmente se quitan los conectores como: la, los, en, hacia, donde, etc.

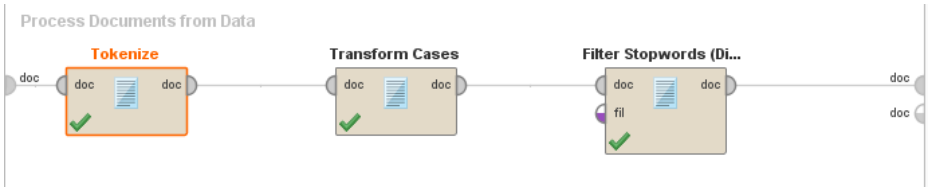

Figura 9. Pre procesamiento de datos Rapidminer

Los datos procesados son pasados al bloque *Cross Validation*, en donde se aloja el algoritmo SVM; en esta etapa es donde el kernel del algoritmo es configurado, como se describe en la figura 10.

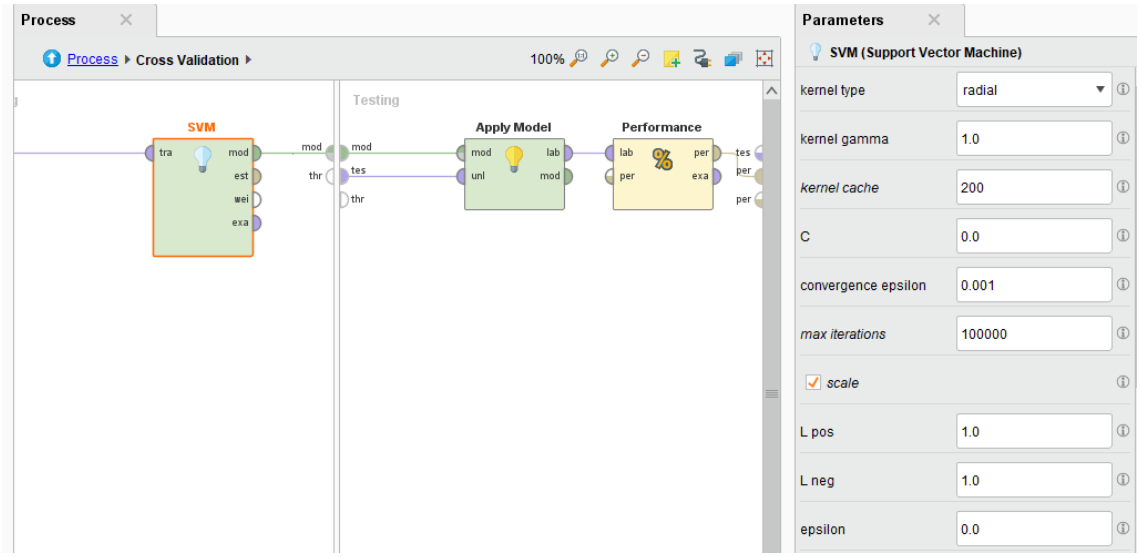

Figura 10. Cross Validation Rapidminer

4.3.1.2 Resultados

Una vez analizados los datos, la plataforma permite ver en una tabla los resultados del análisis predictivo. En la figura 11, se presenta una muestra de 10 datos de los 324 analizados, la cual contiene el número de fila, como el algoritmo clasificó el texto ingresado, la probabilidad de que este sea clasificado como positivo, la probabilidad de que el texto sea clasificado como negativo y en la última columna el texto analizado.

Row No.	prediction(p...	confidence(...	confidence(...	text
1	negativo	0.731	0.269	ecu reporta mayo max uhle llamada indica accidente tránsito unidades circulación vehicular habilitada
2	positivo	0.423	0.577	tránsitoecu visualiza circulación vehicular normal panamericana sur camino rayoloma cuencac
3	positivo	0.269	0.731	tránsitoecu visualiza circulación vehicular baja av américas héroes verdeloma cuenca
4	positivo	0.423	0.577	tránsitoecu visualiza circulación vehicular normal max uhle pumapungo cuenca
5	positivo	0.423	0.577	tránsitoecu visualiza circulación vehicular baja latroncalw
6	positivo	0.423	0.577	tránsitoecu visualiza circulación vehicular baja azoguesa
7	positivo	0.423	0.577	cuenca señalización emov ep av abelardo andrade mejora circulación vehicular tráfico pesado
8	positivo	0.423	0.577	señalización observar mejoramiento circulación vehicular colas interminables av abelardo andrade to...
9	negativo	0.731	0.269	azuay cuenca obraspublicasec informa cierre tránsito vehicular vía descanso puente
10	positivo	0.423	0.577	tránsitoecu visualiza circulación vehicular baja av américas turuhuayco cuencaq

Figura 11. Resultados de predicción de Rapidminer

4.3.1.2.1 Matriz de confusión

En la tabla 9, se presenta la matriz de confusión, donde 2751 es el número de verdaderos positivos; verdaderos positivos o *true positive* (TP), se denomina a el número de *tweets* que el algoritmo clasificó como positivos, cuando estos eran positivos realmente; 287 representa a falsos negativos o *false positive* (FP) son aquellos que eran positivos, pero el algoritmo los clasificó como negativos, *false negative* (FN) o falsos negativos; en este caso 135 son datos que siendo negativos, el algoritmo los clasificó como positivos; finalmente verdaderos negativos o *true nevative* (TN), 451 son aquellos datos que originalmente son negativos y el algoritmo los clasificó como negativos.

Tabla 9. Matriz de confusión de Rapidminer

	Condición positivo	Condición negativo
Predicción positivo	2751	287
Predicción negativo	135	451

En la tabla 10, se indican los resultados *Recall, Precision, F-Measure, Kappa*; obtenidos a partir de los valores de la matriz de confusión, en ella se describe cada uno de las variables obtenidas y la fórmula para obtener dicho valor.

Tabla 10. Resultados Recall, Precision, F-Measure, Kappa de Rapidminer

	Valor	Descripción
Recall	95,32	Es el porcentaje de los que en realidad son positivos, cuantos al final fueron clasificados como positivo (TP/TP+FN)
Precision	90,55	Es el porcentaje de los clasificados como positivos, cuales al final realmente son positivos (TP/TP+FP)
F-Measure	92,88	Media armónica entre Recall y Precision 2*(Recall * Precision) / (Recall + Precision)
Kappa	0.61	El grado de concordancia es **moderado**

4.3.1.3 Gráfico de resultados de Rapidminer

Como se observa en la figura 12, el análisis predictivo dicta que el 88% de los *tweetts* pertenece a una congestión vehicular negativa y solo un 12% de los datos habla de que el tráfico vehicular es positivo.

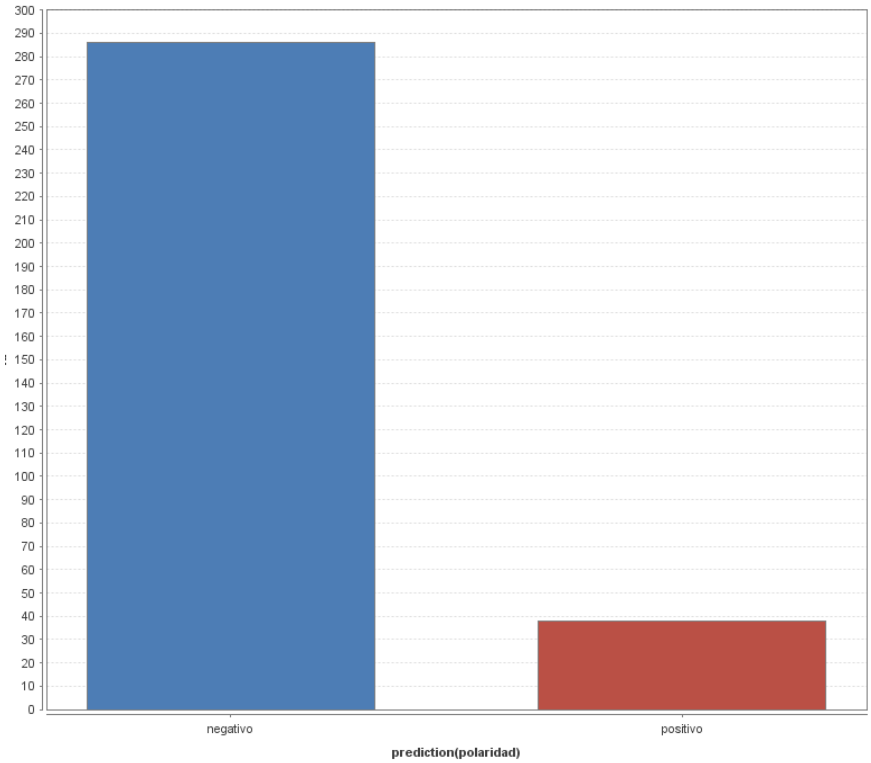

Figura 12. Representación gráfica de los datos en Rapidminer

4.3.1.4 Modelo de Evaluación de calidad de software de Rapidminer

En la tabla 11, se observa la calificación en el modelo de calidad de software concedida a la herramienta Rapidminer; los valores de calificación asignados a cada factor corresponden a una previa experimentación con los datos de evaluación sobre la herramienta.

Tabla 11. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Rapidminer

Factores	Calificación	Descripción
Flexibilidad	9	Permite la manipulación de los datos, algoritmos y procesos
Portabilidad	9	Es compatible con más de un sistema operativo
Interoperabilidad	9	Tiene conexión directa con Twitter
Manual técnico	9	Descripción competa de los complementos del software
Manual de usuario	9	Basta información en línea y consta de un manual integrado en la plataforma
Ayuda en línea	8	Existe varis ejemplos, blogs, video, tutoriales
Fácil de instalar	9	Proceso de instalación guiado
Fácil de configurar	9	Proceso de configuración guiado
Amigable	10	El software es bástate interactivo, muy intuitivo, facial de comprender y utilizar, consta de pre procesos de ayuda ya realizados
Íconos con ayuda	8	Consta con ayuda en todas sus herramientas
Seguridad	8	Si brinda seguridad de datos
Actualizaciones	9	Fácil instalación de actualizaciones desde la barra de menú
Soporte técnico	9	Tiene contactos de soporte técnico
Independencia de hardware	7	Completamente compatible, aunque la plataforma pide requerimientos mínimos para operar correctamente

4.3.1.5 Modelo de Evaluación de funcionabilidad de software de Rapidminer

La tabla 12, describe la calificación en el modelo de funcionabilidad de software concedida a la herramienta Rapidminer, los valores de calificación asignados a cada factor corresponden a una previa experimentación con los datos de evaluación sobre la herramienta.

Tabla 12. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Rapidminer

Factores	Calificación	Descripción
Procesamiento de Datos	4	Permite la evaluación de datos de una manera gráfica y sencilla por medio de bloques de procesos
Modelos De análisis de sentimiento	5	Permite configura el algoritmo de una manera sencilla, y consta de varios tipos de kernel
Métodos de aprendizaje de datos	4	Con la ayuda de las soluciones pre programadas de la plataforma, se torna sencillo el aprendizaje del algoritmo ya que no es complicado entender el proceso para el de configuración del mismo
Validación del Modelo	5	Valores precisos
Graficación	5	No es necesario un bloque de graficación, dispone de varios modelos de gráficos.

4.3.2 Knime

4.3.2.1 Modelado

Para la construcción de este modelo se utiliza los ejemplos pre diseñados del software, los cuales son de gran ayuda al momentos de realizar un nuevo análisis; el modelado en esta plataforma es por bloques de proceso y la gran ventaja de estos es la capacidad de ser ejecutados por separado y no todo el modelo en conjunto, reduciendo el tiempo de procesamiento. El modelo se presenta en la figura 13, en la primera parte se lee el archivo CSV, como segunda parte en el bloque *Document Creation* se convierte el archivo en texto, posteriormente se selecciona la columna con la cual se va a evaluar; el bloque *Preprocessing* cumple con las funciones de limpieza de los datos, para luego ser divididos por palabras relevantes y clasificados; con estos datos ya procesados a continuación se le asigna una color, para continuar con el paso de *Partitioning*, donde la matriz de datos se divide en: datos de aprendizaje y datos de prueba. En Knime el algoritmo consta de: el aprendiz y el predictor; una vez procesados los datos es necesario agregar un bloque que grafique los datos y un bloque que muestre los resultados.

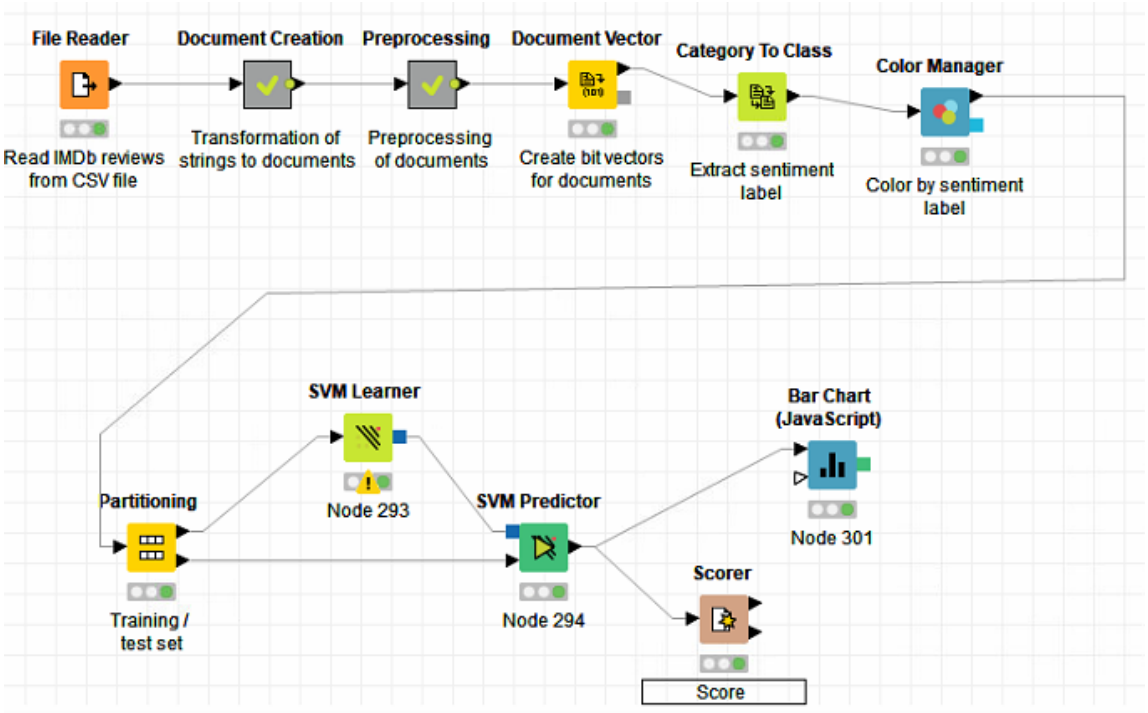

Figura 13. Modelado de Knime

La figura 14, muestra como es el pre procesamiento de datos en esta plataforma; donde a los datos, con la ayuda de esta herramienta, elimina signos de puntuación, números y convierte todo en musculas; a continuación se retiran las palabras conectores, ingresadas en una lista previamente cargada; por último, se realiza el proceso de *tokenizer*, es importante señalar que la herramienta ocupa varios bloque de procesamiento para realizar esta acción.

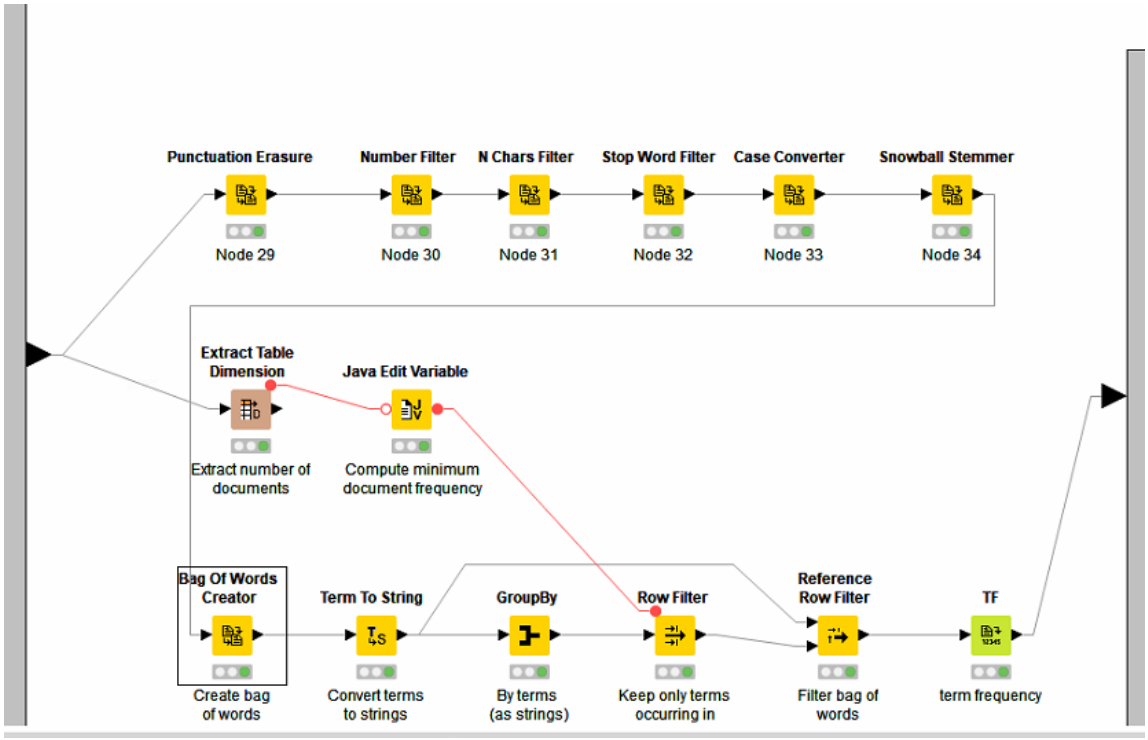

Figura 14. Modelado de Knime pre procesamiento

4.3.2.2 Resultados

4.3.2.2.1Matriz de confusión

En la tabla 13, se presenta la matriz de confusión, donde 562 es el número de (TP), 177 representa a (FP), (FN) en este caso 65, finalmente (TN) que son 2820.

Tabla 13. Matriz de confusión de Knime

	Condición positivo	Condición negativo
Predicción positivo	562	177
Predicción negativo	65	2820

En la tabla 14, se muestran los resultados *Recall, Precision, F-Measure, Kappa*; obtenidos a partir de los valores de la matriz de confusión; en ella se describe cada una de las variables obtenidas y la fórmula para obtener dicho valor.

Tabla 14. Resultados Recall, Precision, F-Measure, Kappa de Knime

	Valor	Descripción
Recall	89,63	Es el porcentaje de los que en realidad son positivos, cuantos al final fueron clasificados como positivo (TP/TP+FN)
Precision	76,05	Es el porcentaje de los clasificados como positivos, cuales al final realmente son positivos (TP/TP+FP)
F-Measure	82,28	Media armónica entre Recall y Precision 2*(Recall * Precision) / (Recall + Precision)
Kappa	0.78	El grado de concordancia es **sustancial**

4.3.2.3 Gráfico de resultados de Knime

En la figura 15, se observa que el mayor porcentaje de los datos analizados son negativos con el 83.38%, demostrando el descontento de los ciudadanos sobre el tráfico vehicular; contrario al 16.62%, que representan tweets positivos.

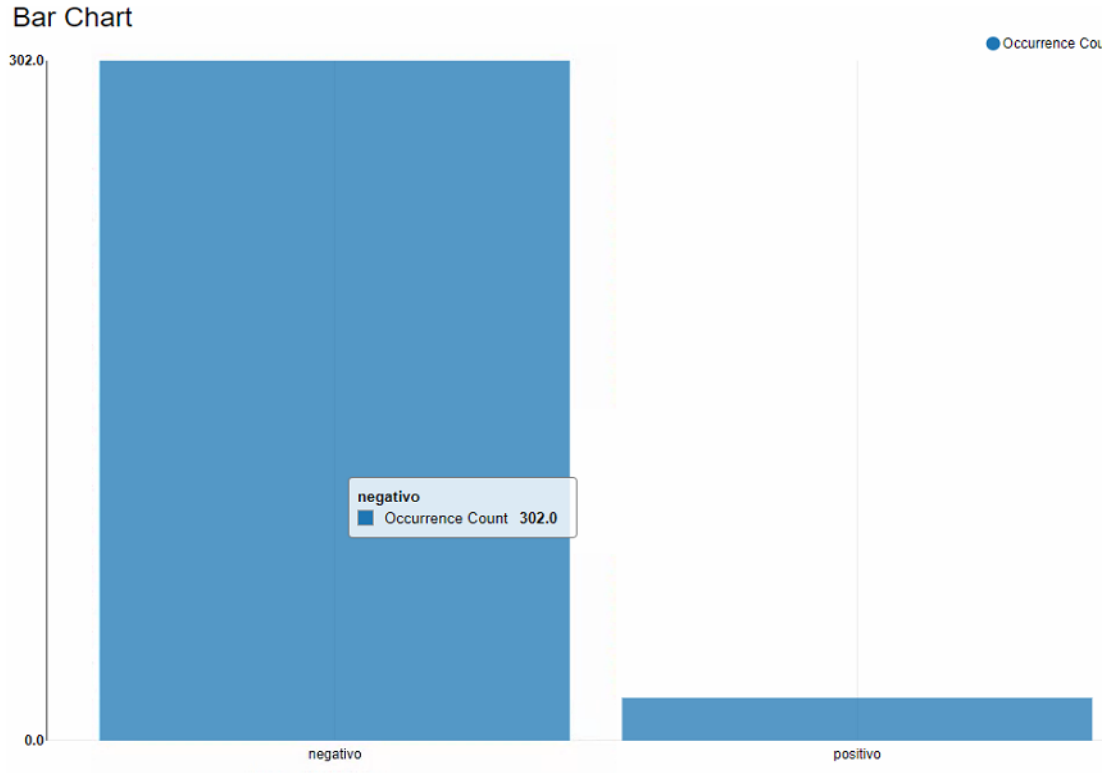

Figura 15. Representación gráfica de los datos en Knime

4.3.2.4 Modelo de Evaluación de calidad de software de Knime

En la tabla 15, se observa la calificación en el modelo de calidad de software concedida a la herramienta Knime, los valores de calificación asignados a cada factor corresponden a una previa experimentación con los datos de evaluación sobre la herramienta.

Tabla 15. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de knime

Factores	Calificación	Descripción
Flexibilidad	9	Permite la manipulación de los datos, algoritmos, procesos
Portabilidad	9	Es compatible con más de un sistema operativo
Interoperabilidad	9	Tiene conexión directa con Twitter
Manual técnico	9	Descripción completa de los complementos del software
Manual de usuario	9	Basta información en línea y consta de un manual integrado en la plataforma
Ayuda en línea	7	Existe ejemplos, blogs, video, tutoriales
Fácil de instalar	9	Proceso de instalación guiado
Fácil de configurar	9	Proceso de configuración guiado
Amigable	8	El software no es muy fácil de comprender, pero sus procesos son más complejos
Íconos con ayuda	8	Consta con ayuda en todas sus herramientas
Seguridad	8	Si brinda seguridad de datos
Actualizaciones	7	Complejo para agregar actualizaciones
Soporte técnico	7	El soporte técnico se brinda en línea
Independencia de hardware	7	Completamente compatible, aunque la plataforma pide requerimientos mínimos para operar correctamente

4.3.2.5 Modelo de Evaluación de funcionabilidad de software de Knime

La tabla 16, describe la calificación en el modelo de funcionabilidad de software concedida a la herramienta Knime, los valores de calificación asignados a cada factor corresponden a una previa experimentación con los datos de evaluación sobre la herramienta.

Tabla 16. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de knime

Factores	Calificación	Descripción
Procesamiento de Datos	3	Es complicado el pre procesamiento de datos , es necesario varios bloques de procesos para limpiar los datos
Modelos De análisis de sentimiento	4	Permite configura del algoritmo, pero no tiene todos los tipos de kernel solo consta de tres
Métodos de aprendizaje de datos	4	Tiene cierta dificultad la limpieza de datos de aprendizaje, se requiere más bloque de procesos
Validación del Modelo	5	Valores precisos
Graficación	5	Es necesario un bloque de traficación, y uno para mostrar los resultados

4.3.3 Alteryx.

4.3.3.1 Modelado

Una de las principales ventajas de este software, es su interfaz gráfica muy intuitiva para el usuario, sus herramientas con ayuda en línea hacen que el modelado en esta plataforma reduzca el tiempo de trabajo, esto potenciado gracias al acceso directo a sus componentes de análisis. En la figura 16, se describe el modelo de Alteryx, donde en primer lugar se lee el archivo, para luego pasar estos datos por filtros sencillos de configurar para cumplir con la limpieza de datos, se invierte la matriz dividiéndola en palabras relevantes, para a continuación realizar el proceso de *stopword*, que se realiza mediante una unión del conjunto de conectores y la matriz de palabras relevantes; posteriormente se filtra la matriz resultante con el propósito de encontrar la matriz que necesita el algoritmo para clasificar los datos; luego se dividen los datos para aprendizaje y pruebas; como última parte se ocupa el algoritmo, para lo cual es necesario colocar visores y un bloque que muestre los resultados.

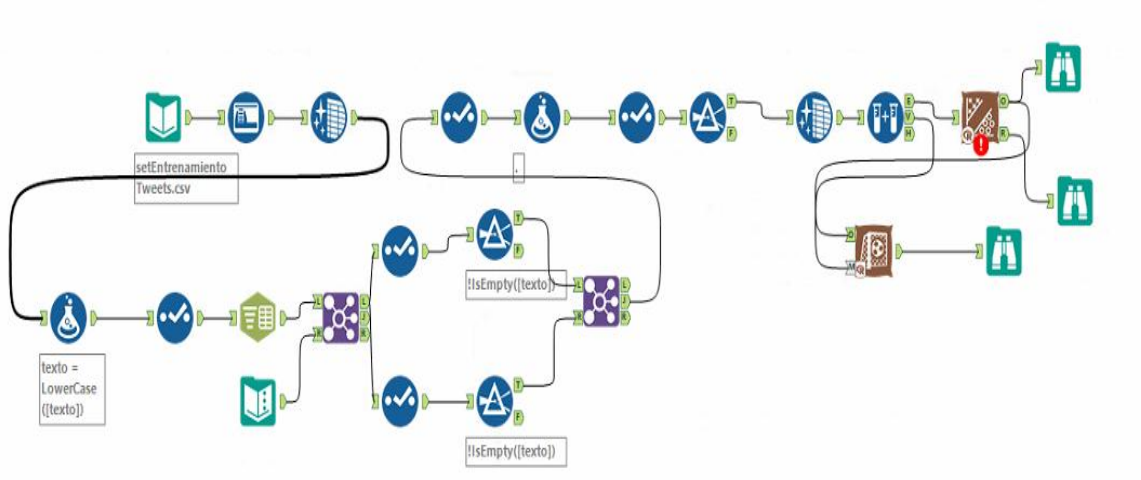

Figura 16. Modelado de Alteryx

4.3.3.2 Resultados

4.3.3.2.1Matriz de confusión

En la tabla 17, se presenta la matriz de confusión, donde 2713 es el número de (TP), 334 representa a (FP), (FN) en este caso 173, finalmente (TN) que son 404.

Tabla 17. Matriz de confusión de Alteryx

	Condición positivo	Condición negativo
Predicción positivo	2713	334
Predicción negativo	173	404

En la tabla 18, se pueden observar los resultados *Recall, Precision, F-Measure, Kappa*; obtenidos a partir de los valores de la matriz de confusión, en ella se describe cada uno de las variables obtenidas, y la fórmula para obtener dicho valor

Tabla 18. Resultados Recall, Precision, F-Measure, Kappa de Alteryx

	Valor	Descripción
Recall	94,01	Es el porcentaje de los que en realidad son positivos, cuantos al final fueron clasificados como positivo (TP/TP+FN)
Precision	89,04	Es el porcentaje de los clasificados como positivos, cuales al final realmente son positivos (TP/TP+FP)
F-Measure	91,45	Media armónica entre Recall y Precision 2*(Recall * Precision) / (Recall + Precision)
Kappa	0,53	El grado de concordancia es **moderado**

4.3.3.3 Gráfico de resultados de Alteryx

En la figura 17, se presenta el gráfico obtenido del análisis de la herramienta Alteryx, donde se muestra que existe una significativa diferencia de polaridades, es importante señalar que la herramienta genera automáticamente esta gráfica como resultados del algoritmo SVM; con el 87.03% de tweets negativos y el 12.97% de publicaciones, en la red social Twitter, positivas.

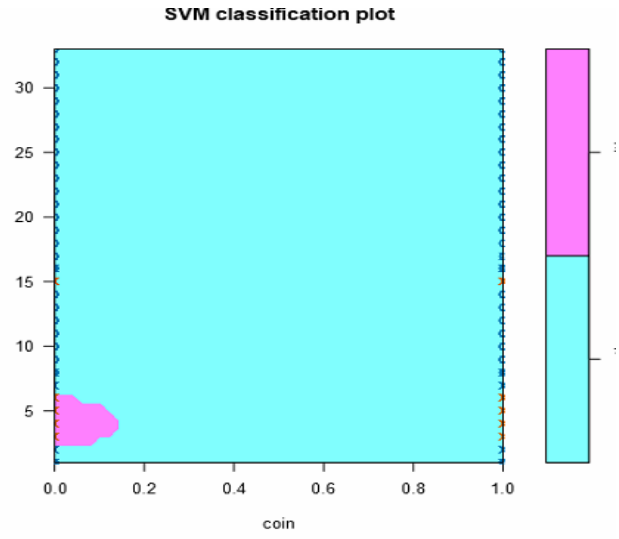

Figura 17. Representación gráfica de los datos en Alteryx

4.3.3.4 Modelo de Evaluación de calidad de software

En la tabla 19, se observa la calificación en el modelo de calidad de software concedida a la herramienta Alteryx, los valores de calificación asignados a cada factor corresponden a una previa experimentación con los datos de evaluación sobre la herramienta.

Tabla 19. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Alteryx

Factores	Calificación	Descripción
Flexibilidad	9	Permite la manipulación de los datos, algoritmos, procesos
Portabilidad	9	Es compatible con más de un sistema operativo
Interoperabilidad	8	No tiene conexión directa con Twitter, es compatible con R
Manual técnico	8	Descripción no detallada de los complementos del software
Manual de usuario	8	Información relacionada en línea no completa
Ayuda en línea	6	Existe pocos ejemplos, blogs, video, tutoriales
Fácil de instalar	7	Proceso de instalación guiado, se torda difícil descargar le herramienta correcta ya que Alteryx brinda distintas soluciones
Fácil de configurar	9	Proceso de configuración guiado
Amigable	9	El software muy intuitivo, se hace fácil su manejo y comprensión por la clasificación grafica de sus herramientas
Íconos con ayuda	9	Consta con ayuda en todas sus herramientas
Seguridad	8	Si brinda seguridad de datos
Actualizaciones	6	Es difícil actualizar nuevas funciones al software, ya que hay distintas solones ofrecidas por el desarrollador, y no todas las distribuciones de Alteryx brindan las mimas herramientas, se descargó dos tipos de herramientas de la página oficial, ya que uno de los software descargados no contaba con las herramientas de predicción
Soporte técnico	9	Consta de contactos para soporte técnico
Independencia de hardware	7	Completamente compatible, aunque la plataforma pide requerimientos mínimos para operar correctamente

4.3.3.5 Modelo de Evaluación de funcionabilidad de software

La tabla 20, describe la calificación en el modelo de funcionabilidad de software concedida a la herramienta Alteryx, los valores de calificación asignados a cada factor corresponden a una previa experimentación con los datos de evaluación sobre la herramienta.

Tabla 20. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de Alteryx

Factores	Calificación	Descripción
Procesamiento de Datos	3	Es complicado el pre procesamiento de datos , es necesario varios bloques de procesos para limpiar los datos
Modelos De análisis de sentimiento	4	Permite configura el algoritmo, de una manera gráfica
Métodos de aprendizaje de datos	4	Permite la limpieza de todos por medio de bloques fáciles de configurar
Validación del Modelo	4	Solo muestra la matriz de confusión
Graficación	5	Es necesario un bloque de graficación

4.3.4 SAP

4.3.4.1 Modelado

La herramienta SAP se compone de dos partes para el análisis de datos: SAP HANA Studio, que corre sobre la plataforma Eclipse y SAP HANA Data Base *express edition*, que se puede descargar desde la página de SAP, previamente registrando un usuario *developer*. La versión de SAP HANA *express edition*, es una plataforma de datos en memoria, para el correcto funcionamiento de la plataforma; la base de datos requiere 24 GB de memoria RAM, un *hard drive disk* (HDD) 120GB y 4 núcleos de procesamiento; otras versiones de prueba ya no están disponibles por el motivo que desde el lanzamiento de SAP HANA, SAP no tiene conexión con otras bases de datos y es estrictamente necesaria a integración con SAP HANA Data Base, por ultimo esta SAP *Academy*, para ingresar en las versiones de prueba de esta distribución, es necesario obtener un usuario

tipo S o súper administrador; la obtención de ese usuario pagar por la licencia. La plataforma usa código plano en lenguaje SQL, puesto que todo análisis es realizado en su propia base de datos. En la figura 18, se indica la distribución SAP, una vez configurada.

Figura 18. Consola SAP Hana

4.3.4.2 Resultados

No se evaluó esta herramienta por motivos de licenciamiento, se probó con versiones anteriores como: MINISAP, sin resultados, porque desde el lanzamiento de SAP HANA, no se puede conectar a otra base de datos, que no sea SAP HANA Data Base *express edition,* la cual tiene una versión de prueba, pero es necesario muchos recursos de hardware; para el funcionamiento de SAP Studio es necesario SAP HANA Date base.

4.3.4.3 Modelo de Evaluación de calidad de software de SAP

En la tabla 21, se observa la calificación en el modelo de calidad de software concedida a la herramienta SAP, los valores de calificación asignados a cada factor corresponden a una previa experimentación con de la herramienta.

Tabla 21. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAP

Factores	Calificación	Descripción
Flexibilidad	0	No es posible evaluar
Portabilidad	8	Es compatible con más de un sistema operativo, trabaja conjuntamente con el software desarrollador de Java Eclipse
Interoperabilidad	0	No es posible evaluar
Manual técnico	6	Pocos detalles de complementos del software
Manual de usuario	8	Poca información relacionada en línea
Ayuda en línea	5	Existe muy pocos ejemplos, blogs, video, tutoriales
Fácil de instalar	5	Es necesario instalar Java, el desarrollador Eclipse, luego adjuntar el complemento SAP Hana, SAP HANA Date base necesita de muchos recursos de software
Fácil de configurar	8	Proceso de configuración guiado
Amigable	4	El software es confuso no es por bloques de conexión es por líneas de código
Íconos con ayuda	0	No es posible evaluar
Seguridad	0	No es posible evaluar
Actualizaciones	6	Existe muy poca información sobre actualizaciones y ayuda de la plataforma
Soporte técnico	8	Consta de soporte técnico en su página oficial
Independencia de hardware	3	Requiere altos recursos para su correcto funcionamiento

4.3.4.4 Modelo de Evaluación de funcionabilidad de software de SAP

La tabla 22, describe la calificación en el modelo de funcionabilidad de software concedida a la herramienta SAP, esta herramienta no pudo ser evaluada, ya que el software no pudo ser instalado, ni probado; para este modelo por cuestiones de licenciamiento y recursos mínimos que exige el programa para su correcto funcionamiento.

Tabla 22. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAP

Factores	Calificación	Descripción
Procesamiento de Datos	0	No es posible evaluar
Modelos De análisis de sentimiento	0	No es posible evaluar
Métodos de aprendizaje de datos	0	No es posible evaluar
Validación del Modelo	0	No es posible evaluar
Graficación	0	No es posible evaluar

4.3.5 SAS

4.3.5.1 Modelado

La configuración de esta herramienta es muy diferente al resto de otras plataformas, puesto que SAS, en su distribución de prueba no ofrece una solución de flujo de bloques; la distribución SAS *Enterprise Miner* permite el modelado por bloque, pero esta herramienta necesita de una licencia pagada; para el análisis se usará SAS Viya, la cual es online y tiene una versión de prueba por 15 días. Para cumplir el modelo general, la figura 19, se muestra el menú principal de la herramienta, en la que se ocupará las opciones: Administrar datos, Preparar datos, Explorar y visualizar datos y Construir modelos.

Figura 19. Menú principal de SAS Viya

4.3.5.1.1 Administrar Datos

Como muestra la figura 20, esta función del software es la importación de los datos. Es muy importante señalar que, por cuestiones de licenciamiento, esta herramienta solo permite la manipulación de hasta 100 datos, por este motivo la matriz de *tweets* de entrenamiento se ha cortado para cumplir con este número de datos permitidos.

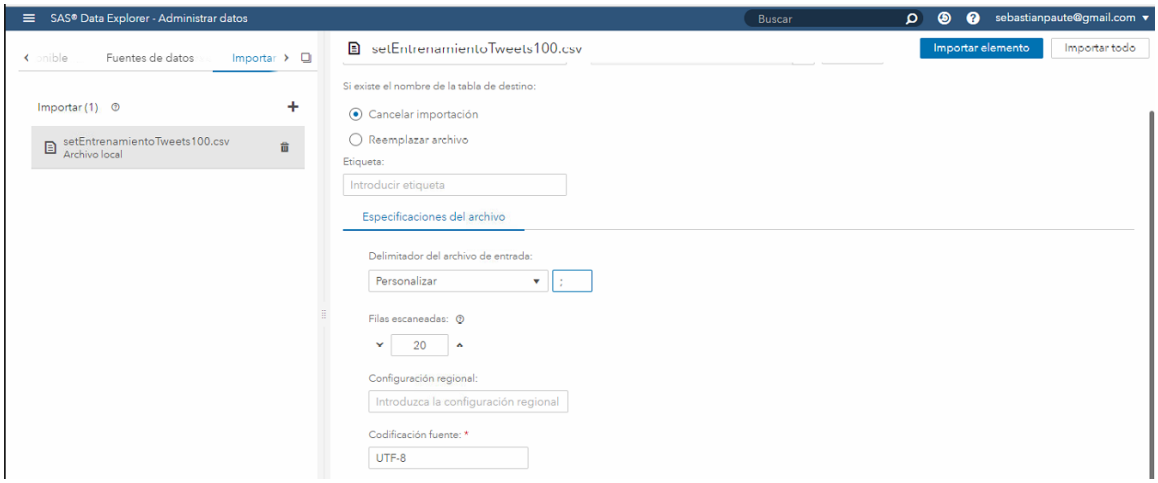

Figura 20. Administrar datos de SAS Viya

4.2.5.1.2Preparar datos

Esta función es la de limpieza de datos, la cual permite crear reglas de pre procesamiento, la figura 21, muestra el primer paso, convertir todo el texto a minúsculas, la parte de *stopword*, no se puede realizar en esta distribución, porque la aplicación no permite subir un diccionario propio, tampoco consta de un diccionario en español. Como se indica en la figura 22, los datos al no ser limpiados correctamente generan predictores de irrelevantes, y un volumen de datos para evaluar superior a lo permitido por la aplicación; haciendo que esta falle y se cierre, como se muestra en la figura 23.

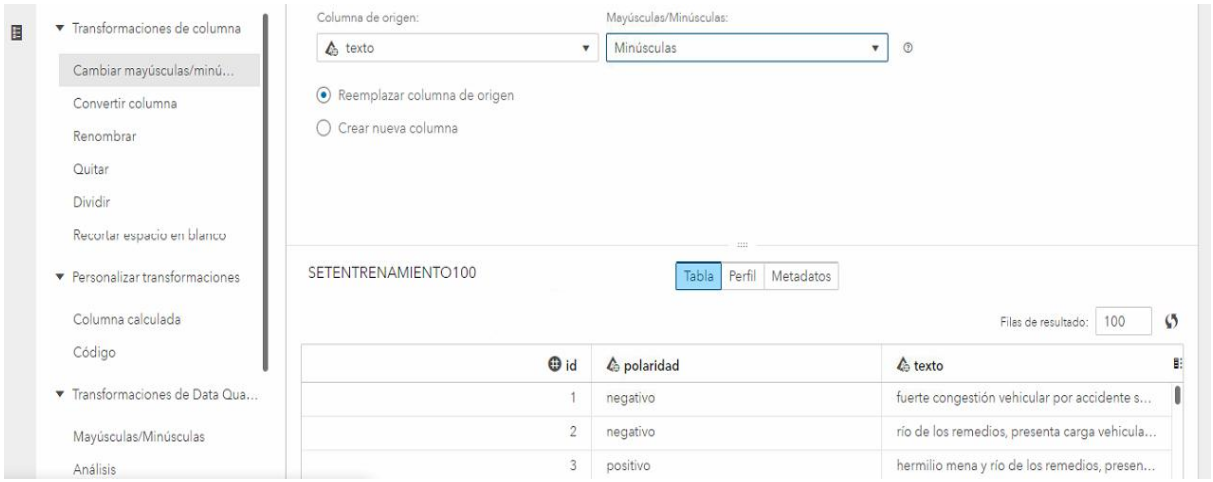

Figura 21. Convertir en minúsculas de SAS Viya

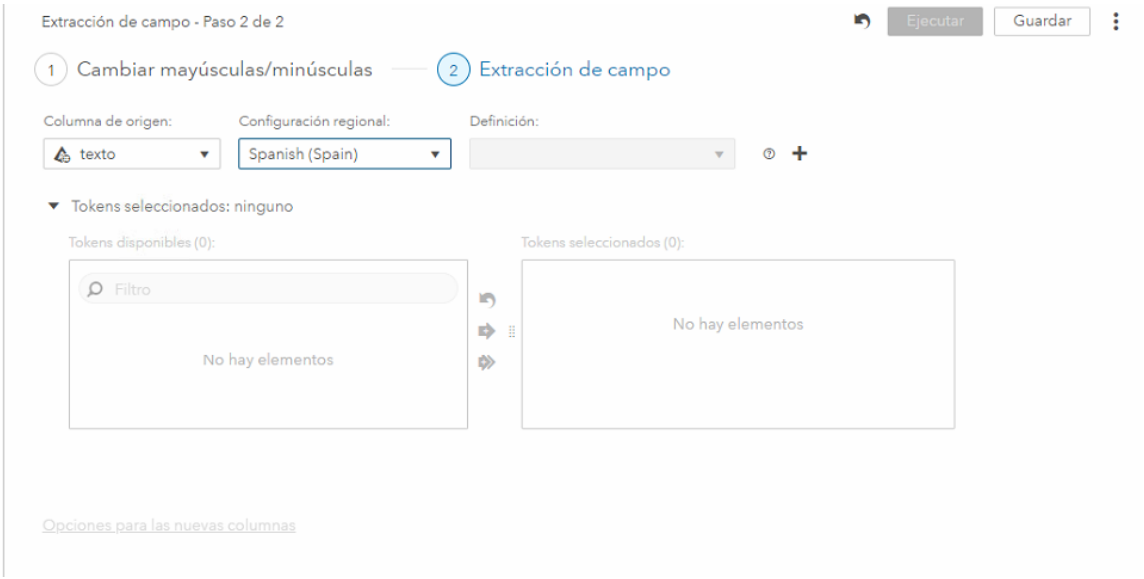

Figura 22. Stopword de SAS Viya

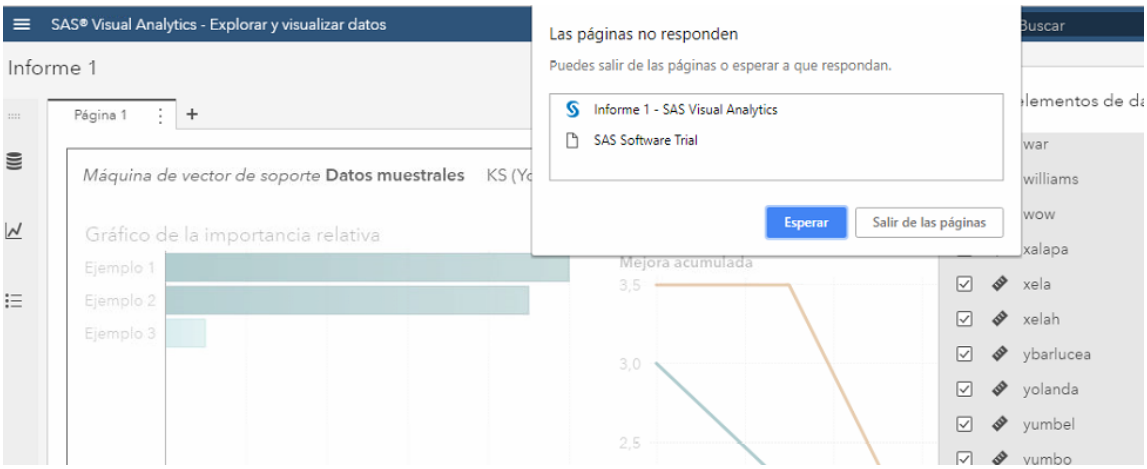

Figura 23. Error de desbordamiento de SAS Viya

Para resolver este problema y analizar la herramienta, se utilizan datos pre procesados por la aplicación Rapidminer; seguidamente estos datos pre procesados se cargan en la plataforma SAS directamente a la función de explorar y visualizar datos.

4.2.5.1.3 Explorar y visualizar datos

Esta función permite la parte de pre procesamiento *tokenizer*, también manipular los datos que van a ser procesados de una manera gráfica y muy sencilla, además en esta parte se configura las variables el algoritmo SVM como muestra la figura 24:

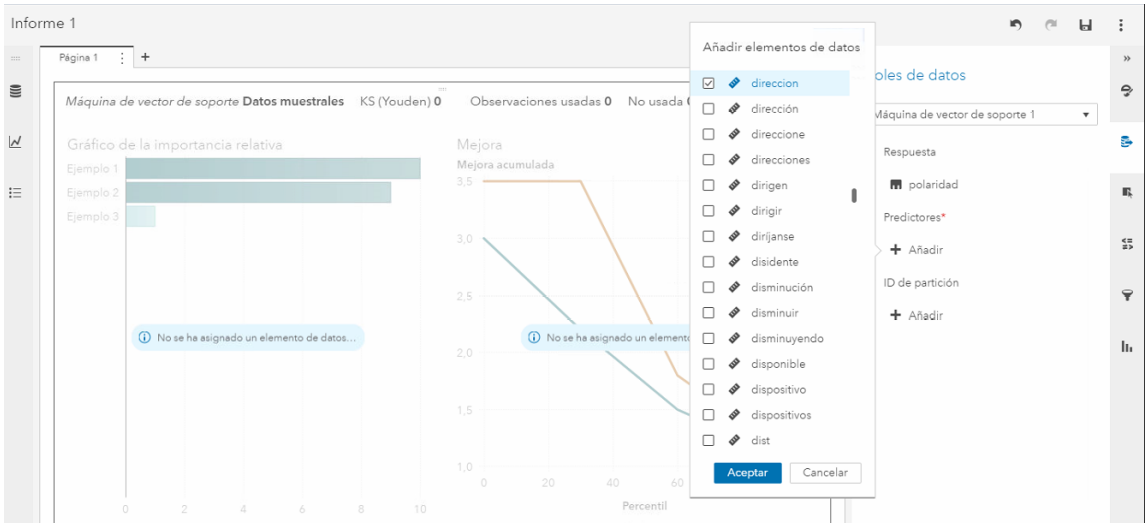

Figura 24. Configuración del algoritmo SVM de SAS Viya

4.2.5.1.4 Construir Modelos

Por último, se construye el modelo como muestra la figura 25, esta aplicación no permite opciones de graficación complestas de los resultados del análisis, es sencillo el modelo describe la entrada de datos, la limpieza de datos o pre procesamiento, la aplicación del algoritmo y resultados finales.

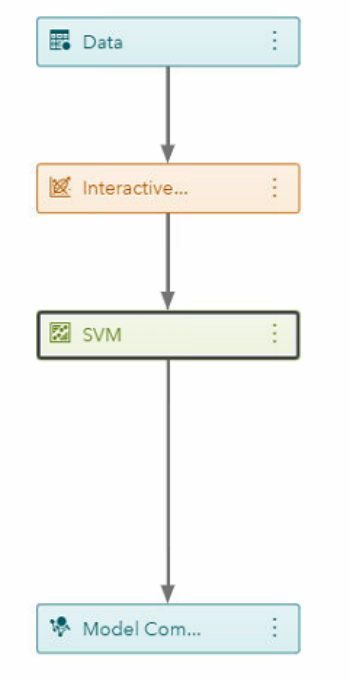

Figura 25. Modelo de solución de SAS Viya

4.3.5.2 Resultados

4.3.5.2.1 Matriz de confusión de SAS

En la tabla 23, se presenta la matriz de confusión donde 17 es el número de (TP), 0 representa a (FP), (FN) en este caso 1, finalmente (TN) que son 82. Es importante resaltar que esta herramienta se probó con 100 *tweets* pre procesados, ya que la versión de prueba de la plataforma no permite más datos.

Tabla 23. Matriz de confusión de SAS

	Condición positivo	Condición negativo
Predicción positivo	17	0
Predicción negativo	1	82

La tabla 24, contiene los resultados *Recall, Precision, F-Measure, Kappa*. Obtenidos a partir de los valores de la matriz de confusión, en ella se describe cada uno de las variables obtenidas, y la fórmula para obtener dicho valor

Tabla 24, Resultados Recall, Precision, F-Measure, Kappa de SAS

	Valor	Descripción
Recall	94,44	Es el porcentaje de los que en realidad son positivos, cuantos al final fueron clasificados como positivo (TP/TP+FN)
Precision	100,00	Es el porcentaje de los clasificados como positivos, cuales al final realmente son positivo (TP/TP+FP)
F-Measure	97,14	Media armónica entre Recall y Precision 2*(Recall * Precision) / (Recall + Precision)
Kappa	0,97	El grado de concordancia es **casi perfecto**

4.3.5.3 Gráfico de resultados de SAS

La figura 26, indica los resultados obtenidos de la experimentación con la plataforma SAS; se observa una mayor presencia de polaridad negativa, con el 82.82% y el 17.18% restante, son texto positivo, de los *tweets* ingresados para su análisis. La figura 26, corresponde a la evaluación de 100 datos.

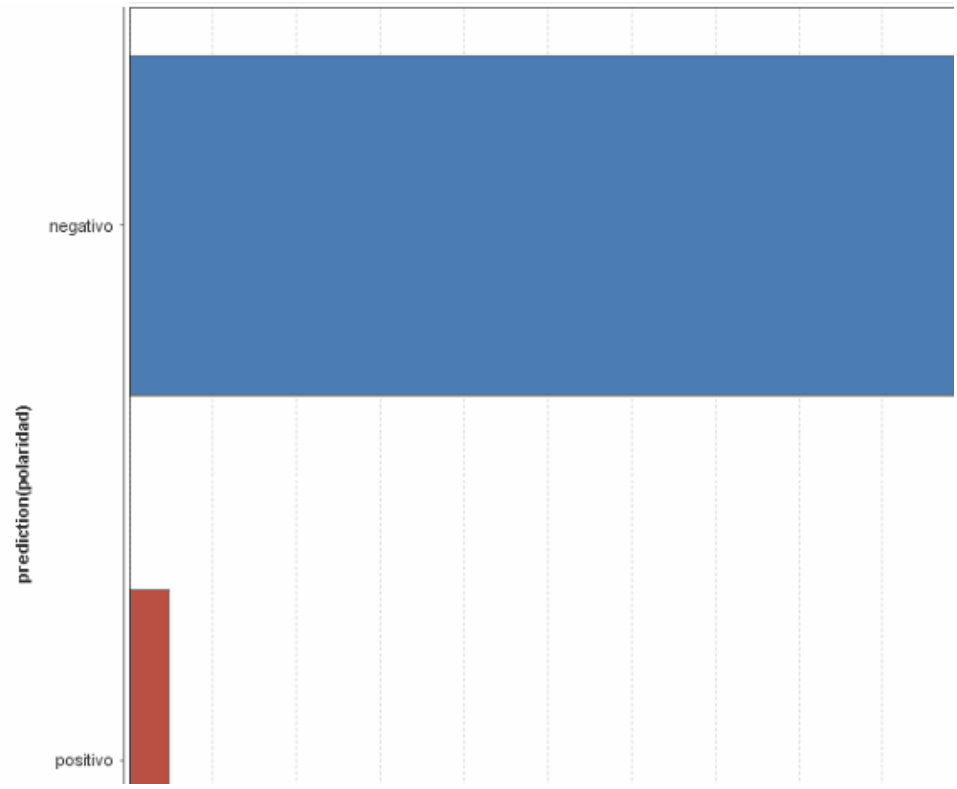

Figura 26. Gráfico de resultados de SAS Viya

4.3.5.4 Modelo de Evaluación de calidad de software de SAS

En la tabla 25, se observa la calificación en el modelo de calidad de software concedida a la herramienta SAS, los valores de calificación asignados a cada factor corresponden a una previa experimentación con de la herramienta.

Tabla 25. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAS

Factores	Calificación	Descripción
Flexibilidad	9	Permite la manipulación de los datos, algoritmos, procesos
Portabilidad	9	Es compatible con más de un sistema operativo, está en internet
Interoperabilidad	9	Tiene conexión directa con Twitter
Manual técnico	7	Pocos detalles de complementos del software
Manual de usuario	7	Poca información relacionada en línea
Ayuda en línea	6	No existe ejemplos claros de la aplicación ya que SAS cuenta con varias distribuciones de software, las cuales son dedicadas a problemas específicos y no existe una experimentación grande con esta herramienta, ya que la misma tiene costos de licenciamiento
Fácil de instalar	6	La distribución de SAS ocupada para la evaluación por cuestiones de licenciamiento no se puede instalar directamente sobre la máquina física, será probada en línea
Fácil de configurar	6	no permite configuración de la plataformas
Amigable	6	El software no es intuitivo, se maneja como una página web, con menús desplegables
Íconos con ayuda	7	consta con poca ayuda en todas sus herramientas
Seguridad	8	si brinda seguridad de datos
Actualizaciones	6	existe muy poca información y ayuda de la plataforma
Soporte técnico	7	consta de contactos para soporte técnico en su página oficial, pero solo para la versión de pago
Independencia de hardware	7	Está en la red

4.3.5.5 Modelo de Evaluación de funcionabilidad de software de SAS

La tabla 26, describe la calificación en el modelo de funcionabilidad de software concedida a la herramienta SAS, los valores de calificación asignados a cada factor corresponden a una previa experimentación con los datos de evaluación sobre la herramienta.

Tabla 26. Descripción de la calificación obtenida por la herramienta en el modelo de Evaluación de calidad de software de SAS

Factores	Calificación	Descripción
Procesamiento de Datos	2	Limitación en la cantidad de datos por licenciamiento
Modelos De análisis de sentimiento	5	Tiene procesos básicos pre armados, permite la configuración del algoritmo
Métodos de aprendizaje de datos	4	Permite la limpieza de datos, excepto stopword, no analiza datos en español
Validación del Modelo	4	Se ocupó solo una parte de todos los Tweets por no tener la licencia profesional
Graficación	5	Los gráficos se generan automáticamente

4.4 Conclusión

En base a la experimentación, se tornó difícil realizar el modelado de procesos, pues las herramientas presentan una gran gama de funciones; así mismo, es importante resaltar que no todas las herramientas cuentan con los mismos kernel, con configuración lineal del algoritmo, para lo cual se ha implementado la configuración de kernel radial. Posterior a la experimentación se procede al análisis de los resultados, con los criterios establecidos.

CAPÍTULO V: ANÁLISIS DE RESULTADOS

5.1 Evaluación de calidad de software

En lo que respecta a la calidad de software la plataforma Rapidminer es la que obtuvo mayor puntaje, a partir de lo cual puede deducirse que esta es la mejor, esto gracias a que el programa es bastante amigable e intuitivo para el usuario, a diferencia de las otras herramientas evaluadas. Con un puntaje total de 888, esta herramienta lidera la evaluación de calidad de software, como se detalla en la tabla 27.

Tabla 27. Tabla de puntuación en las herramientas en calidad de software

	Herramientas de *Text Mining*					
Factores	Rapidminer	Knime	Alteryx	SAP	SAS	Peso
Flexibilidad	9	9	9	0	9	8
Portabilidad	9	9	9	8	9	2
Interoperabilidad	9	9	8	0	9	10
Manual técnico	9	9	8	6	7	5
Manual de usuario	9	9	8	8	7	7
Ayuda en línea	8	7	6	5	6	15
Fácil de instalar	9	9	7	5	6	4
Fácil de configurar	9	9	9	8	6	5
Amigable	10	8	9	4	6	20
Íconos con ayuda	8	8	9	0	7	5
Seguridad	8	8	8	0	8	2
Actualizaciones	9	7	6	6	6	10
Soporte técnico	9	7	9	8	7	2
Independencia de hardware	7	7	7	3	7	5
Total	888	809	783	408	688	

5.2 Evaluación de la funcionabilidad de software

En la tabla 28, se puede evidenciar que nuevamente la herramienta Rapidminer obtiene mayor puntuación en relación a las otras plataformas evaluadas, esto se debe a que el algoritmo es completo al momento de configurar sus variables por el motivo de que cuenta con una amplia configuración de kernel, los tweets, así también, la validación del modelo es confiable.

Tabla 28. Puntuación en las herramientas en funcionabilidad de software

		Herramientas de Text Mining				
Factores	Rapidminer	Knime	Alteryx	SAP	SAS	Peso
Procesamiento de Datos	4	3	3	0	2	15
Modelos De análisis de sentimiento	5	4	4	0	5	20
Métodos de aprendizaje de datos	4	4	4	0	4	20
Validación del Modelo	5	5	4	0	4	30
Graficación	5	5	5	0	5	15
Total	465	430	400	0	405	

5.3 Evaluación de métricas estadísticas para determinar la exactitud del modelo

Después de analizar los resultados y compararlos en la tabla 29. Se concluye que el modelo más exacto y de mayor precisión pertenece a la herramienta Rapidminer, es importante recordar que los resultados de la herramienta SAS, pertenecen al análisis de solo 100 datos pre procesados.

Tabla 29. Tabla de métricas estadísticas para determinar la exactitud del modelo

	Herramientas de Text Mining				
Factores	Rapidminer	Knime	Alteryx	SAP	SAS
Recall	95,32	89,63	94,01	0	94,44
Precision	90,55	76,05	89,04	0	100,00
F-Measure	92,88	82,28	91,45	0	97,14
Kappa	0,61	0,78	0,53	0	0,97

5.4 Conclusión

Rapidminer se define como la mejor herramienta, puesto que en la evaluación de calidad, su puntaje es de 888 puntos sobre 1000 ; esto se debe, principalmente a lo amigable que es el software y dado que es muy intuitivo; de igual forma, su interfaz y procesos predefinidos ayudan al usuario a agilitar la minería de texto, reduciendo el tiempo de modelado de procesos frente a las otras herramientas que se evaluaron; donde Knime obtuvo una calificación de 809, la herramienta Alteryx 783 puntos, la plataforma SAP 408 puntos; esta calificación baja del programa, es por esta razón que no se evaluó en su totalidad; pues se intentó instalar todas sus versiones de prueba, pero todas necesitan de la base de datos de SAP, la cual no se instaló por motivo de excesivos requerimientos del software. Por último la herramienta SAS, con una calificación de 688 puntos de esta herramienta; es importante resaltar que se experimentó con una versión en línea con muy pocas opciones de minería de texto.

De igual forma, como en el modelo anterior, Rapidminer es líder en la evaluación de funcionabilidad, con un valor de 465 puntos, gracias a que la veracidad de su modelo es bastante alta. Consta de una gran cantidad de gráficas para mostrar los resultados, la limpieza de datos es fácil de configurar y hacer; las configuraciones de su algoritmo son bastante completas y fáciles de definir frente a las otras herramientas que se emplearon; donde Knime, obtuvo una calificación de 430, el programa Alteryx 400 puntos, la plataforma SAP 0 puntos, dado los problemas de instalación. Finalmente la plataforma SAS en su versión de prueba *online* con una calificación de 405 puntos.

En cuanto a la métrica *de F- measure*, se observa que es superior, con un valor de 92,88. La facilidad del uso de sus bloques de proceso, ayuda a que la limpieza sea más eficaz y rápida; esto se logra gracias a que sus bloques son bastante intuitivos, frente a las otras herramientas que se probaron, donde Knime obtuvo una calificación de 82,28, el programa Alteryx 91,45, la plataforma SAP 0, por el motivo que esta herramienta no se evaluó. Finalmente la plataforma SAS obtuvo un valor de *f- measure* 97,14. Es importante resaltar que este valor se obtuvo de la evaluación de 100 datos pre procesados en Rapidminer, por tal motivo el valor es elevado.

Otra característica que hace de Rapidminer una herramienta líder en la evaluación, es la capacidad y facilidad de integrar otras librerías desde su buscador; el campo de graficación permite elegir una gran variedad de modelos, sin necesidad de agregar otro bloque de datos o un proceso *plot*. La cantidad de información, ejemplos en blogs de otros usuarios y en la página oficial de la herramienta, permite que los problemas sean solucionados de manera más rápida. La documentación en línea es vasta en comparación a las otras herramientas.

CAPÍTULO VI: VENTAJAS Y DESVENTAJAS DE HERRAMIENTAS

6.1 Tabla general de ventajas

La tabla 30 describe en general las ventajas de cada una de las herramientas, estas ventajas se obtuvieron luego de la experimentación y manejo de las plataformas.

Tabla 30. Ventajas de herramientas

	Herramientas				
Ventajas	Rapidminer	Knime	Alteryx	SAP	SAS
variedad de algoritmos	X				
fácil integración de datos	X	X			X
fácil pre procesamiento de datos	X				
Entorno grafico amigable	X	X	X		
Ayuda en la red (ejemplos, documentación, videos)	X	X			
fácil instalación	X				
integración con otras plataformas	X	X	X	X	
utilización online					X

6.2 Tabla general de desventajas

La tabla 31, describe de forma general las desventajas de las herramientas utilizadas en el análisis.

Tabla 31. Desventajas de herramientas

	Herramientas				
Desventajas	Rapidminer	Knime	Alteryx	SAP	SAS
Muy poca ayuda en la red				X	X
Difícil compresión de funcionamiento				X	X
Difícil instalación y configuración				X	X
Problemas con licencias			X	X	X
Falta de opciones de configuración del algoritmo		X			
cantidad límite de datos para procesar	X		X	X	X

6.3 Descripción detallada de ventas y desventajas

6.3.1 Rapidminer

Ventajas:

- Extensión de la plataforma: la herramienta cuenta con una gran cantidad de algoritmos, fácil integración de fuentes de datos, capacidad de modelado flexible, fácil preparación y limpieza de datos, esta plataforma no es dedicada a una sola área de minería.
- Facilidad de entorno gráfico: soluciones pre-canalizadas, desarrollo de modelos, facilidad de aprendizaje, velocidad en desarrollo de modelos, análisis avanzados, variedad de herramientas y complementos de lenguaje de programación.
- Extensa cantidad de ayuda en la web: soporte al usuario, blog, foros, ejemplos descargables.
- Muy fácil descarga e instalación del software.

Desventajas:

- Cantidad límite de datos procesados en su versión libre, Rapidminer soluciona este inconveniente con la aplicación de licencias.
- Problemas con ejemplos incompletos o muy básicos.

6.3.2 Knime

Ventajas:

- Plataforma de código abierto lo cual la hace flexible, tiene una gran apertura y extensibilidad.
- Gran capacidad de manejo de datos: mezcla de datos, verificación de calidad de datos, modificación de datos (limpieza, partición, generación de características, agregación de valores).
- Ejemplos y ayuda en la red por parte de los clientes de esta plataforma.

Desventajas:

- Kernel de algoritmos insuficientes para comparación de herramientas, solo se pueden configurar los más usados por los clientes.
- Gestión de bloques de modelado: al ser simples, es necesario ocupar varios para un proceso largo; lo cual genera que el modelo incremente su tamaño y sea confuso por la cantidad de líneas y bloques.
- Ejemplos muy confusos, por la cantidad de bloques, falta de documentación en línea sobre el programa que permita la ayuda a nuevos clientes.

6.3.3 Alteryx

Ventajas:

- Software muy intuitivo, con una interfaz de usuario muy buena y fácil de modelar, íconos de pre procesamiento que agilitan el modelado.
- La ayuda e información técnica de la plataforma son de fácil acceso.
- Facilidad de aprendizaje, manejo de datos, modificación de datos, presentación de resultados y graficación.

Desventajas:

- Difícil de instalar, consta de una gran gama de soluciones dedicadas, para lo cual se tiene que pedir específicamente una licencia, si se busca una

solución; pues si por error se descarga otra (o la más básica) no permite actualizar la plataforma ni descargar complementos.

- Tiempo de espera por una licencia: al descargar una solución de predicción de datos desde la página oficial, Alteryx se toma un tiempo para emitir la licencia de prueba, la cual es notificada al correo para posteriormente, con un link de redirección, poder abrir otra página donde se pide el registro del cliente y ahí sí poder pasar a la página de descarga del software.
- Muy poca documentación, ejemplos o blogs, lo que se debe a la cantidad de soluciones dedicadas que ofrece la plataforma.

6.3.4 SAP

Ventajas:

- Integración con otras plataformas como Eclipse, R.
- Capacidad de soporte asistido y construcción de modelos automatizados.

Desventajas:

- Difícil instalación, ya que necesita de otros programas para funcionar, como Eclipse.
- Difícil programación por falta de ejemplos claros, esto se debe a que SAP ofrece muchas soluciones en diversos campos del aprendizaje automático.
- Dificultades en la comprensión de funcionamiento, a la vez que muy poco intuitivo.
- Muy poca documentación y ayuda en línea.
- Se necesita muchos recursos del software para poder probar la herramienta.

6.3.5 SAS

Ventajas:

- Portabilidad: al estar en la nube, es fácil de acceder a la plataforma desde cualquier parte.

- Datos en la nube.

Desventajas:

- Muy poca documentación, difícil de modelar.
- Gran cantidad de soluciones dedicadas.

CAPÍTULO VII: ANÁLISIS DE SENSIBILIDAD CON LA MEJOR HERRAMIENTA

La herramienta con mejor puntaje es *Rapidminer* por varias características, principalmente su facilidad de uso; al ser una plataforma muy usada en el mundo, la documentación de la misma esa amplia, así como: foros, ejemplos y videos explicativos; esto hace que aprender a usar el software no se torne complejo.

En la figura 27, se representa el sentimiento de los usuarios cuencanos, así como de entidades de control y del Gobierno, como el ECU 911, que ocupan la red social *Twitter*. En cuanto al tráfico vehicular se sabe que tiene varios factores como: la congestión en horas pico, los accidentes, cierre de vías, trabajos en la urbe, como el tranvía de los 4 ríos de Cuenca, los cuales incrementan el malestar de la población; como este resultado da a conocer.

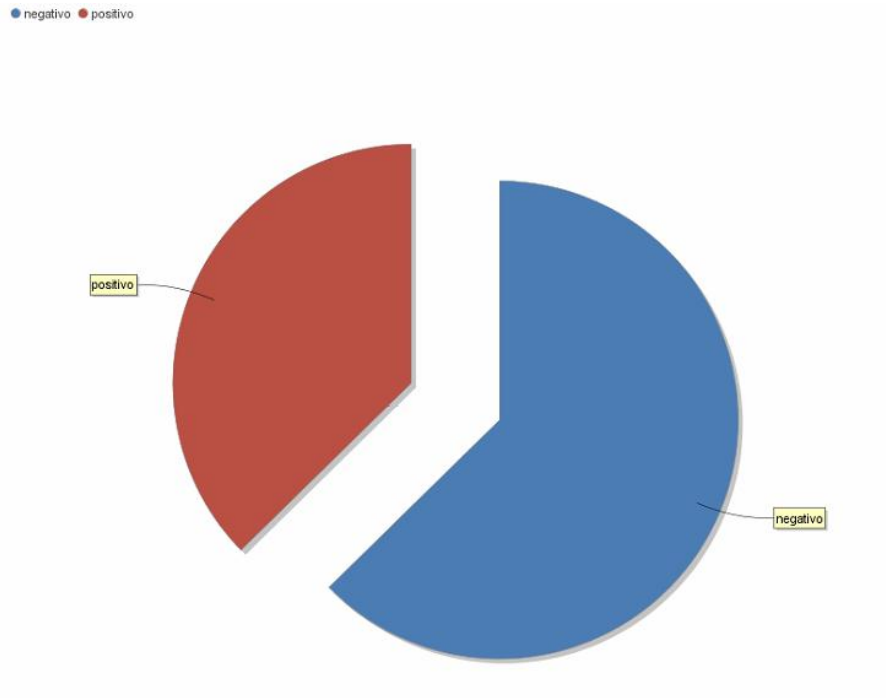

Figura 27. Sentimiento con respecto a la congestión vehicular

CAPÍTULO VIII: CONCLUSIONES

Se analizó, probó y comparó 5 herramientas diferentes de minería de texto en la investigación: Rapidminer, Knime, Alteryx, SAS y SAP; con un conjunto de datos extraídos de Twitter, por medio de la API oficial de la misma empresa, se recopiló 3949 tweets aproximadamente, donde solo 325 datos útiles pertenecen a la ciudad de Cuenca-Ecuador, es importante resaltar esto, porque en la fase principal del proyecto, los tweets de Cuenca - España, dieron problemas en las primeras pruebas.

De igual forma se ha usado el algoritmo support vector machine (SVM), basándose en varios estudios similares de minería de texto sobre una red social, donde se determina que el mejor para estas pruebas es este algoritmo.

Se ha establecido que la mejor herramienta es Rapidminer, puesto que con esta plataforma se obtienen: mejor resultados, más confiables y en menor tiempo; porque al tener una interfaz muy bien estructurada se hace de fácil uso para los usuarios.

El análisis de la herramienta está basado en tres ejes principales, como: la calidad, usabilidad del software y los índices de confiabilidad que son generados propiamente por el software; se determinó ciertas herramientas a partir de otros análisis de software mucho más específicos, corroborando el liderazgo de estas herramientas en sus categorías.

En la parte de usabilidad de software, se pone en consideración que, Alteryx es una buena opción para usuarios que estén empezando a minar texto, porque es una herramienta sencilla compatible con R y muy gráfica, pero no ofrece análisis a profundidad ni datos bastante significativos, en comparación con Rapidminer.

En cuanto a los datos para futuras investigaciones, es necesario ampliar la cantidad de datos de aprendizaje, así se afina y mejora el algoritmo de predicción y los resultados son más específicos y útiles.

Referencias

Alteryx, I. (22 de marzo de 2017). Obtenido de https://www.alteryx.com/

Arias, M. B. (24 de febrero de 2016). Mineria de texto en medios sociales caso de estudio del proyecto tranvía Cuenca. Cuenca, Azuay, Ecuador.

Bannister, K. (10 de 02 de 2015). Obtenido de Entendiendo el análisis de sentimiento: qué es y para qué se usa: https://www.brandwatch.com/es/blog/analisis-de-sentimiento/

Batrinca, B., & Treleaven, P. C. (2014). Social media analytics: a survey of techniques, tools and platforms. *AI and Society*, 89-116.

Bonzanini, M. (06 de 2016). *kdnuggets*. Recuperado el 9 de 2017, de Mining Twitter Data with Python Part 1: Collecting Data: https://www.kdnuggets.com/2016/06/mining-twitter-data-python-part-1.html

Byrd, K., Mansurov, A., & Baysal, O. (2016). Mining twitter data for influenza detection and surveillance. *Proceedings - International Workshop on Software Engineering in Healthcare Systems, SEHS 2016*, 43-49.

Data Flair. (12 de 8 de 2017). *Kernel Functions-Introduction to SVM Kernel & Examples.* Obtenido de https://data-flair.training/blogs/svm-kernel-functions/

David Black, J. T. (22 de 1 de 2016). *Gartner.* Obtenido de https://www.gartner.com/doc/3188318?ref=SiteSearch&sthkw=evaluation%20criteria&fnl=search&srcId=1-3478922254

De Groot, R. (2012). Data Mining for Tweet Sentiment Classification. 63.

Eleonora D'Andrea, P. D. (2015). *Real-Time Detection of Traffic From.*

Fillottrani, P. R. (2007). *Calidad en el Desarrollo de Software, Modelos de calidad de software.* Universidad Nacional del Sur.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 137-144.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information and Management*, 801-812.

He, W., Zha, S., & Li, L. (2013). *International Journal of Information Management*, 464-472.

Hospital Universitario Ramón y Cajal. (2018). Obtenido de Índices de concordancia: http://www.hrc.es/bioest/errores_2.html

Javier Saldarini, C. C. (2017). *Un Modelo de Calidad Mixto como Soporte a la Mejora de los Productos Software con Impacto en los Procesos Organizacionales* . Cordoba: Universidad Tecnológica San Francisco.

José C. Riquelme, R. R. (2015). *Minería de Datos: Conceptos y Tendencias.* Sevilla: Universidad de Sevilla.

Ken Collier, Bernard Carey, Donald Sautter, Curt Marjaniemi. (2015). *A Methodology for Evaluating and Selecting Data Mining Software.* Arizona.

KMINE. (2018). *KNIME para científicos de datos*. Obtenido de https://www.knime.com/knime

Linden, A., Krensky, P., Hare, J., Idoine, C., & Sicular, S. (14 de Febrero de 2017). *Magic Quadrant for Data Science Platforms.* Obtenido de https://www.gartner.com/doc/reprints?id=13TKR16P&ct=170215&st=sg

Mustafa Sofean, M. S. (2012). *A Real-Time Architecture for Detection of Diseases using.*

Ortíz, A., & Martín, M. (2015). *Detección automática de Spam utilizando Regresión Logística.* Jaén, Andalucía: Departamento de Informática universidad de Jaén.

Parikh, R. (2013). ET : Events from Tweets. 613-620.

Piatetsky, G. (2017). *KDnuggets.* Obtenido de https://www.kdnuggets.com/2017/04/forrester-gartner-data-science-platforms-machine-learning.html

Rapidminer. (2018). *Rapidminer*. Obtenido de https://rapidminer.com/products/

Raschka, S. (2017). *kdnuggets.* Obtenido de https://www.kdnuggets.com/2016/06/select-support-vector-machine-kernels.html

Rochina, P. (25 de 04 de 2017). *Revistadigital inesem*. Obtenido de https://revistadigital.inesem.es/informatica-y-tics/text-mining/

SAP. (2018). *SAP*. Obtenido de https://www.sap.com/latinamerica/about.html

Sapir, H. (5 de 4 de 2016). *stackoverflow.* Obtenido de https://stackoverflow.com/questions/28057430/what-is-the-access-token-vs-access-token-secret-and-consumer-key-vs-consumer-s

SAS. (2018). *SAS.* Obtenido de https://www.sas.com/es_mx/home.html

Smith, K. (07 de junio de 2016). *Brandwatch Analytics*. Recuperado el 06 de noviembre de 2017, de brandwatch.com: https://www.brandwatch.com/es/2016/06/44-estadisticas-twitter-2016/

Xindong Wu, V. K. (2007). *Top 10 algorithms in data mining.*

Doctora Jenny Ríos Coello, Secretaria de la Facultad de Ciencias de la Administración de la Universidad del Azuay

C E R T I F I C A:

Que, el Consejo de Facultad en sesión del 29 de mayo de 2017, conoció la petición del estudiante **IVÁN SEBASTIÁN PAUTE CÁRDENAS** con código **49170**, que presenta el diseño de su trabajo de titulación denominado: "***EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE TEXTOS EN LOS MENSAJES DE LA RED SOCIAL TWITTER***", presentado previa a la obtención del título de Ingeniero de Sistemas y Telemática.- El Consejo de Facultad acogió el informe de la Junta Académica de Ingeniería de Sistemas y Telemática y resolvió aprobar el diseño. Designa como **Director al ingeniero Marcos Orellana Cordero** y como miembros del Tribunal Examinador al ingeniero Francisco Salgado Arteaga, Ph.D. e ingeniera María Inés Acosta Urigüen.- En esta misma sesión el Consejo de Facultad fija como plazo para la entrega del trabajo de titulación, seis meses contados desde la fecha de su aprobación, esto es hasta el **29 de noviembre de 2017**, debiendo el Director presentar a la Junta Académica, dos informes bimensuales del desarrollo del trabajo de titulación.

Cuenca, mayo 30 de 2017

Dra. Jenny Ríos Coello
Secretaria de la Facultad de
Ciencias de la Administración

CONVOCATORIA

Por disposición de la Junta Académica de **Ingeniería de Sistemas y Telemática**, se convoca a los Miembros del Tribunal Examinador, a la sustentación del Protocolo del Trabajo de Titulación: **"EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE TEXTOS EN LOS MENSAJES DE LA RED SOCIAL TWITTER"**, presentado por el estudiante **Iván Sebastián Paute Cárdenas**, previa a la obtención del grado de **Ingeniero en Sistemas y Telemática**, para el día **LUNES 15 DE MAYO DE 2017 A LAS 07h40.** La sustentación se realizará en el **laboratorio del IERSE.**

Cuenca, 10 de mayo de 2017

Dra. Jenny Ríos Coello
Secretaria de la Facultad

Ing. Marcos Orellana Cordero

Ing. María Inés Acosta Uriguen

Dr. Francisco Salgado Arteaga

mjmr/

Comunicado 12-05-17

Oficio Nro. 061-2017-DIST-UDA

Cuenca, 12 de mayo de 2017

Señor Ingeniero
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
Presente.-

De nuestras consideraciones:

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 12 de mayo del 2017, recibió el proyecto de tesis titulado "Evaluación de las herramientas de minería de textos en los mensajes de la red social Twitter", presentado por Iván Sébastián Paute Cárdenas estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por el Ing. Marcos Orellana, previo a la obtención del título de Ingeniero de Sistemas y Telemática.

Por lo expuesto, y de conformidad con el Reglamento de Graduación de la Facultad, recomendamos como director y responsable de aplicar cualquier modificación al diseño del trabajo de graduación posterior al Ing. Marcos Orellana y como miembros del Tribunal a Francisco Salgado Ph.D. e Ing. María Inés Acosta.

Atentamente,

Ing. Marcos Orellana Cordero
Cordinador Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

UNIVERSIDAD DEL AZUAY

ACTA

SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

1.1 Nombre del estudiante: Iván Sebastián Paute Cárdenas
1.2 Director sugerido: Ing. Marcos Orellana Cordero
1.3 Codirector (opcional): _____
1.4 Tribunal: Ing. María Inés Acosta Uriguen / Dr. Francisco Salgado Arteaga
1.5 Título propuesto: "EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE TEXTOS EN LOS MENSAJES DE LA RED SOCIAL TWITTER"
1.6 Resolución:

1.6.1 Aceptado sin modificaciones ✓

1.6.2 Aceptado con las siguientes modificaciones:

_____

_____

_____

1.6.3 Responsable de dar seguimiento a las modificaciones: Ing. Marcos Orellana Cordero

1.6.4 No aceptado
- Justificación:

_____
_____
_____

Tribunal

Ing. Marcos Orellana Cordero

Ing. María Inés Acosta Uriguen

Dr. Francisco Salgado Arteaga

Sr. Iván Sebastián Paute Cárdenas

Dra. Jenny Ríos Coello
Secretario de Facultad

Fecha de sustentación: día LUNES 15 DE MAYO DE 2017 A LAS 07h40

RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN

1.1 Nombre del estudiante: Iván Sebastián Paute Cárdenas
1.2 Director sugerido: Ing. Marcos Orellana Cordero
1.3 Codirector (opcional):
1.4 Título propuesto: "EVALUACIÓN DE LAS HERRAMIENTAS DE MINERÍA DE TEXTOS EN LOS MENSAJES DE LA RED SOCIAL TWITTER"
1.5 Revisores (tribunal): Ing. María Inés Acosta Uriguen / Dr. Francisco Salgado Arteaga
1.6 Recomendaciones generales de la revisión:

	Cumple totalmente	Cumple parcialmente	No cumple	Observaciones (*)
Línea de investigación				
1. ¿El contenido se enmarca en la línea de investigación seleccionada?	✓			
Título Propuesto				
2. ¿Es informativo?	✓			
3. ¿Es conciso?	✓			
Estado del arte				
4. ¿Identifica claramente el contexto histórico, científico, global y regional del tema del trabajo?	✓			
5. ¿Describe la teoría en la que se enmarca el trabajo	✓			
6. ¿Describe los trabajos relacionados más relevantes?	✓			
7. ¿Utiliza citas bibliográficas?	✓			
Problemática y/o pregunta de investigación				
8. ¿Presenta una descripción precisa y clara?	✓			
9. ¿Tiene relevancia profesional y social?	✓			
Hipótesis (opcional)				
10.¿Se expresa de forma clara?				
11.¿Es factible de verificación?				
Objetivo general				
12.¿Concuerda con el problema formulado?	✓			
13.¿Se encuentra redactado en tiempo verbal infinitivo?	✓			
Objetivos específicos				

14.¿Concuerdan con el objetivo general?	✓			
15.¿Son comprobables cualitativa o cuantitativamente?	✓			
Metodología				
16.¿Se encuentran disponibles los datos y materiales mencionados?	✓			
17.¿Las actividades se presentan siguiendo una secuencia lógica?	✓			
18.¿Las actividades permitirán la consecución de los objetivos específicos planteados?	✓			
19.¿Los datos, materiales y actividades mencionadas son adecuados para resolver el problema formulado?	✓			
Resultados esperados				
20.¿Son relevantes para resolver o contribuir con el problema formulado?	✓			
21.¿Concuerdan con los objetivos específicos?	✓			
22.¿Se detalla la forma de presentación de los resultados?	✓			
23.¿Los resultados esperados son consecuencia, en todos los casos, de las actividades mencionadas?	✓			
Supuestos y riesgos				
24.¿Se mencionan los supuestos y riesgos más relevantes?	✓			
25.¿Es conveniente llevar a cabo el trabajo dado los supuestos y riesgos mencionados?	✓			
Presupuesto				
26.¿El presupuesto es razonable?	✓			
27.¿Se consideran los rubros más relevantes?	✓			
Cronograma				
28.¿Los plazos para las actividades son realistas?	✓			
Referencias				
29.¿Se siguen las recomendaciones de normas internacionales para citar?	✓			
Expresión escrita				
30.¿La redacción es clara y fácilmente comprensible?	✓			
31.¿El texto se encuentra libre de faltas ortográficas?	✓			

(*) Breve justificación, explicación o recomendación.

- Opcional cuando cumple totalmente,
- Obligatorio cuando cumple parcialmente y NO cumple.

.....

.....

.....

Ing. Marcos Orellana Cordero Ing. María Inés Acosta Uriguen Dr. Francisco Salgado Arteaga

UNIVERSIDAD DEL AZUAY

Escuela Sistemas y Telemática	Oficio Estudiante: Solicitud aprobación de Protocolo de Trabajo de Titulación	IST-RE-EST-02 Versión 01 04/04/2017 Página 1 de 1

Lugar de Almacenamiento	Retención	Disposición Final
F: Archivo Secretaría de la Facultad	5 años	Almacenar en archivo pasivo de la Facultad

Cuenca, 11 de Mayo del 2017

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Estimado Señor Decano, yo **Iván Sebastián Paute Cárdenas** con C.I. **0104967773**, código estudiantil 49170; estudiante de la Carrera de Sistemas y Telemática, solicito muy comedidamente a usted y por su intermedio al Consejo de Facultad, la aprobación del protocolo de trabajo de titulación con el tema **"Evaluación de las herramientas de minería de textos en los mensajes de la red social Twitter"** previo a la obtención del título de Ingeniero en Sistemas y Telemática para lo cual adjunto la documentación respectiva.

Por la favorable acogida que brinde a la presente, anticipo mi agradecimiento.

Atentamente:

Iván Sebastián Paute Cárdenas

Estudiante de la Carrera de Sistemas y Telemática

Edición autorizada de 10.000 ejemplares Del 788.501 al 798.500 Nº 0796065

UNIVERSIDAD DEL AZUAY

DOCTORA JENNY RIOS COELLO, SECRETARIA DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN DE LA UNIVERSIDAD DEL AZUAY

CERTIFICA:

Que, el Señor **Iván Sebastián Paute Cárdenas** registrado con código **49170** estudiante de la Escuela de Ingeniería de Sistemas y Telemática tiene aprobado más del 80% de su Pensum de estudios.

Que, el Señor **Iván Sebastián Paute Cárdenas** le falta aprobar las siguientes asignaturas para finalizar sus estudios:

INGENIERIA DE SOFTWARE II

DESARROLLO Y GESTION DE PROYECTOS INFORMATICOS

PROYECTOS TELEMATICOS

SISTEMA DE INFORMACION GERENCIAL

PRODUCCION II

DISEÑO DE TRABAJO DE GRADUACION

Cuenca, 11 de Mayo de 2017

FACULTAD DE ADMINISTRACION SECRETARIA

Derecho 156920
ver.-

Edición autorizada de 10.000 ejemplares Del 788.501 al 798.500 Nº 079602

Cuenca, 11 de Mayo del 2017

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Yo, **Marcos Patricio Orellana Cordero** informo que he revisado el protocolo de trabajo de titulación previo a la obtención del título de Ingeniero en Sistemas y Telemática, denominado " **Evaluación de las herramientas de minería de textos en los mensajes de la red social Twitter**", realizado por el estudiante **Iván Sebastián Paute Cárdenas**, con código estudiantil 49170, protocolo que a mi criterio, cumple con los lineamientos y requerimientos establecidos por la carrera.

Por lo expuesto, me permito sugerir que sea considerado para la revisión y sustentación del mismo,

Sin otro particular, suscribo.

Atentamente

Ing. Marcos Orellana Cordero

UNIVERSIDAD DEL AZUAY

UNIVERSIDAD DEL AZUAY
INGENIERIA EN SISTEMAS Y TELEMATICA
DISEÑO DE TESIS

1. DATOS GENERALES

1.1 Nombre del estudiante: Paute Cárdenas Iván Sebastián

1.1.1 Código: 49170

1.1.2 Contacto:

Teléfono convencional: 2849405

Celular: 0992552666

Correo electrónico: sebastianpaute@gmail.com

1.3 Director sugerido: Orellana Cordero Marcos Patricio Ing.

1.2.1 Contacto:

Teléfono convencional:

Celular: 0999955611

Correo electrónico: marore@uazuay.edu.ec

1.3 Co-director sugerido:

1.4 Asesor metodológico: PhD Francisco Salgado.

1.5 Tribunal designado:

1.6 Aprobación: Junta Académica:

Consejo de Facultad:

1.7 Línea de Investigación de la carrera:

1.7.1 Código UNESCO: 1203 Informática de computadores

1203.17 Informática

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 079720

1.7.2 Tipo de trabajo: Investigación aplicada.

1.8 Área de estudio: minería de texto, análisis de sensibilidad de redes sociales.

1.9 Título propuesto: Evaluación de las herramientas de minería de textos en los mensajes de la red social Twitter.

1.10 Subtítulo:

1.11 Estado del proyecto: La investigación no propone generar herramientas nuevas, consta de un análisis de herramientas existentes, con la finalidad de encontrar la sensación de las personas sobre el tráfico de ciudades grandes

2. CONTENIDO

2.1 Motivación de la investigación:

En la actualidad existe una gran cantidad de herramientas para minar textos de la red social Twitter, y cada una de estas herramientas cuenta con factores, como: gama de algoritmos disponibles, usabilidad, seguridad, viabilidad, velocidad, debido a esta razón la necesidad de realizar un análisis para encontrar la mejor herramienta para la búsqueda de texto específico.

Encontrar la mejor herramienta para minar texto ayudará a realizar un análisis de sensibilidad acerca del tráfico en las ciudades grandes, y que como lo viven los usuarios de la red social en cuestión. Este análisis servirá para una investigación más exhaustiva acerca de la movilidad, que a futuro y con la integración de software adecuado podría dar a conocer de forma autómata el tráfico dentro de una ciudad.

2.2 Problemática:

Con la gran cantidad de datos que se generan en la red social Twitter, y con una variedad de herramientas para minar textos, es necesario encontrar una herramienta que procese esta información de manera ágil y presente resultados significativos para lo que se esté sondeando en

la red social, así con los resultados obtenidos se podrá realizar un análisis de sentimientos más confiable y eficiente.

2.3 Pregunta de investigación:

¿Cuál es la mejor herramienta para minar textos y realizar un análisis de sensibilidad tomando en cuenta factores, como: gama de algoritmos disponibles, usabilidad, seguridad, viabilidad, velocidad en el tráfico de las grandes ciudades?

2.4 Resumen:

La investigación consiste en comparar diferentes herramientas de minería de textos hasta encontrar la mejor opción, y mostrar las herramientas en un cuadro comparativo, de esta manera elegir la mejor a ser aplicada en un análisis de sensibilidad, acerca del tráfico que se vive en grandes ciudades.

Se experimentará con datos generados por la red social Twitter, catalogados como datos no estructurados. En un siguiente paso se procesará los datos con la extracción, almacenamiento y depuración de los elementos texto. Se realizarán pruebas de algoritmos en herramientas de minería de texto, sobre datos de tráfico previamente obtenidos y se realizará una selección según los resultados obtenidos por cada uno.

2.5 Indagación exploratoria:

En nuestro medio actual, las redes sociales son significativamente importantes para la investigación social computacional que, busca patrones, investiga preguntas, o quiere saber el comportamiento de la humanidad (Batrinca & Treleaven, 2014) .Y con aproximadamente más de 320 millones de usuarios al mes en la actualidad, Twitter está produciendo una gigantesca cantidad

de datos para explotar y extraer ideas, pensamientos, y sentimientos, los cuales son aplicables a una gran parte de áreas de estudio (Byrd, Mansurov, & Baysal, 2016).

Una de las características más significativas de Twitter es su portabilidad, al estar en la web y en dispositivos móviles permite a los usuarios de esta red publicar en cualquier lugar y momento. Esta facilidad de agregar información a la red, hace que sea un repositorio de datos que revelan eventos del mundo real. Muchos eventos sucedidos fueron señalados por los usuarios de esta red, por poco al mismo tiempo o antes de que los medios tradicionales de comunicación como televisión o radio pudieran darlos a conocer (Parikh, 2013). Otra característica de Twitter, que marca diferencia frente a otras redes sociales, es el tamaño limitado de los mensajes puestos en esta red, con un tamaño máximo de 140 caracteres, limita a los usuarios a expresarse en palabras o frases clave, haciendo esta información más significativa para su posterior tratamiento, una de las principales ventaja de Twitter es que lo utilizan todas las clases sociales, ampliando la visión de lo que está pasando en todo el mundo (De Groot, 2012). Con el aumento del uso de las redes sociales, Text mining es una tecnología ascendiente que pretende extraer información valiosa de datos textuales no estructurados (He, Zha, & Li, 2013). Por ejemplo según el estudio citado por (He, Zha, & Li, 2013) señala que alrededor del el 80% de la información de una institución está comprometida en documentos de texto, tales como informes, ordenes de pedido de los clientes, facturas, correos electrónicos y memorandos. Según este informe los análisis de texto permiten a las empresas convertir grandes cantidades de textos generado por sus empleados, clientes y competencia, en resúmenes significativos, que respaldan la toma de decisiones con un fundamento en evidencias (Gandomi & Haider, 2015).

Para la obtención de información valiosa de un gran banco de información textual con eficacia, es indispensable el uso de métodos informáticos automatizados (He, Zha, & Li, 2013), siendo el

objetivo principal de la minería de texto hallar tendencias, guías, patrones ocultos, normas, o modelos provechosos, todo esto extraídos de datos textuales no estructurados, como por ejemplo correos, redes sociales, correos electrónicos y archivos web (He, Zha, & Li, 2013). Mediante la recopilación de sentimientos público de redes sociales, se puede lograr un análisis de mercado para llevar a cabo negocios, también conocer cómo se ve una empresa en el medio o por que una empresa fracasa. Este sitio proporciona un medio rápido para que sus usuarios se expresen, pero al mismo tiempo con esta información crear una valiosa base de datos (Wakade, Shekar, Liszka, & Chan, 2012).

Las aplicaciones más importantes de la minería de texto comprenden resumen de texto, análisis de enlaces y agrupación. El análisis de agrupamiento en especial tiene la ventaja de revelar tendencias, correlaciones o patrones imprevistos de los datos (He, Zha, & Li, 2013).

En la actualidad, hay una gran variedad de herramientas para realizar minería de textos, como Leximancer, Nvivo 9, SAS Enterprise Miner o SPSS Modeler (anteriormente Clementine). Estas herramientas usan complejos modelos de programación, como árboles de decisión, agrupación, programación lógica y algoritmos estadísticos para encontrar patrones en datos textuales no estructurados (He, Zha, & Li, 2013).También Bošnjak dio a conocer una herramienta de nombre TwitterEcho que es un rastreador de código abierto. TwitterEcho es utilizado para recuperar información de la red social Twitter. Permite a los buscadores de datos recopilar información de sean de interés específico. Otra herramienta de minería de texto es TweerPos, un programa adecuado para estudiar y buscar el origen geográfico de los tweets, y para revelar la evolución

geografía de la popularidad de los temas en la red social Twitter, esta herramienta ofrece una visualización mixta que muestra datos basados en mapas de calor y geográficos, permite un estudio a detalle y extraer textos con respecto a la distribución geoespacial de datos en específico (Wijnants, Blazejczak, Quax, & Lamotte, 2015).

La Universidad de Salamanca desarrollo PIAR una plataforma capaz de reconocer una puntuación de influencia alternativa, basada en los contactos y seguidores de un usuario de Twitter, la cual tiene un enorme potencial para el uso de académicos y profesionales a pesar que fue desarrollada con fines de seguridad nacional (Lahuerta-Otero & Cordero-Gutierrez, 2016).

La minería de textos aplicada a redes sociales, puede entregar interesantes resultados como el comportamiento humano y la interacción entre personas (He, Zha, & Li, 2013). Según (He, Wu, Yan, Akula, & Shen, 2015), el análisis del sentimiento se refiere a la extracción automática de opiniones positivas o negativas de los textos, ya que normalmente los textos contienen una mezcla de sentimientos positivos, negativos o neutros. El análisis de sentimientos se utiliza para determinar la actitud de los clientes y usuarios en temas específicos, tales como revisiones de productos (por ejemplo, electrónica, software, ropa), revisiones de servicios, finanzas y estado de ánimo en general de la población.

2.6 Objetivo general:

Desarrollar un cuadro comparativo de las distintas herramientas que existen en el mercado para minería de texto y recomendar la mejor para el análisis de sensibilidad sobre la congestión vehicular en las grandes ciudades.

2.7 Objetivos específicos:

1. Realizar una investigación del estado de arte sobre las diferentes herramientas de minería de texto.
2. Probar dichas herramientas en un entorno real sobre la red social Twitter y encontrar las mejores para la investigación.
3. Realizar un análisis comparativo de las herramientas de minería de texto a fin de seleccionar la mejor alternativa y generar un cuadro analítico de herramientas de minería de texto, mostrando las ventajas y desventajas de dichas herramientas.
4. Recomendar la mejor herramienta para el análisis de sensibilidad sobre la congestión vehicular.

2.8 Metodología:

Se investigará las bases teóricas del tema, apoyado en una investigación bibliográfica en diferentes fuentes: revistas, bibliotecas digitales, libros y conferencias, se elegirá las fuentes de información más relevantes y con esta información se redactará un documento que denote el estado del arte y un análisis sobre herramientas de minería de texto.

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 07972

Con la información necesaria se procederá a la experimentación de diferentes algoritmos sobre distintas herramientas de minería de texto, documentando cada prueba realizada para obtener los elementos necesarios y establecer un cuadro comparativo de herramientas de minería de texto para así recomendar la mejor opción.

2.9 Alcances y resultados esperados:

Finalizada la primera parte del proyecto se espera obtener un estado del arte claro e información necesaria sobre herramientas de minería de texto, obteniendo una mejor visión del alcance y los que se espera del proyecto

Para la segunda fase del proyecto se pretende realizar pruebas con distintos algoritmos sobre distintas herramientas de minería de texto, y documentar sus resultados y desempeño.

Terminado el proyecto se espera la selección de una o más herramientas acorde a la minería de textos la red social Twitter, un cuadro comparativo de las herramientas que existen en minería de textos y un análisis de sensibilidad sobre la congestión vehicular en una ciudad.

UNIVERSIDAD DEL AZUAY

2.10 Supuestos y riesgos:

Riesgos	Probabilidad	Alternativas de solución
Falta de conocimiento en minería de texto, análisis de sensibilidad y manejo de herramientas de minería de texto.	Media	Asesoría con profesionales entendidos o expertos en el tema.
Poca o nula información en libros sobre herramientas de minería de texto	Media	Investigar en fuentes alternas como revistas tecnológicas, publicaciones de universidades, bibliotecas digitales
Costos elevados por herramientas mejor puntuadas según pruebas de otras investigaciones.	Media	Buscar versiones de prueba de dichas herramientas
Investigaciones similares ya realizadas, sobre comparación de herramientas y aplicación de las mismas sobre tráfico de ciudades grandes	Baja	Buscar una problemática diferente, pero con la misma temática de minería de texto
Poca información en Twitter para minar, por falta de publicaciones sobre tráfico de los usuarios	Baja	Buscar un tema con más afluencia de información en Twitter

2.11 Esquema tentativo:

PORTADA

DEDICATORIA

AGRADECIMIENTOS

ABSTRACT

ÍNDICE DE IMÁGENES

ÍNDICE DE TABLAS

ÍNDICE DE CONTENIDOS

INTRODUCCIÓN

OBJETIVO GENERAL

Edición autorizada de 10.000 ejemplares Nº 07971

OBJETIVOS ESPECÍFICOS

JUSTIFICACIÓN

CAPÍTULO 1 - MARCO TEÓRICO

¿Qué es la red social Twitter?

¿Qué es Minería de Texto?

¿Qué son las Herramientas de minería de Texto?

¿Qué son los análisis de sensibilidad?

CAPÍTULO 2 – Prueba de las mejores herramientas de minería de texto.

CAPITULO 3 – Desarrollo del análisis de sensibilidad.

CAPITULO 4 – RESULTADOS

Cuadro comparativo entre Herramientas de Minería de Texto

CONCLUSIONES

RECOMENDACIONES

BIBLIOGRAFÍA

ANEXOS

2.12 Cronograma:

CRONOGRAMA

Objetivo Específico	Actividad	SEMANAS 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Realizar una investigación del estado de arte sobre las diferentes herramientas de minería de texto.	Revisión de documentación y conceptos de fuentes válidas para utilizarse dentro del desarrollo de la investigación																					
Probar dichas herramientas en un entorno real sobre la red social Twitter y encontrar las mejores para la investigación.	Experimentación con un grupo de trabajo sobre datos reales de Twitter, de diferentes algoritmos sobre distintas herramientas de minería de texto, documentado cada prueba realizada.																					
Realizar un análisis comparativo de las herramientas de minería de texto a fin de seleccionar la mejor alternativa y generar un cuadro analítico de herramientas de minería de texto, mostrando las ventajas y desventajas de dichas herramientas.	Con la información obtenida de las pruebas realizadas establecer un cuadro comparativo de herramientas de minería de texto y recomendar la mejor opción.																					
Recomendar la mejor herramienta para el análisis de sensibilidad sobre la congestión vehicular.	Con las pruebas realizadas recomendar la mejor opción para un análisis de sensibilidad.																					

2.13 Referencias:

Batrinca, B., & Treleaven, P. C. (2014). Social media analytics: a survey of techniques, tools and platforms. *AI and Society*, 89-116.

Bošnjak, M. (2012). TwitterEcho. *Support Open Research with Twitter Data.*

Byrd, K., Mansurov, A., & Baysal, O. (2016). Mining twitter data for influenza detection and surveillance. *Proceedings - International Workshop on Software Engineering in Healthcare Systems, SEHS 2016*, 43-49.

Byun, C., Lee, H., & Kim, Y. (2012). Automated Twitter data collecting tool for data mining in social network. *Proceedings of the 2012 ACM Research in Applied Computation Symposium on - RACS '12*, 76.

De Groot, R. (2012). Data Mining for Tweet Sentiment Classification. 63.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 137-144.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information and Management*, 801-812.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 464-472.

Lahuerta-Otero, E., & Cordero-Gutierrez, R. (2016). Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. *Computers in Human Behavior*, 575-583.

Parikh, R. (2013). ET : Events from Tweets. 613-620.

Wakade, S., Shekar, C., Liszka, K. J., & Chan, C.-c. (2012). Text Mining for Sentiment Analysis of Twitter Data. *Ike*, 5.

Wijnants, M., Blazejczak, A., Quax, P., & Lamotte, W. (2015). Geo-spatial trend detection through twitter data feed mining. *Lecture Notes in Business Information Processing*, 212-227.

UNIVERSIDAD DEL AZUAY

2.14 Anexos: para casos en los que se requiera respaldar el proyecto.

2.15 Firma de responsabilidad (estudiante)

Sebastián Paute

2.16 Firma de responsabilidad (director sugerido)

Ing. Marcos Orellana

2.17 Firma de Asesor metodológico (asesor sugerido)

PhD Francisco Salgado.

2.18 Fecha de entrega: 10 de mayo 2017

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500 Nº 079723