UCUENCA

Facultad de Ingeniería
Carrera de Ingeniería de Sistemas

Generación de un corpus para detección de competidores en el idioma español mediante minería de opiniones comparativas. Caso de estudio: Sector Textil en la provincia del Azuay

Trabajo de titulación previo a la obtención del título de Ingeniero de Sistemas

Autores:

Néstor Ariel Bravo Chuqui
CI: 0105272421
Correo electrónico: arielbravo.ec@gmail.com

Ángel Patricio Fajardo Cárdenas
CI: 0106243751
Correo electrónico: apatricio.fajardoc@gmail.com

Director:

Ing. Andrés Vinicio Auquilla Sangolquí
CI: 0103557369

Codirector:

Ing. Paúl Fernando Vanegas Peña
CI: 0102596186

Cuenca, Ecuador

02-agosto-2022

Resumen:

En la actualidad con el avance de la tecnología y más aún con la llegada de la pandemia el uso de las plataformas digitales se ha incrementado. Un estudio presentado por la Cámara de Comercio Electrónico Ecuatoriana del año 2020 demuestra que el comercio electrónico ha incrementado en al menos 15 veces con respecto al 2019 el uso de plataformas digitales online con la llegada de la pandemia. Debido a esto, las empresas para hacer estudios de mercado deben buscar nuevas fuentes de información. Por lo tanto, el internet se ha convertido en un insumo intangible de toda estrategia comercial. Una parte fundamental de una estrategia comercial es analizar a la competencia, este análisis en años anteriores según la literatura se realizaba generalmente mediante encuestas, pero con la llegada de las plataformas digitales ha cambiado este método y hoy por hoy se puede extraer los datos de la web para luego implementar un proceso de Inteligencia Competitiva (CI), la cual permite hacer un análisis completo para tener una ventaja competitiva. CI comprende de varios pasos, esta investigación aborda todos estos pasos, pero se enfoca principalmente en el paso inicial, la recolección y análisis de datos, que es un paso fundamental para CI, donde actualmente existen problemas como: falta de corpus en español especializado para CI, por lo cual los investigadores no tienen la facilidad de implementar modelos de aprendizaje automático que les ayuden a tener una ventaja competitiva. El presente trabajo de investigación presenta una metodología para la creación de un corpus en el idioma español que permita entrenar algoritmos con el fin de realizar detección de competidores en el contexto del sector textil. Se han generado dos resultados principales: 1) Una metodología utilizando técnicas de minería de textos (minería de opiniones comparativas y reconocimiento de entidades nombradas) para construir corpus enfocado hacia la Inteligencia Competitiva. 2) Un corpus en español, dentro del dominio de comentarios de redes sociales, el cual sirve de base para futuras investigaciones relacionadas con la inteligencia competitiva, específicamente en la detección de competidores en el lenguaje español, donde la CI estaba estrictamente restringida por la falta de un corpus. Por último, se ha evaluado la utilidad del corpus desarrollado mediante un Dashboard creado en base a un caso de estudio llevado a cabo en el contexto del sector textil en redes sociales. Se ha demostrado que efectivamente es de utilidad para el sector textil, sin embargo, se recomienda hacer una nueva validación con empresas que estén directamente relacionadas al sector textil y así obtener una validación más directa, también se recomienda evaluar en otros sectores.

Palabras claves: : Inteligencia competitiva (**CI**). Aprendizaje automático. Reconocimiento de entidades nombradas (**NER**). Minería de textos. Minería de opiniones comparativas. Análisis de competencia. Redes sociales. Sector textil.

Abstract:

Currently, with the advancement of technology and even more so with the arrival of the pandemic, the use of digital platforms has increased. A study presented by the Ecuadorian Chamber of Electronic Commerce for the year 2020 shows that electronic commerce has increased the use of online digital platforms by at least 15 times compared to 2019 with the arrival of the pandemic. Due to this, companies to do market research must look for new sources of information. Therefore, the internet has become an intangible input for any business strategy. A fundamental part of a commercial strategy is to analyze the competition, this analysis in previous years according to the literature was generally carried out through surveys, but with the arrival of digital platforms this method has changed and today the data can be extracted from the web to then implement a Competitive Intelligence (CI) process, which allows a complete analysis to have a competitive advantage. CI comprises several steps, this research addresses all these steps, but focuses mainly on the initial step, data collection and data analysis, which is a fundamental step for CI, where there are currently problems such as: lack of corpus in Spanish specialized for CI, so researchers do not have the facility to implement machine learning models that help them to have a competitive advantage. This research presents a methodology for the creation of a corpus in the Spanish language that allows algorithms to be trained in order to detect competitors in the context of the textile sector. Two main results have been generated: 1) A methodology using text mining techniques (comparative opinion mining and named entity recognition) to build a corpus focused on Competitive Intelligence. 2) A corpus in Spanish, within the domain of social network comments, which serves as a basis for future research related to competitive intelligence, specifically in the detection of competitors in the Spanish language, where the CI was strictly restricted by the lack of a corpus. Finally, the usefulness of the corpus developed has been evaluated through a Dashboard created based on a case study carried out in the context of the textile sector in social networks. It has been shown that it is indeed useful for the textile sector, however, it is recommended to carry out a new validation with companies that are directly related to the textile sector and thus obtain a more direct validation, it is also recommended to evaluate in other sectors.

Keywords: Competitive intelligence (**CI**). Machine learning. Named entity recognition (**NER**). Text mining. Comparative opinion mining. Competition analysis. Social networks. Textile sector.

INDICE

INDICE DE FIGURAS

INDICE DE TABLAS

UCUENCA

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

Néstor Ariel Bravo Chuqui en calidad de autor/a y titular de los derechos morales y patrimoniales del trabajo de titulación "Generación de un corpus para detección de competidores en el idioma español mediante minería de opiniones comparativas. Caso de estudio: Sector Textil en la provincia del Azuay", de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca, 2 de agosto de 2022

Néstor Ariel Bravo Chuqui

C.I: 0105272421

UCUENCA

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

Ángel Patricio Fajardo Cárdenas en calidad de autor/a y titular de los derechos morales y patrimoniales del trabajo de titulación "Generación de un corpus para detección de competidores en el idioma español mediante minería de opiniones comparativas. Caso de estudio: Sector Textil en la provincia del Azuay", de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca, 2 de agosto de 2022

Ángel Patricio Fajardo Cárdenas

C.I: 0106243751

UCUENCA

Cláusula de Propiedad Intelectual

Néstor Ariel Bravo Chuqui, autor/a del trabajo de titulación "Generación de un corpus para detección de competidores en el idioma español mediante minería de opiniones comparativas. Caso de estudio: Sector Textil en la provincia del Azuay", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autor/a.

Cuenca, 2 de agosto de 2022

Néstor Ariel Bravo Chuqui

C.I: 0105272421

UCUENCA

Cláusula de Propiedad Intelectual

Ángel Patricio Fajardo Cárdenas, autor/a del trabajo de titulación "Generación de un corpus para detección de competidores en el idioma español mediante minería de opiniones comparativas. Caso de estudio: Sector Textil en la provincia del Azuay", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autor/a.

Cuenca, 2 de agosto de 2022

Ángel Patricio Fajardo Cárdenas

C.I: 0106243751

Dedicatoria

A mis padres, María y Vinicio, quienes me han apoyado en cada momento de mi vida. Todo esto es total mérito de ustedes, me han enseñado lo que es el esfuerzo, y motivado a cumplir mis metas. Espero que puedan sentirse orgullosos de mí, ya que todo lo que hago es para ustedes.

A mi hermano, que ha sido siempre mi ejemplo a seguir. Por todos los consejos y buen ejemplo que ha sido para mí, inspirándome a siempre ser mejor cada día.

A mis tíos/ñaños que siempre han estado presentes en mi vida, sin su cariño nada de esto sería posible.

A mis abuelitos, que me brindan sus consejos y amor, los llevo siempre en mi corazón.

Por último, a mi Alejo, espero que veas en mí un ejemplo a seguir, siempre te protegeré y apoyaré en toda tu vida.

Ariel Bravo.

Dedicatoria

A mis padres, Ángel y Blanca, por ser mi apoyo cada día y por todo el esfuerzo que han realizado a lo largo de esta etapa universitaria y por estar siempre pendiente de mí. Me han guiado siempre para ser una buena persona y me han ayudado a siempre conseguir todo lo que me he propuesto, ha sido un camino que hemos recorrido juntos. ¡Lo hemos logrado Papi y Mami!

De la misma manera, a mi Tía Elsa, que siempre me ha apoyado desde el inicio de mi carrera y por motivarme a terminarla, por sus buenos consejos y por siempre estar pendiente de mi en esta etapa universitaria. Ha sido un camino que hemos recorrido juntos. ¡Lo hemos logrado Tía!

A mi hermanos, Mauricio, German y Sebastián y mi hermana Jhoana por siempre estar en todos momentos.

A mis Abuelita Florinda, por su apoyo y porque siempre estuvo pendiente de mí, ahora ella desde el cielo debe está muy feliz verme alcanzar esta meta.

A la Universidad de Cuenca por ser mí segundo hogar. A mis compañeros y docentes, porque gracias a sus enseñanzas he terminado esta etapa y puedo continuar con mi vida profesional.

Patricio Fajardo C.

Agradecimientos

A mis padres, María y Vinicio por formarme como una persona de bien y todo el amor incondicional que he recibido siempre, nunca podré pagarles todo lo que han hecho por mí. De igual manera a mi hermano, por todo su apoyo, consejos y ayudas que me ayudaron a llegar a esta prestigiosa universidad.

A mis tíos/ñaños, especialmente a mi ñaño Julio, que ha sido de gran inspiración a llegar más lejos y saber que no existen límites si nos esforzamos lo suficiente. A mis abuelitos por siempre recibirme en su hogar y comprenderme en los momentos de dificultad.

A nuestros directores de tesis, Ingeniero Andrés Auquilla e Ingeniero Paúl Vanegas, por su valiosa dirección y apoyo, por su paciencia, profesionalismo, críticas constructivas y sobre todo por el tiempo que invirtieron en nosotros.

A mis amigos más cercanos, Edisson, Bryan, Erika, Stefy, Diego, Adrián, Fercho, Patricio por todas las experiencias que vivimos y hacer que la vida universitaria sea de las mejores etapas de mi vida.

Finalmente, a la Universidad de Cuenca, especialmente a la gloriosa Facultad de Ingeniería, por todas las enseñanzas, y permitirme ser un profesional. Siento mucho orgullo, solo los que hemos pasado por esto sabemos lo difícil y el enorme sacrificio que demanda esta profesión.

Ariel Bravo.

Agradecimientos

A Dios por darme la voluntad y la fuerza para seguir adelante en todo este proceso de realización personal.

A mi familia por ser un apoyo constante en todo este proceso. Especialmente a mis padres, mi tía Elsa, mis hermanos y mi hermana quienes me han apoyado cada día, tanto en los buenos y malos momentos.

A nuestros directores de tesis, Ingeniero Andrés Auquilla e Ingeniero Paúl Vanegas, por la confianza brindada al permitirnos contribuir en el desarrollo de este proyecto y por su valiosa dirección y apoyo, por su paciencia, profesionalismo, críticas constructivas, experiencia y educación. Factores que han sido mi fuente de motivación durante este tiempo.

A mis amigos más cercanos, Moisés, Ariel, Jonathan, Lenin, Eduardo por brindarme su apoyo, compartir grandes momentos, experiencias inolvidables y hacer que la vida universitaria sea de las mejores etapas de mi vida.

Y finalmente, a la Universidad de Cuenca y a la Facultad de Ingeniería por prepararnos como buenos profesionales, y brindarnos los mejores años y recuerdos.

Patricio Fajardo C.

LISTADO DE ABREVIATURAS

CI: Inteligencia Competitiva
SCIP: Sociedad de Profesionales de Inteligencia Competitiva
NLP: Procesamiento de Lenguaje Natural
PYMEs: Pequeñas y medianas empresas
NER: Reconocimiento de entidad nombrada
API: Application Programs Interfaces
POS: Part of Speech Tagging
BOW: Bag of Words
SVM: Support Vector Machine
RNA: Red Neuronal Artificial
RNP: Red Neuronal Profunda
VP: Verdaderos Positivos
VN: Verdaderos Positivos
FP: Falsos Positivos
FN: Falsos Negativos
ROC: Receiver Operating Characteristic
AUC: Área bajo la curva
NLTK: Natural Language Toolkit
CRF: Campos Aleatorios Condicionales
PER: Persona
ORG: Organización
LOC: Locación
MISC: Cualquier otra entidad

CAPÍTULO 1: INTRODUCCIÓN

1.1 Contexto

En la actualidad la mayoría de las empresas descuidan la necesidad de una decisión estratégica correcta para crear condiciones favorables que aseguren el éxito futuro en un entorno empresarial aún más desafiante (Gundersen, 2019; Mclean & Woods, 2014). Para realizar un correcto plan estratégico, es importante que las empresas recopilen y analicen información sobre los productos y planes de sus competidores (Mclean & Woods, 2014). Con base en dicha información, una empresa puede conocer las debilidades y fortalezas relativas de sus propios productos, y luego tomar decisiones inteligentes como diseñar nuevos productos y campañas específicas para contrarrestar las de sus competidores (Xu et al., 2011). Hoy en día uno de los aspectos fundamentales dentro del análisis de la competencia es la Inteligencia Competitiva (CI por sus siglas en inglés). Un buen diseño de CI en una empresa puede ayudar en sus procesos de planificación y determinar la intención y la capacidad de sus competidores. Asimismo, CI es parte de la cadena de valor de una empresa, que convierte los datos de los componentes en información utilizable y los resultados de decisiones estratégicas (Stefanikova et al., 2015).

La definición de CI varía según los diferentes autores y enfoques en el campo de los negocios. Según Bulley et al. (2014), Mclean & Woods (2014) y Zanasi (1998) la CI se la puede definir como el proceso de seguimiento del entorno competitivo y competidores de la empresa, en el cual la definición, recopilación de información, análisis y distribución de los resultados obtenidos se realiza de forma paulatina para que puedan apoyar la actividad empresarial eficiente y su capacidad para tomar decisiones calificadas. El objetivo actual y más importante de la actividad de CI es comprender, con la mayor antelación posible, la estrategia tecnológica del competidor y/o las tendencias del mercado (Bose, 2008).

En base al objetivo de CI, se han realizado estudios en algunas empresas como por ejemplo, Kim et al. (2016) extrae inteligencia competitiva en las redes sociales para encontrar el conocimiento del mercado comparando las opiniones de los consumidores con el desempeño de ventas de una empresa, demostrando que los datos de las redes sociales contienen inteligencia competitiva. Otro estudio se presenta en Xue et al. (2018) donde se utilizó inteligencia competitiva para encontrar empresas que son competencia de IBM. Para este estudio se utilizaron 158594 comentarios de la plataforma YouTube sobre la empresa IBM entre en año 2006 y 2013, como resultado presentan un top 10 de competidores de la empresa, análisis detallado de las relaciones competitivas de IBM con algunos de sus principales rivales en la cual destacan que Google es su principal competencia y también presentan una visualización de nube de palabras que resaltan las fortalezas y debilidades de IBM en el periodo de 2010-2011. Por último, en Vera Kristanti Dewi & Sri Darma (2019) se presentó un estudio sobre dos empresas de Indonesia Grab y Go-Je, donde se realizó un análisis de los factores que le permitían a Grab tener ventaja competitiva; los resultados indicaron que entre otros factores que ayudaron a dicha ventaja está la buena implementación de la inteligencia Competitiva.

CI tiene importancia en la gestión y práctica empresarial, debido a que agrega valor a la planificación y la toma de decisiones (Bulley et al., 2014). En Bose (2008) y Gabes (2012) se

mencionan algunos beneficios de CI, uno de los más importantes es la capacidad para crear perfiles de información para ayudar a una empresa a identificar las fortalezas, debilidades, estrategias, objetivos, posicionamiento en el mercado y patrones de reacción probables de su competidor. Estos perfiles de información incluyen los datos necesarios para identificar, clasificar y rastrear efectivamente a los competidores y su comportamiento. Con el uso de dichos perfiles, una empresa comienza a buscar puntos de comparación en cuanto a sus fortalezas y debilidades frente a sus competidores. La importancia de la CI en las empresas se convierte prácticamente en una necesidad y es ampliamente aceptada en las empresas (Amarouche et al., 2015).

De acuerdo a Stefanikova et al. (2015) después de una encuesta a varias empresas hasta un 70% de las empresas dicen que la inteligencia competitiva ha sido un elemento crítico de su estrategia comercial y de marketing. De acuerdo a Bose (2008) uno de los principales sectores interesados es el sector productivo, en donde se encuentran PYMES del sector textil, las cuales la utilizan para entender el mercado y su competencia. Por último, también en otra encuesta realizada en la investigación de Stefanikova et al. (2015) cerca de los 70% de las empresas encuestadas dicen que planea aumentar sus presupuestos de inteligencia competitiva, ya que el 94% de ellas está de acuerdo en que el sistema de inteligencia competitiva les beneficia.

Uno de los puntos a destacar en CI es su proceso, el cual comprende la acción de recopilar, analizar y aplicar información sobre productos, competidores, proveedores, reguladores, socios y clientes para las necesidades de planificación a corto y largo plazo de una organización (Kahaner, 1997). Un proceso de CI efectivo, según la Sociedad de Profesionales de Inteligencia Competitiva (SCIP), se ejecuta en un ciclo continuo, llamado ciclo de CI. El SCIP describe al ciclo de CI como el proceso mediante el cual la información en bruto se adquiere, recopila, transmite, evalúa, analiza y pone a disposición como inteligencia completa para que los formuladores de políticas la utilicen en la toma de decisiones y la acción. Hay cinco fases que constituyen este ciclo: planificación y dirección; recolección; análisis; difusión y retroalimentación. Sin embargo, en Xu et al. (2011) destacan que aún existen problemas grandes en la fase de recolección de datos y en la fase de análisis de datos debido a la falta de fuentes de información suficientes y confiables sobre los competidores, este problema restringe enormemente la capacidad de CI.

1.2 Justificación

CI es un campo de investigación que ha tenido grandes avances, anteriormente, para CI se utilizaba información de perfil comercial e informes editados por profesionales de la CI. La tendencia actual de CI es recopilar información de la competencia a partir de datos en línea. En la actualidad técnicas como la minería de textos y la minería de opiniones ayudan en el proceso de extracción y análisis de este conjunto de datos en línea. A la minería de textos se la define como el proceso de extraer la información útil de datos no estructurados tales como páginas web, artículos de revistas, etc. Esta técnica se ha utilizado ampliamente en las fases de recolección y análisis de CI (Cardoso et al., 2019), por su parte la minería de opiniones es una tarea del procesamiento de lenguaje natural (NLP) que permite identificar opiniones, emociones, actitudes relacionadas en textos (Cedeno-Moreno & Vargas, 2020). Estudios como los presentados en Arora et al. (2017), Gao et al. (2018), He et al. (2013), Jeong et al. (2019), Liu et al. (2019), Tsirakis et al. (2017), Vera Kristanti Dewi & Sri

Darma (2019), Xue et al. (2018), Yadav & Shah (2019) proponen varios modelos para detección de competidores donde utilizan como fuente de datos los comentarios de redes sociales y páginas de comercio electrónico, posteriormente aplican algoritmos basados en redes neuronales, Support Vector Machine (SVM), etc., y por último presentan los resultados de los principales competidores para la empresa que se está analizando. Todos los modelos que se han generado en cada investigación ayudan a una empresa involucrada a tener una ventaja competitiva, sin embargo, todas estas investigaciones están destinadas para grandes empresas, no toman en cuenta el idioma español, y realizan el etiquetado de manera manual al no contar con un corpus para entrenar sus modelos.

Existen otras investigaciones como Baviera Puig (2017), Cedeno-Moreno & Vargas (2020), Martínez Cámara et al. (2011), Miranda et al. (2016), Peñalver-Martínez et al. (2011), Reyes-Ortiz et al. (2017), Salazar Loor & Ponce Intriago (2018), Vilares et al. (2013) que se enfocaron en la aplicación de técnicas de minería de textos y minería de opiniones exclusivamente para el lenguaje español, en estas investigaciones se crearon modelos para determinar sentimientos de comentarios de diferentes plataformas, pero ninguno presentó modelos para realizar detección de competidores al no contar con un corpus para entrenar sus modelos.

La minería de opiniones comparativa es definida en Varathan et al. (2017) como un subcampo de la minería de opiniones que ayuda a identificar y extraer información que se exprese de manera comparativa en un texto, por ejemplo: “El producto A es mejor que el producto B”, “la empresa A tiene mejor atención que la empresa B”, etc. Este campo contribuye a la inteligencia competitiva para ayudar a las organizaciones empresariales a identificar riesgos y mercados potenciales en las primeras etapas (Xu et al., 2011), por lo tanto, este campo es muy importante dentro del análisis de datos de CI especialmente en la detección de competidores y/o productos competitivos, ya que como se menciona en Varathan et al. (2017) la CI a menudo está muy restringida por la falta de suficientes fuentes de información sobre los competidores, por lo que, es importante poder identificar las fuentes de información en que se están mencionando a posibles competidores y/o posibles productos competitivos. En la literatura de la minería de opiniones comparativa no existen investigaciones que tomen en cuenta el idioma español, además, tampoco existen corpus en español que ayuden a esta tarea. Un corpus en un idioma determinado es un conjunto de datos y un único conjunto de datos anotado se denomina corpus anotado. Los corpus anotados se pueden usar para entrenar algoritmos de aprendizaje automático (Pustejovsky & Stubbs, 2013).

En resumen, las áreas de investigación mencionadas son importantes para la detección de competidores en la CI; sin embargo, en la literatura actual, la mayoría de estudios están destinados para grandes empresas y principalmente los modelos están desarrollados para el idioma inglés, por lo que generar un corpus que ayude a la detección de competidores en el idioma español es de gran relevancia; además, como se menciona en Koseoglu et al. (2011), Nenzhelele & Pellissier (2014) se recomienda aplicar CI a pequeñas y medianas empresas (PYMEs) debido a que una PYME es vulnerable a la competencia directa de una empresa más grande. Como se ejemplifica en Ponis & Christou, (2013) el pez grande se come al pequeño, excepto en aquellos casos en los que el pez pequeño muestra una inteligencia significativa. Por todo lo mencionado, generar corpus aptos para

luego poder generar modelos de aprendizaje automático que permitan identificar posibles competidores en el idioma español es un tópico nuevo que debe ser abordado a profundidad.

1.3 Objetivos

El objetivo de esta investigación es la creación de un corpus en el idioma español —para la fase de análisis de datos en el proceso de CI— que permita crear modelos para detectar posibles competidores. Este corpus generado será evaluado en un caso de estudio considerando empresas textiles pertenecientes al proyecto "Incorporating sustainability concepts to management models of textile Micro, Small and Medium Enterprises (SUMA)".

1.3.1 Objetivos Específicos

El presente trabajo tiene los siguientes objetivos específicos:

- Determinar una forma de priorizar los corpus existentes en el idioma español de acuerdo a la relevancia y utilidad en la inteligencia competitiva.
- Generar un corpus en base al ranking de relevancia y la identificación de texto comparativo.
- Crear modelos de aprendizaje automático con el corpus generado para determinar posibles competidores.
- Implementar un Dashboard para visualizar la detección de posibles competidores y realizar una evaluación de la utilidad de estos resultados en un caso de estudio considerando empresas textiles pertenecientes al proyecto SUMA.

CAPÍTULO 2: MARCO TEÓRICO

En este capítulo se describe la definición de corpus, posteriormente se analiza las técnicas para la recolección de datos y plataformas utilizadas en la extracción de los mismos. También se realiza una descripción a profundidad de la Inteligencia Competitiva, su proceso y la detección de competidores. Después, se hace un acercamiento a la minería de textos, tomando en cuenta las técnicas del procesamiento del lenguaje natural y aprendizaje automático que la ayudan a cumplir sus objetivos dentro de la minería de opiniones, minería de opiniones comparativas y el reconocimiento de entidades. Por último, se hace una revisión de las herramientas tecnológicas que se han utilizado para cumplir con los objetivos de este trabajo de titulación.

2.1 Definición de Corpus

Existen varias definiciones de un corpus, una de ellas lo define Atkins et al. (1992) como un subconjunto de una colección de textos electrónicos que están en formato estandarizado con ciertas convenciones relacionadas con el contenido, pero sin restricción de selección rigurosa, que está construido de acuerdo a criterios de diseño explícitos para un propósito específico. Por su parte, Parodi (2008) lo define como una colección o conjunto de textos que comparten ciertos rasgos definitorios, limitados sólo por características inherentes a la naturaleza de los mismos. Así también, Huang & Yao (2015) lo define como una colección de ejemplos de lenguaje en uso que se

seleccionan y compilan de manera basada en principios, la intención es que sea un cuerpo de evidencia representativo para el estudio del lenguaje y el uso del lenguaje. El objetivo de un corpus es servir como una herramienta útil para descubrir muchos aspectos del uso del lenguaje que, de otro modo, podrían pasar desapercibidos (Mike, 2010).

Los corpus pueden ser de diferentes tamaños y deben estar en concordancia con los objetivos de cada investigación, los corpus pequeños (por ejemplo 1 millón de palabras) generalmente sirven para estudios de construcciones sintácticas de alta frecuencia pero suelen ser inadecuados para el estudio de fenómenos léxicos (Jerga, Vulgarismos, entre otros) y semánticos mientras que los "mega corpus" que pueden tener miles de millones de palabras de páginas web fáciles de obtener son a menudo un "mancha" de textos, que no tiene una estructura que se preste al estudio de la variación dialectal (Mark Davies, 2019). En la actualidad existe un creciente interés en la construcción y análisis de Corpus debido a la demostración que el procesamiento estadístico de corpus de texto es un enfoque viable para algunos de los problemas difíciles tradicionales de la lingüística computacional, la traducción automática y la ingeniería del conocimiento (Atkins et al., 1992; Pustejovsky & Stubbs, 2013).

La tarea de construir un corpus, dependiendo de los tipos de preguntas de investigación que se aborden, demanda un trabajo metodológico y puede ser una tarea razonablemente eficiente y restringida, o puede llevar mucho tiempo. Un corpus de alto valor debe cumplir con características de: tipo, tamaño y composición para una investigación delimitada. Sin embargo, una limitante grave en la construcción y uso de corpus es la legislación nacional e internacional sobre derechos de autor, ya que en la mayoría de los casos en necesario obtener un permiso de derechos de autor para que los textos puedan ser computarizados (Atkins et al., 1992; Mike, 2010).

2.2 Técnicas para la Recolección de Datos

En la actualidad con el avance de la tecnología existen algunas técnicas para la extracción de datos, a continuación, se detalla algunas de las principales técnicas: Application Programs Interfaces (API) oficiales de las plataformas digitales, Web Scraping y también existen corpus que están distribuidos a través de repositorios.

2.2.1 API de Plataformas Digitales

Una API es una herramienta que permite a terceros tener la capacidad de consultar y obtener acceso a porciones de información que son datos generados por los usuarios al usar y experimentar la plataforma digital. Cada API tiene detallada sus reglas mediante las cuales el software se va a comunicar, articulando qué elementos se pueden consultar, con qué frecuencia y cómo aparecen los resultados (Acker & Kreisberg, 2020). En estos últimos años debido a la pandemia el comercio electrónico se ha incrementado enormemente. La Cámara Ecuatoriana de Comercio Electrónico demuestra en un estudio que el sector textil tuvo un crecimiento de al menos un 43.73% en el año 2020 en comparación con el año 2019 (E-commerce, 2020). Las principales fuentes de información para estudios relacionados a este sector como el estudio de mercado y el análisis de competencia son las plataformas digitales de comercio electrónico y las redes sociales, debido a que estas

plataformas permiten a los usuarios tener un medio donde pueden expresar sus opiniones positivas o negativas de un producto o una empresa. Esto ha sido un factor importante para que las empresas reconozcan la importancia de estas plataformas para los estudios de mercado (Tripathi & S, 2015). Cada API tiene sus respectivas restricciones para los usuarios, en los últimos años debido a problemas con mal uso de datos, las APIs especialmente de redes sociales se han restringido enormemente (Perriam et al., 2020).

2.2.2 Web Scraping

Web Scraping tiene muchas definiciones, por ejemplo Molina et al. (2015) lo presenta como una solución tecnológica que sirve para extraer datos de sitios web, de manera rápida, eficiente y automatizada, ofreciendo datos de una formato más estructurado y fácil de usar. Estos procesos pueden estar desarrollados en diferentes lenguajes de programación y generalmente se desarrollan para cumplir tareas específicas debido a que un desarrollo con esta técnica conlleva una alta especialización.

En la actualidad existen muchos Bots (aplicación de software que ejecuta tareas automatizadas a través de Internet) desarrollados con web scraping, estos pueden ser buenos y malos. Un Bot bueno es aquel que permite la extracción de datos para investigaciones como comparación de precios para ahorrar dinero del cliente, reconocer el sentimiento en las redes sociales y muchas otras aplicaciones muy útiles, mientras que un Bot malo es aquel que se usa para actividades como la recuperación de datos privados de los consumidores, como números de teléfono y correos electrónicos, secuestro de cuentas, spam y fraude de datos digitales (Krotov & Silva, 2018). Motivo por el cual muchas empresas de tecnología no permiten hacer web scraping, por ejemplo Facebook tiene implementado mecanismos para evitar esta técnica de extracción de datos, incluso son capaces de emprender acciones legales contra quienes están haciendo esta actividad de métodos automatizados dentro de su dominio (Mancosu & Vegetti, 2020).

2.2.3 Corpus Disponibles en Repositorios

Existen también varios corpus orientados hacia diferentes objetivos e incluso de diferentes idiomas que varios investigadores, organizaciones o empresas han puesto para el público bajo diferentes tipos de licencias. En el capítulo 1 se habló de la limitación que se tiene con respecto a este recurso en el idioma español, se determinó que no existen una cantidad grande de corpus en el idioma que los investigadores puedan utilizar para estudios orientados hacia la inteligencia competitiva; sin embargo, esta investigación parte de 4 corpus (orientados en su mayoría a minería de opiniones) que se han encontrado y se ha podido obtener acceso. Se ha tomado la descripción de cada corpus donde se mencionan los objetivos para los que han sido construidos cada uno y un análisis previo de sus datos para hacer la selección de los corpus más idóneos para esta investigación. Los corpus disponibles en español que se han utilizado en esta investigación tienen licencia Creative Commons (CC) la cual es una de las varias licencias públicas de derechos de autor que permiten la distribución gratuita de una "obra" protegida por derechos de autor y los otros dos corpus restantes tiene una licencia propia que permite que los datos estén de manera gratuita para el usuario final.

2.3 Descripción de las Plataformas Digitales

En el artículo de investigación denominado como Plataforma para Análisis de Mercado a través de Datos de Redes Sociales (Fajardo Cárdenas et al., 2021) se presenta como resultado que las principales plataformas digitales para el sector textil son la redes sociales, entre ellas está Facebook, Twitter, y YouTube, por lo tanto, a continuación se presenta un descripción de cada una de ellas y sus limitantes con respecto a la API de cada plataforma respectivamente.

2.3.1 Facebook

Facebook es una de las plataformas más grandes de redes sociales ocupando el primer lugar en las redes sociales más utilizadas en Ecuador (Del Alcázar, 2021), donde muchas personas publican contenido a cada instante, al mismo tiempo también dan retroalimentaciones a través de "me gusta", comentarios o debates sobre publicaciones de su interés. Incluso según Similar Web (SimilarWeb, 2021), Facebook es la red social más visitada en el año 2021 a nivel mundial. Sin embargo, su acceso a datos mediante la API está bastante limitada, especialmente después del año 2018, cuando fue involucrada con la Consultora británica Cambridge Analytics que estaba haciendo uso de datos sin consentimiento de los usuarios de la plataforma de red social (ur Rehman, 2019). Debido al problema mencionado la empresa ha actualizado sus políticas de seguridad de datos y en la actualidad para acceder a datos de Facebook mediante la API los desarrolladores necesitan tener desarrollada la aplicación, también tener definidas las políticas de uso de datos en dicha aplicación, luego registrarse en la plataforma y entregar su aplicación para que sea revisadas y sí cumple con las políticas de Facebook será aprobada y podrá hacer uso de sus datos (Facebook, 2018).

2.3.2 Twitter

Twitter es una de las redes sociales que permite a usuarios publicar e interactuar con mensajes conocidos como "tweets", los usuarios registrados también pueden dar "me gusta" y "retuitear". Esta red social es la segunda más utilizada a nivel mundial según Similar Web (SimilarWeb, 2021), y según (Del Alcázar, 2021) en Ecuador es la quinta red social más utilizada. Twitter dispone de tres APIs para la extracción de datos. La API Search es un servicio gratuito que permite al software enviar búsquedas automáticamente a Twitter y recuperar tweets coincidentes, la API stream por su parte permite recuperar tweets en tiempo real, y la API firehouse es una versión pagada de las dos anteriores, donde el precio según la documentación oficial de Twitter es de $149 por mes para 500 peticiones (Twitter, 2019). La diferencia radica en que las versiones gratuitas solo permiten acceder a tweets históricos de máximo una semana, mientras que en la API firehouse se pueden obtener tweets de cualquier intervalo de fechas. Otra diferencia es la limitación que se tiene con el número de peticiones a la API que se pueden realizar.

2.3.3 YouTube

YouTube es una red social que está centrada en la publicación de videos para que otros usuarios los miren y reaccionen con “me gusta”, “no me gusta”, emojis y también pueden hacer comentarios, los cuales son del texto generado por el usuario que será de interés para esta investigación. Esta red social también es una de las más visitadas en el mundo y es la segunda más visitada en Ecuador según Similar Web (SimilarWeb, 2021). YouTube dispone de una API para la extracción de datos de videos llamada API Data V3, en la cual se puede obtener información de videos como el título, comentarios, número de vistas, “me gusta”, “no me gusta”, fecha en que se ha publicado el video, fecha de publicación de los comentarios, entre otros. La API permite buscar videos que coincidan con términos de búsqueda específicos, ubicaciones y fechas de publicación (Google Developers, 2021).

2.4 Inteligencia Competitiva (CI)

La definición de CI varía según diferentes autores y enfoques en el campo de los negocios. En Zanasi (1998) se la define como datos oportunos y basados en hechos en los que la dirección (líderes de la empresa) puede basarse para la toma de decisiones y el desarrollo de estrategias. Según Bulley et al. (2014) la CI es un proceso que conduce a la generación de la información de la competencia y el entorno industrial para la planificación y toma de decisiones. Otra de varias definiciones indica la CI como el proceso de seguimiento del entorno competitivo y competidores de la empresa, en el cual la definición, recopilación de información, análisis y distribución de los resultados obtenidos se realiza de forma paulatina para que puedan apoyar la actividad empresarial eficiente y su capacidad para tomar decisiones calificadas (Mclean & Woods, 2014).

CI en general puede ser considerado un proceso de planificación y gestión estratégica para toda empresa (Bose, 2008). Un proceso de CI efectivo, según SCIP, se ejecuta en un ciclo continuo, llamado ciclo de CI (**Figura 1**). El SCIP describe al ciclo de CI como el proceso mediante el cual la información en bruto se adquiere, recopila, transmite, evalúa, analiza y pone a disposición como inteligencia completa para que los formuladores de políticas la utilicen en la toma de decisiones y la acción. Según el SCIP, existen cinco fases que constituyen este ciclo:

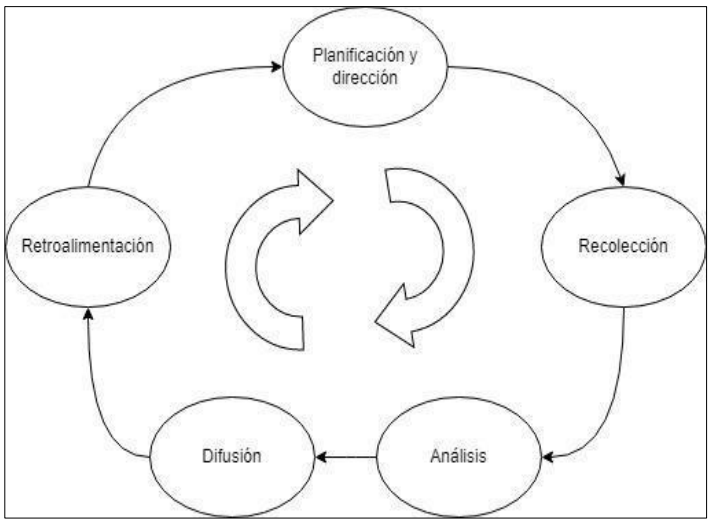


***Figura 1**: Proceso de Inteligencia Competitiva.*

Fuente: Construcción Propia

1. **Planificación y Dirección. -** En esta fase se definen los requisitos de la empresa en términos de la información que se necesita: utilidad y disponibilidad (Bose, 2008). En esta etapa se involucran analistas de CI y tomadores de decisiones, además, implica trabajar para descubrir sus necesidades de inteligencia y luego traducir esas necesidades en sus requisitos de inteligencia específicos o "temas clave de inteligencia" (KITs) (Nasri, 2011).

2. **Recolección. -** Este paso es uno de los más importantes en el proceso de CI (Araujo et al., 2017). Aquí se identifican todas las fuentes potenciales de información, luego se investiga y recopila los datos correctos legal y éticamente de todas las fuentes disponibles y ponerlos en forma ordenada (Herring, 1998). Algunas de las fuentes de información disponibles para esta fase son: sitios de compras en línea, como Amazon; sitios de reseñas de clientes, como opiniones; blogs; sitios de redes sociales; y correos electrónicos (Bose, 2008).

3. **Análisis. –** Es un paso crucial en el proceso de CI (Nasri, 2011), las actividades en esta fase implican convertir la información en "inteligencia procesable" sobre la cual se pueden tomar decisiones estratégicas y tácticas (Gilad & Gilad, 1985; Herring, 1998). El análisis abarca un examen sistemático de los datos, la información y el conocimiento recopilados, para determinar su aplicabilidad o importancia, y la transformación de los resultados en inteligencia procesable que mejora la planificación y la toma de decisiones o permitirá el desarrollo de estrategias que ofrezcan una ventaja competitiva sostenible (Bose, 2008).

4. **Difusión (reporte, informe). -** Es el producto terminado o la inteligencia competitiva comunicada a los tomadores de decisiones en un formato que es fácilmente comprensible (Nasri, 2011). A menudo, la comunicación de los hallazgos toma la forma de un informe, un Dashboard o una reunión (Bose, 2008). Estos hallazgos se utilizan como entradas para realizar análisis adicionales, como la elaboración de perfiles de la competencia, la planificación de escenarios y el análisis de escenarios (Nasri, 2011). En esta fase, la información debe presentarse de manera simple, destacando sólo lo que el gerente o tomador de decisiones necesite saber con una buena visualización de datos (Araujo et al., 2017).

5. **Retroalimentación (evaluación). -** Esta fase implica medir el impacto de la inteligencia que se proporcionó a los tomadores de decisiones (Bose, 2008). Se intenta dar respuestas a las siguientes preguntas: ¿Fue usado?, ¿Cómo o por qué no?, ¿Resultó en hacer un trato?, ¿Ahorró dinero?, ¿Para impulsar la reputación de la empresa?, ¿Cómo se puede ajustar el proceso?. Por lo tanto, proporcionan al analista áreas importantes para la mejora continua o una mayor investigación (Araujo et al., 2017).

El proceso descrito es la forma más general de hacer CI. Este, a su vez, es un proceso cíclico y no una secuencia lineal. Si el gerente o tomador de decisiones necesita nuevos datos se debe repetir el proceso y siempre ir mejorando en cada fase (Araujo et al., 2017; Bose, 2008). Otra consideración importante a tomar en cuenta es si se debe incorporar una unidad de inteligencia competitiva

centralizada o descentralizada, y si la inteligencia competitiva requiere atención a tiempo completo (Kahaner, 1997). Havenga & Botha (2003) han demostrado que la función de CI debe ubicarse lo más alto posible en la organización y debe tener acceso directo, sin filtros, al director ejecutivo.

2.5 Detección de Competidores

La detección de competidores implica reconocer quiénes son las posibles empresas competidoras en el sector o rubro del negocio en que participa la empresa y quiénes no lo son. En la actualidad, con el auge de las redes sociales y otras plataformas como los sitios de comercio electrónico, cada día existen una gran cantidad de datos generados por usuarios que pueden ser de interés para detectar a la posible competencia y así tomar decisiones informadas, tales como: i) conocer las debilidades y fortalezas relativas de sus propios productos, y ii) tomar acciones como diseñar nuevos productos y campañas específicas para contrarrestar las de sus competidores (Xu et al., 2011). Según Gao et al. (2018) generalmente el análisis para realizar detección de competidores se utiliza encuestas con análisis de contenido aunque también en una investigación presentada por Peng & Liang (2016) se realiza con datos extraídos de redes sociales, pero solo se enfocan en empresas grandes como Microsoft, Apple, Samsung etc., y en el idioma inglés.

2.6 Conceptos de Aprendizaje Automático

Las técnicas de NLP, minería de textos y aprendizaje automático trabajan juntas para clasificar y descubrir automáticamente patrones en textos (Aurangzeb et al., 2011). El Aprendizaje Automático es una rama en constante evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno que los rodea (El Naqa & Murphy, 2015). Gran parte de la investigación del aprendizaje automático está inspirada en problemas importantes de la biología, la medicina, las finanzas, la astronomía, etc. y la sociedad (El Naqa & Murphy, 2015; Wagstaff, 2012).

El objetivo principal del aprendizaje automático es modelar la relación entre un conjunto de cantidades observables (entradas) y otro conjunto de variables que están relacionadas con estas (salidas) (Baştanlar & Özuysal, 2014). Una vez que se determina dicho modelo matemático, es posible predecir el valor de las variables deseadas midiendo los observables, es decir realiza predicciones exitosas utilizando experiencias pasadas (Bonaccorso et al., 2018). Desafortunadamente, muchos fenómenos del mundo real son demasiado complejos para modelarlos directamente como una relación de entrada-salida de forma cerrada (Baştanlar & Özuysal, 2014). El aprendizaje automático proporciona técnicas que pueden construir automáticamente un modelo computacional de estas relaciones complejas mediante el procesamiento de los datos disponibles y la maximización de un criterio de rendimiento dependiente del problema. El proceso automático de creación de modelos se denomina "entrenamiento" y los datos utilizados con fines de entrenamiento se denominan "datos de entrenamiento" (El Naqa & Murphy, 2015). El modelo entrenado puede proporcionar nuevos conocimientos sobre cómo las variables de entrada se asignan a la salida y se puede usar para hacer

predicciones para nuevos valores de entrada que no formaban parte de los datos de entrenamiento (Baştanlar & Özuysal, 2014).

Las técnicas de aprendizaje automático tienen muchas clasificaciones, pero las dos categorías principales son aprendizaje supervisado y aprendizaje no supervisado (Baştanlar & Özuysal, 2014) (**Figura 2**). Debido al alcance y objetivos de este trabajo de titulación, nos enfocaremos en el aprendizaje supervisado.

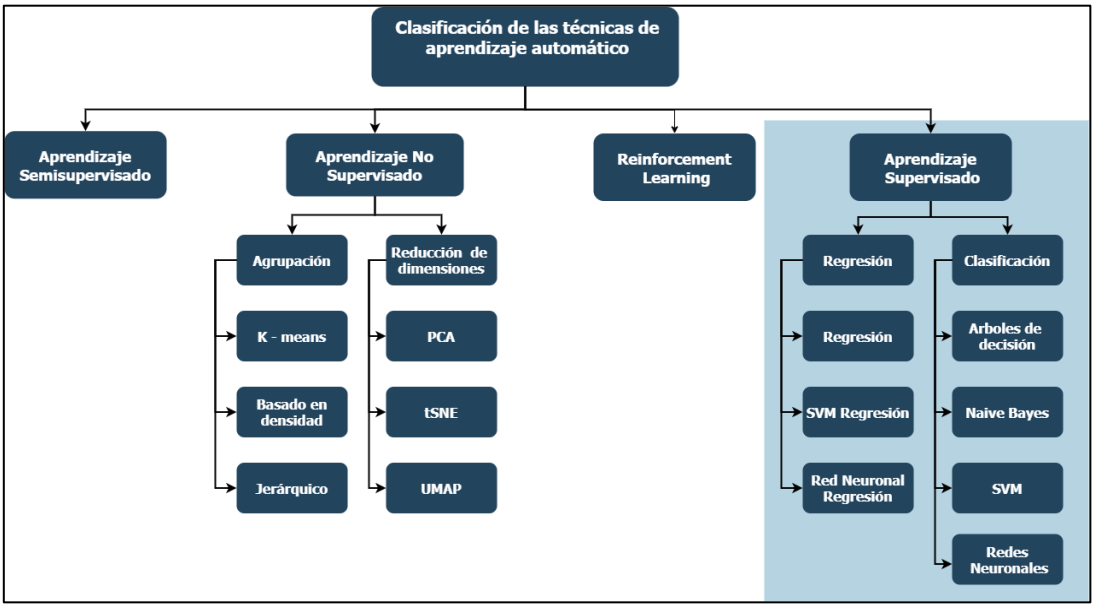


Figura 2: *Descripción general de los principales algoritmos de aprendizaje automático. Fuentes: (Badillo et al., 2020; Mahesh, 2018). Construcción Propia*

2.6.1 Aprendizaje Supervisado

El aprendizaje supervisado es la tarea de aprendizaje automático capaz de aprender una función que asigna una entrada a una salida en base a pares de entrada-salida de ejemplo. Infiere una función a partir de datos de entrenamiento etiquetados que consisten en un conjunto de ejemplos de entrenamiento (Badillo et al., 2020). El conjunto de datos de entrada se divide en conjunto de datos de entrenamiento y prueba.

2.7 Conceptos de Minería de Textos

Para identificar los riesgos potenciales, es importante que las empresas recopilen y analicen información sobre los productos y planes de sus competidores (Xu et al., 2011). Minería de Textos (Text Mining) definido por Cardoso et al. (2019) como un proceso que sirve para la extracción de información interesante y no trivial de textos no estructurados como por ejemplo datos de plataformas digitales ayudan a cumplir estos objetivos. Según Allahyari et al. (2017), Berry & Castellanos (2008),Hotho et al. (2005), Tandel et al. (2019) la minería de textos hace uso del NLP y el aprendizaje automático para llevar a cabo técnicas como la minería de opiniones, minería de opiniones comparativas y NER.

2.7.1 Procesamiento de Lenguaje Natural (NLP)

NLP es un área de investigación y aplicación que explora cómo se pueden usar las computadoras para comprender y manipular el texto o el habla en lenguaje natural para hacer cosas útiles (Gobinda, 2003). Los métodos de NLP se pueden aplicar para analizar el lenguaje en dos niveles diferentes: análisis sintáctico y análisis semántico. El análisis sintáctico transforma y analiza la sintaxis de las oraciones. El análisis semántico tiene como objetivo identificar y analizar el significado de palabras, frases y oraciones (Varathan et al., 2017).

2.7.2 Minería de Opiniones

La minería de opiniones o también denominada análisis de sentimientos es el estudio computacional de las opiniones, valoraciones, actitudes y emociones de las personas hacia entidades, individuos, problemas, eventos, temas y sus atributos. Estos estudios generalmente se realizan ya que las empresas siempre quieren conocer las opiniones ya sean positivas o negativas del público o de los consumidores sobre sus productos y servicios. Por parte de los clientes potenciales también es de interés conocer las opiniones de los usuarios existentes antes de utilizar un servicio o comprar un producto (B. Liu et al., 2012). Por su parte, Cedeno-Moreno & Vargas (2020) lo define como una tarea del PLN que sirve para identificar posturas de opiniones relacionadas con un objeto dentro de un contexto común. Esta tarea descubre de forma eficaz el conocimiento a través de los comentarios expresados, especialmente en el contexto de la web, donde extrae las opiniones, sentimientos y demandas de los usuarios de los textos y distingue su polaridad (Hemmatian & Sohrabi, 2019). La minería de opiniones, especialmente su subcampo llamado minería de opiniones comparativas es ampliamente utilizada en las fases de recolección y análisis de CI.

2.7.3 Minería de Opiniones Comparativas

La minería de opiniones comparativas definida por Varathan et al. (2017) como un como un subcampo de minería de opiniones, este subcampo se ocupa de identificar y extraer información que se expresa de forma comparativa, contribuye a la inteligencia competitiva para ayudar a las organizaciones empresariales a identificar riesgos y mercados potenciales en las primeras etapas (Xu et al., 2011). Las opiniones de los clientes suelen ser una rica fuente de opiniones comparativas. Por ejemplo, los usuarios generalmente prefieren comparar varios productos de la competencia con funciones similares, una demostración de esto son los siguientes comentarios extraídos de una plataforma digital:

- *El Nokia N95 tiene una señal más fuerte que el iPhone.*
- *El iPhone tiene mejor apariencia, pero un precio mucho más alto que el Samsung.*

Estas opiniones comparativas son valiosas fuentes de información para identificar las fortalezas y debilidades relativas de los productos, analizar el riesgo empresarial y las amenazas de la

competencia y seguir diseñando nuevos productos y estrategias comerciales (Xu et al., 2011). La extracción de tales opiniones comparativas no es una tarea trivial debido a la gran cantidad de comentarios de los clientes y su estilo informal.

Debido a que gran parte de los datos web son textos no estructurados, la minería de textos plantea una solución ideal a este problema por su capacidad para descubrir conocimientos y patrones a partir de una gran cantidad de datos de texto (Feldman & Sanger, 2007). En la detección de competidores es importante diferenciar los datos que contienen opiniones comparativas, para esto se utiliza una gran variedad de técnicas de minería de opiniones, NLP y aprendizaje automático.

2.7.4 Reconocimiento de Entidades Nombradas (NER)

NER es la tarea de localizar y categorizar sustantivos importantes y/o nombres propios en un texto como por ejemplo personas, ubicaciones, organizaciones, fármacos, tiempos, etc. Los sistemas NER se utilizan a menudo en la recuperación de información, la resolución de correferencias, traducción automática, el modelado de temas, entre otros. A lo largo de los años se han desarrollado numerosos sistemas y recursos de datos para el NER. Además, ha habido varios foros y programas de evaluación centrados en NER y otras tareas relacionadas (Mohit, 2014; V. Yadav & Bethard, 2018).

2.8 Técnicas y Métodos de NLP que se aplican en Minería de Textos.

La minería de textos para poder llevar a cabo sus tareas y objetivos usa muchas técnicas y métodos del NLP, a continuación, se detalla algunas de ellas.

2.8.1 Normalización

Normalización o escalado de atributos es una fórmula para escalar los datos cuando atributos diferentes tienen una alta diferencia en sus valores máximos y mínimos. Esto es un requisito común para muchos modelos de aprendizaje automático: pueden proveer resultados negativos si las características individuales no se asemejan a los datos estándar normalmente distribuidos (Sieminski et al., 2018). Entre las técnicas de normalización en la minería de textos se encuentran los stopwords, stemming, y la lematización.

2.8.2 StopWords

StopWords puede identificarse como una palabra o término que tiene la misma probabilidad de aparecer en aquellos documentos que no son relevantes para una consulta que en aquellos documentos relevantes para la consulta (Wilbur & Sirotkin, 1992). Para garantizar la precisión y la eficiencia de tareas de Minería de textos como la clasificación de textos, reconocimiento de entidades, las StopWords, a menudo denominadas "palabras vacías", se eliminan en el paso de preprocesamiento ya que estas generan ruido (Sarica & Luo, 2021). El ruido se puede entender como

un contenido que no aporta o que no es aplicable a la tarea a realizar (Nisbet et al., 2017). Algunos ejemplos de StopWords en el lenguaje español incluye: "tal", “la”, y "el".

2.8.3 Stemming y Lematización

Stemming es un paso previo al procesamiento en las aplicaciones de minería de textos, así como un requisito muy común de las funciones de NLP. Busca el manejo automático de las terminaciones de las palabras reduciendo las palabras a sus raíces, en el momento de la indexación y la búsqueda (Anjali et al., 2007).

La lematización es otra técnica de normalización, para cada forma de palabra en un documento o texto, se identifica su forma básica, el lema (Balakrishnan & Lloyd-Yemoh, 2014). Tanto el Stemming como la Lematización juegan un papel muy importante cuando se trata de aumentar la relevancia y obtener mejores resultados en tareas de minería de textos (Balakrishnan & Lloyd-Yemoh, 2014). Por ejemplo, cuando se lematiza la palabra “ganar”, su índice puede usarse para “gana, ganamos, ganan”, etc.

La Lematización es similar a Stemming, pero no requiere producir una raíz de la palabra sino reemplazar el sufijo de una palabra, que aparece en texto libre, con un sufijo de palabra diferente para obtener la forma de la palabra normalizada (Plisson et al., 2004). Por ejemplo, al analizar el conjunto de palabras: “gano”, “ganas”, “gana”, “ganamos”, “ganan” el Stemming y Lematización se presenta en **Tabla 1**.

Tabla 1: *Aplicación de Stemming y Lematización a un conjunto de palabras.*

Palabra Original	Lematización	Stemming
gano	gano	gan
ganas	gana	gan
gana	ganar	gan
ganamos	ganar	gan
ganan	ganar	gan

Como se puede observar en **Tabla 1**, el Stemming tiene como ventaja el reconocimiento de relaciones entre palabras de distinta clase. Por ejemplo, podría reconocer que picante y picar tienen como raíz “pic”. Una desventaja del Stemming es que puede “cortar” demasiado la raíz y encontrar relaciones de palabras inexistentes. Lematización garantiza que los lemas tengan relación, pero en cambio no siempre reconoce todas las relaciones como se puede ver en el ejemplo.

2.8.4 Part of Speech Tagging (POS)

En gramática, POS es una categoría lingüística definida por su comportamiento sintáctico o morfológico. Las categorías comunes de POS son: sustantivo, verbo, adjetivo, adverbio, pronombre, preposición, conjunción e interjección. Las etiquetas POS importantes para este trabajo y sus categorías son: NN: Sustantivo, NNP: Sustantivo propio, PRP: Pronombre, VBZ: Verbo, tiempo

presente, 3a persona del singular, JJR: Adjetivo comparativo, JJS: Adjetivo superlativo, RBR: Adverbio comparativo, RBS: Adverbio superlativo (Jindal & Liu, 2006a). POS ha sido utilizado en algunos procedimientos de minería de textos, por ejemplo Hu & Liu (2004) utilizó esta técnica para extraer los atributos del producto, y luego juzgar las polaridades de las frases de opinión sobre los atributos en función de la información del contexto. Las etiquetas POS de palabras, como adjetivos y adverbios, han sido buenos indicadores para la detección de subjetividad y clasificación de polaridad de sentimientos (Argamon et al., 2007; Turney, 2002). POS también ha tenido varios usos en la minería de opiniones comparativas, Kessler & Kuhn (2014) lo utiliza para identificar el predicado comparativo en un texto, aquí se menciona que este predicado es la parte central de cualquier comparación. (Kessler & Kuhn, 2013) también utilizó POS para la identificación de entidades y por otra parte Xu et al. (2009) usó POS para entrenar un modelo destinado a identificar y categorizar las relaciones comparativas.

2.8.5 Bag of Words (BOW)

El BOW es una representación simplificada que se utiliza en el procesamiento del lenguaje natural. En este modelo un texto se representa como una colección desordenada de sus palabras, sin tener en cuenta la gramática. En el caso de la clasificación de texto, a una palabra en un documento se le asigna un peso de acuerdo con su frecuencia en el documento y la frecuencia entre diferentes documentos (George K & Joseph, 2014). Lu (2015) y Xing et al. (2018) han utilizado este modelo para transformar textos a una representación entendible para los algoritmos de aprendizaje automático utilizados en la minería de textos.

2.9 Algoritmos de Aprendizaje Supervisado en Minería de Textos

El aprendizaje automático se basa en diferentes algoritmos para resolver problemas de datos (Mahesh, 2018). El aprendizaje supervisado es utilizado ampliamente en la detección de competidores (Zhang, 2020), por lo que se definirán algunos de los algoritmos más utilizados.

2.9.1 Random Forest

Random Forest es un algoritmo de aprendizaje automático basado en conjuntos y se utiliza tanto para tareas de regresión como de clasificación. Los resultados de Random Forest están enfocados en cada árbol de decisión. Cuanto mayor sea el número de árboles de decisión en Random Forest, mejor será la generalización (Nayak, 2016). En la **Figura 3** se presenta un árbol de decisión básico y el proceso que sigue el algoritmo es el siguiente: el árbol comienza con variable_1 y se divide en función de criterios específicos. Cuando es 'sí', el árbol de decisión sigue el camino representado, cuando es 'no', el árbol de decisión sigue el otro camino. Este proceso se repite hasta que el árbol de decisión llega al nodo hoja y se decide el resultado final.

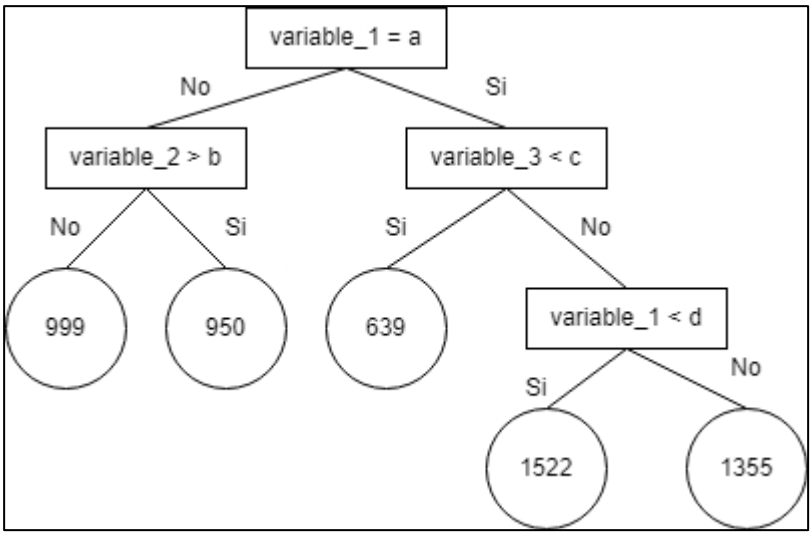


Figura 3: *Diagrama de un árbol de decisión básico.*
Fuente. Construcción Propia

2.9.2 Support Vector Machine (SVM)

SVM es un algoritmo de aprendizaje supervisado que se puede emplear para para regresión o clasificación binaria. Este algoritmo es un tipo de clasificador de gran margen. SVM está basado en un espacio vectorial donde el objetivo es encontrar una frontera de decisión entre dos clases que estén lo más alejadas posible de cualquier punto de los datos de entrenamiento (posiblemente descontando algunos puntos como valores atípicos o ruido). Este algoritmo presenta dos casos: i) cuando el conjunto de datos de dos clases son separables por un clasificador lineal (**Figura 4**) y ii) cuando el conjunto de datos no es separable de manera lineal, a estos se los denomina problemas multiclase y modelos no lineales (Manning et al., 2008).

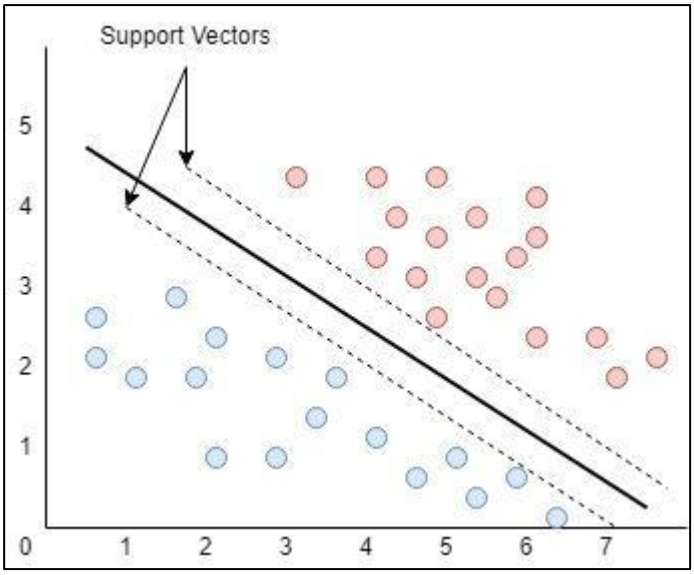


Figura 4*: Hiperplano en 2D de Support Vector Machine.*

Fuente: Construcción Propia

La ecuación del hiperplano para cada clase y, y puntos x tiene las siguientes restricciones:

$$(\mathbf{w} * \mathbf{xi} + \mathbf{b}) \geq \mathbf{1}, \mathbf{Si\ yi} = \mathbf{1} \quad \textbf{(1)}$$

$$(\mathbf{w} * \mathbf{xi} + \mathbf{b}) \leq -\mathbf{1}, \mathbf{Si\ yi} = -\mathbf{1} \quad \textbf{(2)}$$

Donde tanto en la ecuación (1) y ecuación (2) la b es una constante, w se denomina vector de pesos y ||w|| se minimiza para maximizar la separación entre las clases. Al implementar modelos SVM, se deben proporcionar parámetros como C, gamma, etc., para obtener la máxima precisión, teniendo en cuenta la compensación entre sesgo y varianza (Ismail et al., 2016).

El dilema de sesgo-varianza según Sokolova et al. (2006) se trata con frecuencia en los algoritmos de aprendizaje supervisado, ya que tiende a generalizar más allá de los datos de entrenamiento. El error de sesgo conduce a un ajuste inadecuado de los datos, ya que pierde una interacción importante entre las características y las clases. Por otro lado, el error de varianza conduce a un sobreajuste debido a que es muy sensible al ruido y las fluctuaciones que pueden estar presentes en el conjunto de entrenamiento. La precisión de un modelo depende en gran medida de estos parámetros, por lo tanto, los valores óptimos se encuentran realizando un tuneado de hiper parámetros (Ismail et al., 2016).

2.9.3 Naive Bayes

El Naive Bayes es un algoritmo de aprendizaje simple que utiliza la regla de Bayes (Ecuación (3)) junto con una fuerte suposición de que los atributos son condicionalmente independientes dada la clase. Aunque esta suposición de independencia se viola a menudo en la práctica, el Naive Bayes suele ofrecer una precisión de clasificación competitiva. Esto, unido a su eficiencia computacional y a muchas otras características deseables, hace que Naive Bayes se aplique ampliamente en la práctica (Webb, 2016).

$$\boldsymbol{P(y \mid x) = P(y)P(x \mid y) \ / p(x)} \quad \textbf{(3)}$$

Naive Bayes es usado con frecuencia por parte de los investigadores en la minería de opiniones comparativas (Varathan et al., 2017). Este clasificador es uno de los clasificadores más populares utilizados para la clasificación de texto debido a que se basa en el teorema bayesiano y funciona bien cuando hay un gran número de dimensiones (Mohri et al., 2012).

2.9.4 Regresión Logística

La regresión logística tiene como objetivo clasificar los datos en diferentes clases (etiquetas) de polaridad en función de los conjuntos de datos de entrenamiento y prueba (**Figura 5**). Predice a qué clase de polaridad pertenece el dato. La regresión logística se considera el clasificador de predicción más rápido, hace que el modelo sea menos difícil y puede permitirle una mejor generalización, es decir, evitar el sobreajuste, y funciona muy bien con los nuevos datos (Ismail et al., 2016).

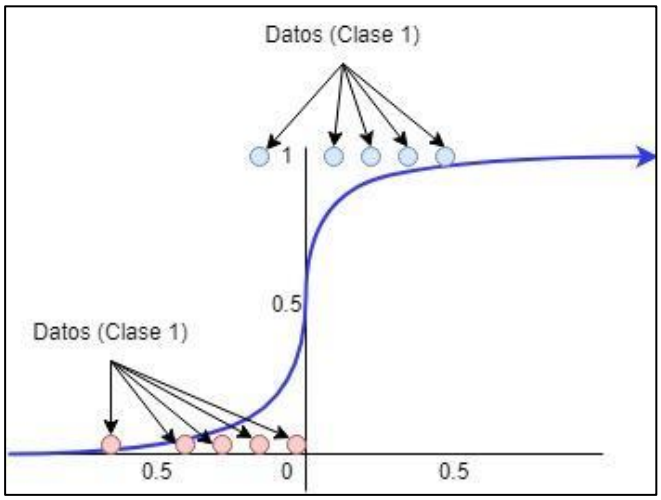


Figura 5: *Representación de la Regresión Logística.*
Fuente: Construcción Propia

2.9.5 Redes Neuronales Artificiales (RNA)

RNA es un modelo matemático que intenta simular la estructura y las funcionalidades de las redes neuronales biológicas. El componente básico de toda RNA es la neurona artificial; es decir, un modelo matemático simple (función). Este modelo tiene tres conjuntos simples de reglas: multiplicación, suma y activación. En el inicio de la neurona artificial las entradas se ponderan, lo que significa que cada valor de entrada se multiplica con un peso individual. En la sección media de la neurona artificial se encuentra la función que suma todas las entradas ponderadas y el sesgo. A la salida de la neurona artificial, la suma de las entradas ponderadas y el sesgo pasa por la función de activación, también llamada función de transferencia (**Figura 6**) (Krenker et al., 2011).

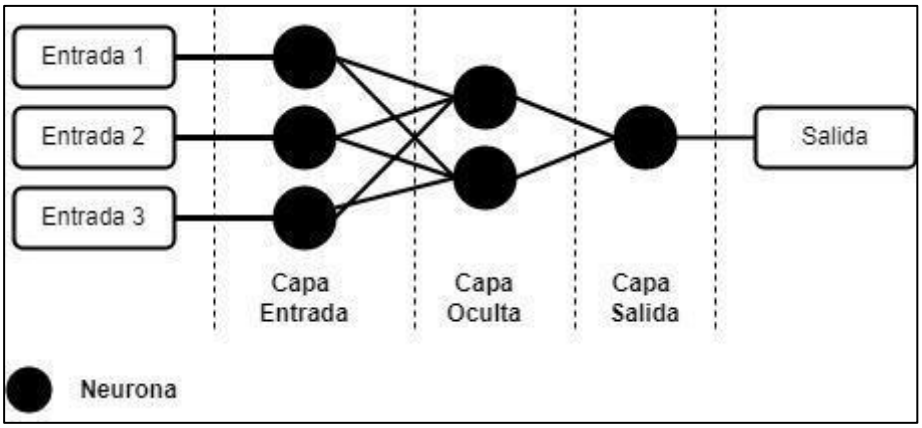


Figura 6: *Representación de las Redes neuronales artificiales.*
Fuente: Construcción Propia

2.9.6 Redes Neuronales Profundas (RNP)

RNP es una RNA con múltiples capas ocultas entre las capas de entrada y salida (**Figura 7**). Al igual que las RNA poco profundas, las RNP pueden modelar relaciones no lineales complejas. El objetivo principal de una red neuronal es recibir un conjunto de entradas, realizar cálculos progresivamente complejos en ellas y dar salida para resolver problemas del mundo real como la clasificación (K. Huang et al., 2019).

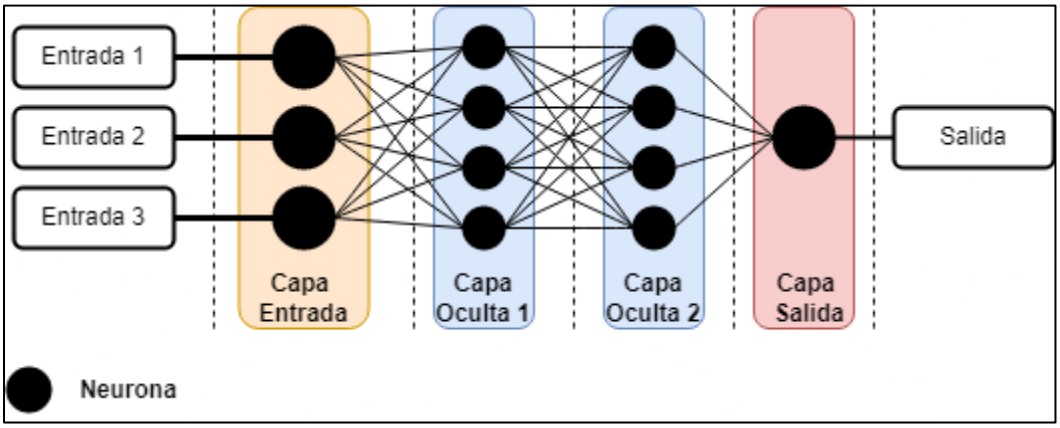


Figura 7: *Representación de las Redes Neuronales Profundas.*
Fuente: Construcción propia.

2.9.7 Exponential Smoothing

Exponential Smoothing es un algoritmo para hacer predicciones de series de tiempo, según Arroyo et al. (2007) este algoritmo obtiene pronósticos como el promedio móvil ponderado de todas las observaciones pasadas donde los pesos asignados disminuyen exponencialmente. En otras palabras, el Exponential Smoothing para datos de series de tiempo asigna ponderaciones exponencialmente decrecientes para las observaciones más recientes que a las más antiguas. Por lo tanto, cuanto más antiguos son los datos, menos prioridad ("peso") se les da a los datos, los datos más nuevos se consideran más relevantes y se les asigna más peso. Los parámetros de suavizado (constantes de suavizado), generalmente indicados por α, determinan los pesos de las observaciones. En Gardner (2006) presenta una descripción a detalle de este algoritmo.

En Arroyo et al. (2007) se menciona que este algoritmo generalmente se usa para hacer pronósticos a corto plazo, debido a que en los pronósticos a más largo plazo con esta técnica pueden ser bastante poco confiables. En la **Figura 8** se presenta una representación del este algoritmo.

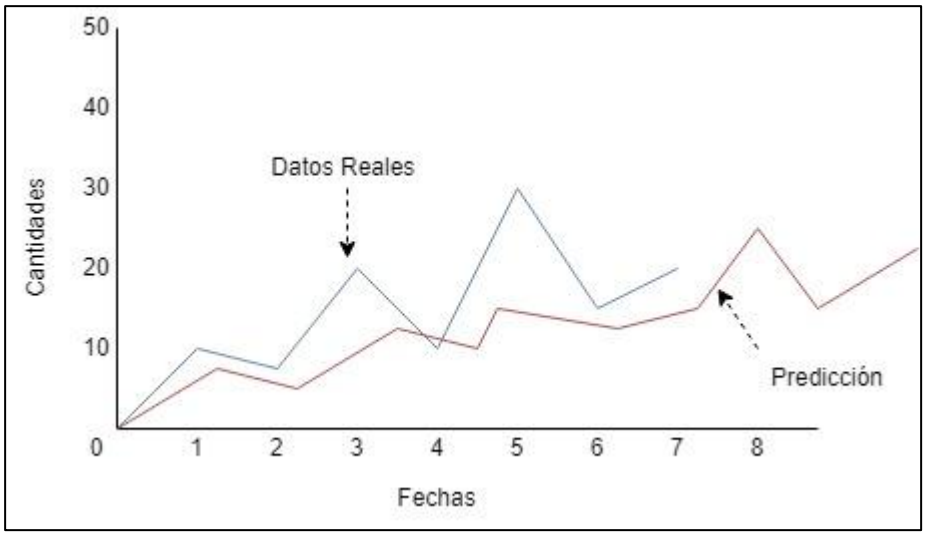


***Figura 8**: Representación la predicción de datos con Exponential Smoothing*
Fuente: Construcción propia.

2.10 Métricas para evaluar un modelo de Aprendizaje Automático

Realizar un modelo de aprendizaje automático es importante en una investigación, pero también es fundamental medir el rendimiento del modelo entrenado para conocer qué tan bueno o que tan malo es el modelo generado. Existen dos métodos para evaluar modelos de aprendizaje automático cuando se trata de un problema de clasificación, la evaluación cuantitativa y la evaluación cualitativa, la cualitativa se realiza mediante la experiencia de los usuarios en el sistema mientras que la cuantitativa permite tener una forma mecánica de cuantificar los resultados (Dalianis, 2018). El objetivo de la evaluación es estimar la precisión de la generalización de un modelo sobre los datos futuros (no vistos/fuera de muestra). En esta investigación se utilizará la evaluación cuantitativa en cada modelo utilizado.

Un concepto importante para la evaluación de modelos es la matriz de confusión **(Figura 9)**. La matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje automático, en esta matriz las filas representan la clase predicha, mientras que las columnas representan la clase real. Entonces, VP y VN denotan el número de instancias positivas y negativas que se clasifican correctamente. Mientras tanto, FP y FN denotan el número de instancias negativas y positivas mal clasificadas, respectivamente (M & M.N, 2015). A partir de la **Figura 9** se pueden generar varias métricas de uso común.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 9: *Matriz de Confusión.*
Fuente: Construcción propia.

De la matriz de confusión se pueden obtener muchas métricas, las principales y necesarias para esta investigación se describen a continuación.

2.10.1 Accuracy

La métrica más común para la evaluación del clasificador, evalúa la efectividad general del algoritmo al estimar la probabilidad del valor verdadero de la etiqueta de clase (Bekkar et al., 2013). La formulación para esta métrica se presenta en la ecuación (4).

$$\textbf{Accuracy} = \frac{\textbf{VP} + \textbf{VN}}{\textbf{VP} + \textbf{VN} + \textbf{FP} + \textbf{VN}} \quad \textbf{(4)}$$

2.10.2 Recall

El recall se utiliza para medir la fracción de patrones positivos que se clasificaron correctamente. Esta métrica suele no ser muy valorada en la recuperación de información (debido al supuesto de que hay muchos documentos relevantes que en realidad no importa qué subconjunto encontremos, no podemos saber nada sobre la relevancia de los documentos que nos son devueltos). El recall tiende a ser descuidado o promediado en aprendizaje automático y lingüística computacional (donde la atención se centra en la confianza que podemos tener en la regla o el clasificador). Sin embargo, en un contexto de lingüística computacional/traducción automática, se ha demostrado que recall tiene un peso importante (Powers & Ailab, 2020). La fórmula que se utiliza para esta métrica se presenta en la ecuación (5) la cual está basada en la matriz de confusión.

$$\textbf{Recall (R)} = \frac{\textbf{VN}}{\textbf{VN} + \textbf{FP}} \quad \textbf{(5)}$$

2.10.3 Precisión

La precisión se utiliza para medir los patrones positivos que se predicen correctamente a partir del total de patrones predichos en una clase positiva (M & M.N, 2015). La fórmula que se utiliza para esta métrica se describe en la ecuación (6) la cual también está basada en matriz de confusión.

$$\textbf{Precision } (\mathbf{P}) = \frac{\mathbf{VN}}{\mathbf{VN} + \mathbf{FN}} \quad \textbf{(6)}$$

2.10.4 F Score

El valor F Score se utiliza para combinar las medidas de precisión y recall en un solo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y el recall entre varias soluciones. El F-Score se define como el promedio ponderado de precisión y recall según la función de ponderación β (Dalianis, 2018). Una formula general para el F-score se presenta en la ecuación (7).

$$\mathbf{F-Score:}\ \ \mathbf{F_b} = (\mathbf{1} + \mathbf{b^2}) * \frac{\mathbf{P} * \mathbf{R}}{\mathbf{b^2} * \mathbf{P} + \mathbf{R}} \quad \textbf{(7)}$$

Con b = 1 se obtiene la ecuación (8) para el f-score.

$$\mathbf{F-score:}\ \mathbf{F_1} = \mathbf{2} * \frac{\mathbf{P} * \mathbf{R}}{\mathbf{P} + \mathbf{R}} \quad \textbf{(8)}$$

2.10.5 Kappa Score

El estadístico kappa de Cohen es una medida ampliamente utilizada que maneja correctamente los problemas de clases múltiples como las clases desbalanceadas. Kappa compara la precisión observada con la precisión esperada. Revela el punto de concordancia entre las clases verdaderas y clasificaciones. Para calcular el valor de kappa se utiliza la ecuación (9).

$$\boldsymbol{\kappa} = \frac{\boldsymbol{\rho_0} - \boldsymbol{\rho_e}}{\mathbf{1} - \boldsymbol{\rho_e}} = \mathbf{1} - \frac{\mathbf{1} - \boldsymbol{\rho_a}}{\mathbf{1} - \boldsymbol{\rho_e}} \quad \textbf{(9)}$$

donde ρ_0 es el acuerdo observado, y ρ_e es el acuerdo esperado.

La interpretación de valor de kappa según Agrawal & Trivedi (2020) se presenta en la **Tabla 2**.

Tabla 2: *Interpretación de valores de kappa*

Valor Kappa	Interpretación
$\boldsymbol{\kappa < 0}$	Concordancia baja
$\boldsymbol{0 < \kappa < 0.20}$	Concordancia leve o nada
$\boldsymbol{0.20 < \kappa < 0.40}$	Concordancia Regular
$\boldsymbol{0.40 < \kappa < 0.60}$	Concordancia Moderada
$\boldsymbol{0.60 < \kappa < 0.80}$	Concordancia sustancial
$\boldsymbol{0.80 < \kappa < 1}$	Concordancia casi acuerdo perfecto

2.10.6 ROC (AUC)

Gráficamente (**Figura 9**), a menudo se representa ROC (Receiver Operating Characteristic) como una curva que obtiene la tasa de verdaderos positivos en función de la tasa de falsos positivos para el mismo grupo; Específicamente, el enfoque ROC implica representar el valor de la sensibilidad en función de (1-especificidad) para todos los valores de umbral posibles y unir los puntos con una curva. Cuanto más inclinada esté la curva hacia la esquina superior izquierda, mejor será la capacidad del clasificador para discriminar entre clases positivas y negativas (Bekkar et al., 2013).

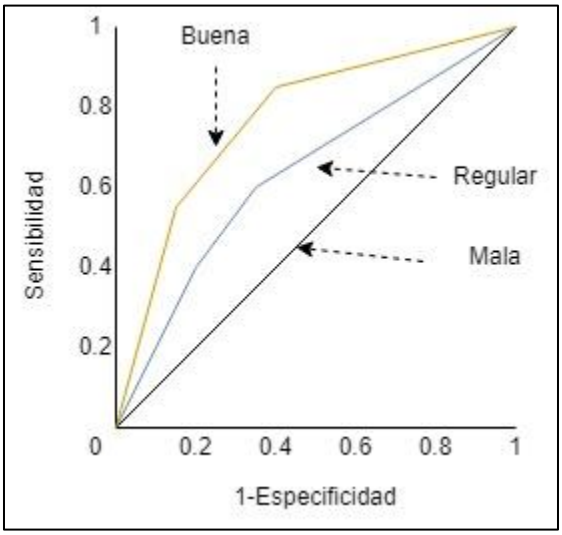


Figura 10: *Representación de la curva ROC.*
Fuente: Construcción propia

Provost & Fawcett (1997) fueron los líderes en el desarrollo del uso de la curva ROC, aconsejando su uso como una alternativa al accuracy en el caso de aprendizaje de datos desbalanceados o desequilibrados. Desde entonces, este enfoque se ha vuelto ampliamente utilizado, con talleres dedicados y con aplicación en varias investigaciones (Elazmeh et al., 2006; Hernàndez-Orallo et al., 2004; J. H. Xue & Titterington, 2008).

El área bajo la curva ROC (Area under curve, AUC) por su parte es un indicador del rendimiento de la curva ROC que puede resumir el rendimiento de un clasificador en una sola métrica. A diferencia de las dificultades encontradas en la comparación de diferentes curvas ROC, especialmente en el caso de intersección, el AUC puede clasificar los modelos por rendimiento general, como resultado, el AUC se considera más en la evaluación de modelos (Batista et al., 2004).

Algunos autores consideran que el AUC puede inducir a error en el rendimiento del modelo, especialmente en el caso de un aprendizaje de datos desbalanceado, ya que cubre una parte del rango de predicción sin utilidad en la práctica (Briggs & Zaretzki, 2008). Otras alternativas fueron propuestas en la literatura para lograr una evaluación más clara, una es el AUC ponderado, la cual

es una variante de AUC que se adapta mejor al caso de aprendizaje de datos desbalanceados (Weng & Poon, 2008).

En la práctica, el valor de AUC varía entre 0,5 y 1. Bekkar et al. (2013) sugiere la siguiente escala para la interpretación del valor AUC (**Tabla 3**).

Tabla 3: *Interpretación del valor AUC*

Valor AUC	Rendimiento del Modelo
0.5 - 0.6	Malo
0.6 - 0.7	Razonable
0.7 - 0.8	Bueno
0.8 - 0.9	Muy Bueno
0.9 - 1.0	Excelente

2.10.7 Prueba de McNemar

La prueba de McNemar se ha usado para determinar qué clasificador es superior a otro en los nuevos ejemplos de prueba (Dietterich, 1998). La prueba de McNemar se basa en una matriz de 2x2 similar a la presentada en la **Tabla 4**. La hipótesis nula establece que la misma proporción de población se clasificará correctamente por el clasificador 1 y el clasificador 2 (De Leeuw et al., 2006). El test utiliza una razón de población $\psi = f_{12} - f_{21}$, que se estima por la razón de muestra $\frac{f_{12}}{f_{21}}$, y además el test se basa en una estadística de chi-cuadrado, calculada con la ecuación (10).

$$\boldsymbol{\chi^2 = \frac{(f_{12} - f_{21})^2}{(f_{12} + f_{21})}} \qquad \textbf{(10)}$$

Si la hipótesis nula es correcta, entonces la probabilidad de que esta cantidad sea mayor que $\chi^2_{1,0.95} = 3.841459$ es menor que 0.05. Entonces, se puede rechazar la hipótesis nula a favor de la hipótesis de que los dos algoritmos tienen un rendimiento diferente cuando se entrenan en el conjunto de entrenamiento particular (De Leeuw et al., 2006).

Tabla 4*: Ejemplo de una tabla de contingencia del test de McNemar.*

	Clasificador 2 (clasificados correctamente)	Clasificador 2 (incorrectos)
Clasificador 1(clasificados correctamente)	f_{22}	f_{21}
Clasificador 1 (incorrectos)	f_{12}	f_{11}

2.10.8 Alfa de Cronbach

En la parte final de la evaluación es importante realizar una validación de la fiabilidad del cuestionario realizado a los evaluadores. Para ello se utiliza un método bastante reconocido en

investigaciones denominado Alfa de Cronbach. Según Rodríguez-Rodríguez & Reguant-Álvarez (2020) el coeficiente alfa de Cronbach es una fórmula general para estimar la fiabilidad de un instrumento en el que la respuesta a los ítems tiene más de dos valores como por ejemplo en una escala de actitudes con respuesta de tipo Likert.

Existen varias fórmulas para hacer el cálculo del coeficiente alfa de Cronbach, una de las más utilizadas se presenta en la Ecuación (11).

$$\alpha = \frac{\kappa}{k-1}\left[1 - \sum \frac{S_i^2}{S_t^2}\right] \quad (11)$$

Donde, k es el número de ítems del instrumento, S_i^2 la varianza de las puntuaciones en el ítem i, y S_t^2 la varianza de las puntuaciones totales del cuestionario o test.

El valor del coeficiente de alfa de Cronbach oscila entre 0 y 1, el valor mínimo aceptable para el coeficiente alfa de Cronbach es 0.70; por debajo de ese valor la consistencia interna de la escala utilizada es baja. Por su parte, el valor máximo esperado es 0.90; por encima de este valor se considera que hay redundancia o duplicación. Cuando el valor supera 0.90, varios ítems están midiendo exactamente el mismo elemento de un constructo; por lo tanto, los ítems redundantes deben eliminarse. Usualmente, se prefieren valores de alfa entre 0.80 y 0.90 (Streiner, 2003). Sin embargo, cuando no se cuenta con un mejor instrumento se pueden aceptar valores inferiores al de alfa de Cronbach, teniendo siempre presente esa limitación (Cortina, 1993).

2.10.9 Prueba de Wilcoxon

Es una prueba no paramétrica que se realiza en la evaluación con el objetivo de contrastar si dos muestras proceden de poblaciones equidistribuidas. Es decir, se compara el rango medio de dos muestras relacionadas y se determina si existen diferencias entre ellas. Si se tiene una muestra aleatoria $X_1, X_2 \dots X_n$, con valor observado $x_1, x_2 \dots x_n$, de una función de distribución continua y simétrica F_x con mediana M_x. Una hipótesis nula sobre el valor de la mediana se escribe como $H_0 : M_x = M_0$, donde M_0 es la mediana de la variable aleatoria X. Si r(.) es el rango de una observación, el estadístico de rango con signo de Wilcoxon se puede escribir simbólicamente como la Ecuación **(12)** (Taheri & Hesamian, 2013).

$$T^+ = \sum_{i=i}^{n} r(|d_i|) I(d_i > 0) \quad (12)$$

Donde $d_i = x_i - M_0$, y I es la función indicadora.

2.10.10 Prueba de Shapiro-Wilk

La prueba de Shapiro y Wilk (Shapiro & Wilk, 1965) es una de las pruebas de hipótesis más empleadas para analizar la normalidad a asimetría o curtosis, o ambas. Originalmente esta prueba se restringió para una un tamaño de muestra de menos de 50. Si se tiene una muestra aleatoria ordenada $y_1 < y_2 \ldots < y_n$ la prueba estadística original de Shapiro-Wilk se define como la ecuación (13. El valor de W se encuentra entre cero y uno. Valores pequeños de W conducen al rechazo de la normalidad mientras que un valor de uno indica normalidad de los datos (Mohd Razali & Bee Wah, 2011) .

$$W = \frac{(\sum_{i=i}^{n} a_i y_i)^2}{\sum_{i=n}^{n}(y_i - \bar{\bar{y}})^2} \qquad \textbf{(13)}$$

Donde y_i es la i^{th} es el estadístico de i-ésimo orden, $\bar{\bar{y}}$ es la media muestral, $a_i = (a_1 \ldots a_n) = \frac{M^T V^{-1}}{(M^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$ y $m = (m_1 \ldots m_n)^T$ son los valores esperados de los estadísticos de orden de variables aleatorias independientes e idénticamente distribuidas muestreadas a partir de la distribución normal estándar y V es la matriz de covarianza de esos estadísticos de orden.

2.11 Herramientas Tecnológicas

En esta subsección se presentan conceptos de algunas herramientas necesarias para el desarrollo de esta investigación.

2.11.1 Python

Python es un lenguaje de programación de alto nivel creado por Guido van Rossum a principios de los años 90. Es un lenguaje con una sintaxis muy limpia, tiene un código legible y permite desarrollar todo tipo de aplicaciones, también es un lenguaje de programación que permite trabajar fácilmente con inteligencia artificial, aprendizaje automático, big data y ciencia de datos, campos que están en auge (Raschka et al., 2020).

2.11.2 Jupyter Notebook

Jupyter Notebook es una herramienta de código abierto basada en navegador que funciona como un cuaderno de laboratorio virtual para admitir flujos de trabajo, código, datos y visualizaciones que detallan el proceso de investigación. Es legible por máquinas y por humanos, lo que facilita la interoperabilidad y la comunicación académica. Estos cuadernos pueden vivir en repositorios en línea y proporcionar conexiones a objetos de investigación como conjuntos de datos, código, documentos de métodos, flujos de trabajo y publicaciones que residen en otros lugares (Randles et al., 2017).

2.11.3 Pandas

Pandas es un paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para que el trabajo con datos "relacionales" o "etiquetados" sea fácil e intuitivo. Su objetivo es ser el bloque de construcción fundamental de alto nivel para realizar análisis de datos prácticos del mundo real en Python (McKinney et al., 2012).

2.11.4 Natural Language Toolkit

Natural Language Toolkit (NLTK), es un conjunto de módulos de programa de código abierto, tutoriales y conjuntos de problemas, que proporciona material didáctico de lingüística computacional listo para usar. NLTK cubre el procesamiento del lenguaje natural simbólico y estadístico, y está interconectado con corpus anotados (Loper & Bird, 2002).

2.11.5 Scikit-learn

Scikit-learn es una librería de Python que integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados (Kramer, 2016). Incluye otras funciones auxiliares que son parte del aprendizaje automático, como pasos de preprocesamiento de datos, técnicas de re-muestreo de datos, parámetros de evaluación e interfaces de búsqueda para ajustar/optimizar el rendimiento de un algoritmo (Pedregosa et al., 2011).

2.11.6 Statsmodel

Statsmodels es una librería de Python que proporciona a SciPy para cálculos estadísticos que incluyen estadísticas descriptivas y estimación de modelos estadísticos. Statsmodel tiene modelos como: regresión lineal, modelos lineales robustos, modelos lineales generalizados y modelos para datos discretos, pero en la última versión de scikits.statsmodels también se incluye algunas herramientas y modelos básicos para el análisis de series temporales (Perktold et al., 2011).

2.11.7 Facebook-Scrapper

Es una librería desarrollada con Python que permite hacer web Scraping de grupos públicos de Facebook, el código fuente con la descripción de cada una de las funcionalidades y las limitaciones está disponible en Github (Hellriegel, 2021).

2.11.8 SpaCy

SpaCy es una librería de Python que permite construir aplicaciones de NLP. SpaCy proporciona modelos pre entrenados con una gran cantidad de datos de diferentes idiomas, lo cual junto a una sintaxis clara hace que sea ideal para NLP. SpaCy permite realizar tokenización, POS, NER, entre otros (Explosion AI, 2017).

2.11.9 Dashboard

Dashboard es una herramienta cognitiva que mejora el "ámbito de control" sobre una gran cantidad de datos comerciales. Ayuda a las personas a identificar visualmente tendencias, patrones y anomalías, razonar sobre lo que ven y ayudar a guiarlos hacia decisiones efectivas. Como tal, estas herramientas deben aprovechar las capacidades visuales de las personas (Brath & Peters, 2004).

2.11.10 Power BI

Power BI permite a las organizaciones convertir sus datos en información útil que impulsa conocimientos comerciales más profundos e informa la toma de decisiones. La plataforma se conecta a los datos almacenados en una variedad de fuentes nativas de Microsoft y de terceros (por ejemplo, Microsoft Common Data Service for Applications (CDS), Excel, SQL, Google Analytics, MindChimp) y permite a las organizaciones visualizar y comprender fácilmente estos datos. a través de paneles e informes interactivos y personalizables (Microsoft, 2022).

CAPÍTULO 3: ESTADO DEL ARTE Y TRABAJOS RELACIONADOS

En esta sección se describen trabajos relacionados en el campo de la detección de competidores y la minería de opiniones comparativas, se presenta su importancia, los métodos más usados y las metodologías que han resultado de las investigaciones realizadas. Para llevar a cabo la búsqueda de trabajos relacionados, se realizaron búsquedas bibliográficas contra las bases de datos de literatura científica Scopus, ACM Digital Library, IEEE y Google Scholar. La búsqueda se realizó contra el título, resumen y palabras clave.

3.1 Identificación de Competidores

La identificación de competidores es uno de los objetivos principales de la CI, en la actualidad con el avance de la tecnología existen algunas técnicas modernas que se pueden aplicar, no obstante, en los primeros estudios de identificación de competidores, se utilizaron métodos tradicionales como las encuestas combinado con análisis de contenido. Por ejemplo, Chen (1996) y Peteraf & Bergen (2003) combinan los recursos y la información del mercado para identificar a los competidores existentes, mientras que descuidan las amenazas y las innovaciones de los competidores potenciales. Estos estudios utilizan métodos manuales para la recolección de información como lo son las encuestas y además solo se enfocan en un campo específico o temas seleccionados, por lo que los métodos propuestos y las conclusiones del estudio no son aplicables a otras industrias.

Otros estudios como los de He et al. (2013), Y. Xue et al. (2018) y S. Yadav & Shah (2019) identifican competidores utilizando técnicas de minería de textos y aprendizaje automático; sin embargo, no consideran el problema de identificar comparaciones en texto generado por el usuario, además son realizados en el lenguaje inglés. Por otra parte, los estudios de Gao et al. (2018) y Y. Liu et al. (2019) consideran en sus métodos minería de opiniones comparativas en el idioma Inglés y

Chino respectivamente para hacer análisis sobre su competencia. Todos los estudios mencionados (Un resumen se presenta en la **Tabla 5** y **Tabla 6**) carecen del análisis en el lenguaje español y hasta la fecha no existen investigaciones de detección de competidores en este idioma. Una limitación de estos estudios para implementarlo en el idioma español es que existen pocos datos, i.e. corpus que estén orientados a estos objetivos y sirvan para hacer un análisis de este tipo.

3.2 Minería de opiniones comparativa

Para lograr el objetivo de poder detectar competidores usando tecnologías web 2.0 que incorporan datos generados por el usuario, primero se debe identificar y clasificar cuáles son los enunciados que pertenecen a las categorías comparativa y no comparativa. La minería de opiniones por sí sola es insuficiente porque esto solo mostrará cuánto habla la gente y cómo se siente acerca de ciertos productos o servicios, lo que conduce a juicios incorrectos (Varathan et al., 2017).

La minería de opiniones comparativas ha sido abordada en estudios con técnicas como: i) Aprendizaje Automático en los estudios Jindal & Liu (2006), Q. Liu et al. (2013), Wang et al. (2015), Xu et al. (2011); ii) NLP en los estudios Q. Liu et al. (2013), Sun et al. (2009), Xu et al. (2011). Más detalles de los estudios previos de minería de opiniones comparativas se muestran en la **Tabla** 7 y

Tabla 8. La identificación de textos comparativos sigue siendo un problema abierto y no abordado en el lenguaje español. Además de este problema, los estudios actuales mencionados anteriormente y en la **Tabla 7** y **Tabla 8** presentan otras debilidades, donde no han podido realizar una evaluación en un caso de estudio real para determinar si los resultados tienen un valor agregado. Algunas de estas investigaciones hacen referencia a este problema debido a la falta de un corpus destinado para extraer opiniones comparativas. Jindal & Liu, 2006b y Li et al. (2011) indican que no existen corpus diseñados específicamente para determinar si un fragmento de texto hace una comparación. Cuando se trata de texto generado por el usuario comparativo, el problema se vuelve aún más desafiante, ya que el texto suele ser informal y corto (Y. Li et al., 2017).

Tabla 5: *Trabajos relacionados de Identificación de Competidores (Parte 1).*

Título	Descripción	Enfoque	Alcance
Competitor analysis and interfirm rivalry: Toward a theoretical integration (Chen 1996).	Propone un marco para el análisis de la competencia basado en la similitud de recursos y la similitud del mercado	Encuesta	Solo para competidores existentes No se aplica comúnmente en la actualidad. Enfoque manual y sin posibilidad de reusar.
Scanning dynamic competitive landscapes: a market-based and resource-based framework (Peteraf & Bergen 2003).	Proporciona un marco basado en el mercado y en los recursos para escanear campos competitivos complejos.	Encuesta	Solo para competidores existentes Enfoque manual y sin posibilidad de reusar. En la actualidad existen otras técnicas modernas como minería de textos para hacer una detección de competidores de manera automática.
Identifying comparative customer requirements from product online reviews for Competitor Analysis (Jin et al., 2016)	Se investiga cómo seleccionar un pequeño número de oraciones obstinadas de reseñas de productos en línea para el análisis de la competencia.	Minería de textos	Los resultados no se visualizan en una interfaz gráfica de usuario interactiva. Lenguaje Inglés No se evalúa con empresas reales. Una limitante de este estudio para el idioma español es encontrar comentarios con menciones de empresas.
Social media competitive analysis and text mining: A case study in the pizza industry (He et al., 2013).	Ayuda a las empresas a comprender cómo realizar un análisis competitivo de las redes sociales.	Minería de textos	Recolección manual de datos. Caso de estudio de grandes cadenas de pizzería, no se abordan las limitaciones de recursos, etc. que se podría tener en las pymes. Lenguaje Inglés. No se aborda el tema de restricción de acceso a datos en redes sociales de la actualidad.

Tabla 6: *Trabajos relacionados de Identificación de Competidores (Parte 2).*

Título	Descripción	Enfoque	Alcance
Mining Competitive Intelligence from social media: A case Study of IBM (Y. Xue et al., 2018)	Proponen un marco innovador para estudiar a los competidores, así como fortalezas y debilidades de la empresa a partir de datos de redes sociales utilizando técnicas de minería de textos.	Minería de textos	No evalúa el uso de la propuesta por parte de los tomadores de decisiones. Caso de estudio de IBM, no se abordan las limitaciones de recursos y cantidad de datos que se podría tener en las pymes Lenguaje Inglés.
Opinion Mining from Customer Reviews for Predicting Competitors (S. Yadav & Shah, 2019)	Evaluar la competitividad siempre utilizando las opiniones de los clientes en términos de reseñas, valoraciones y abundante fuente de información de la web y otras fuentes.	Minería de textos	Lenguaje Inglés Analiza productos en lugar de competidores. Reportes gráficos no interactivos. Para el idioma español una limitante es la carencia de datos para hacer este análisis.
Identifying competitors through comparative relation mining of online reviews in the restaurant industry (Gao et al., 2018)	Propone un modelo para extraer relaciones comparativas de revisiones en línea, y luego construir tres tipos de redes de relaciones de comparación, lo que permite el análisis de competitividad.	Minería de opiniones comparativas	Se enfoca en restaurantes Lenguaje Inglés Para el idioma español una limitante es la falta de datos para poder hacer estos análisis.
Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media (Y. Liu et al., 2019)	Propone un método novedoso para el análisis de la ventaja competitiva, que proporciona una base esencial para la gestión de la calidad y el desarrollo de la estrategia de marketing.	Minería de opiniones comparativas	Lenguaje Chino No se evalúa la generalidad del método. Una limitante en el idioma español es la ausencia de datos para este tipo de análisis.

Tabla 7: *Trabajos relacionados de Minería de Opiniones Comparativas (Parte 1)*

Título	Descripción	Enfoque	Puntos Fuertes	Debilidades
Exploiting machine learning for comparative sentences extraction Wang et al. (2015).	Construye un modelo para clasificar oraciones como comparativas o no comparativas. Su enfoque se aplicó a las reseñas comparativas de los clientes chinos.	Aprendizaje Automático con SVM entrenado con palabras clave y patrones de secuencia	Realiza una combinación de resultados entre SVM y patrones de secuencia obteniendo un F-score de 0.87	No se realiza un análisis a profundidad de los resultados, por ejemplo, determinando posibles competidores o tendencias de productos. El modelo no se ha implementado en un caso de estudio real.
Chinese comparative sentence identification based on the combination of rules and statistics Q. Liu et al. (2013).	Identifica oraciones comparativas chinas.	Aprendizaje Automático con SVM entrenado con palabras comparativas usando reglas secuenciales de clase	Se obtienen resultados que superan a los que la mayoría de investigaciones reportan hasta ese año.	No se pudo hacer un análisis más completo debido a que no existe un corpus destinado para oraciones comparativas. El modelo no se ha implementado en un caso de estudio real.
Product comparison using comparative relations S. Li et al. (2011).	Clasificación de la polaridad de oraciones comparativas.	Aprendizaje Automático con SVM para oraciones comparativas.	Presenta un método efectivo para comparar productos	No se pudo hacer un análisis más completo debido a que no existe un corpus destinado para oraciones comparativas. La información no se presenta de una manera entendible para los tomadores de decisiones. No se realiza una evaluación para determinar si los resultados tienen un valor agregado. El modelo no se ha implementado en un caso de estudio real.

Tabla 8: *Trabajos relacionados de Minería de Opiniones Comparativas (Parte 2)*

Título	Descripción	Enfoque	Puntos Fuertes	Debilidades
Identifying Comparative Sentences in Text Documents Jindal & Liu (2006a).	Estudia el problema de identificar oraciones comparativas en documentos de texto.	Aprendizaje automático con SVM y Naive bayes	Propone un método novedoso para identificar textos comparativos en base a palabras clave y POS. Analizan textos de diferente tipo: reviews, discusiones en foros, artículos.	No se pudo hacer un análisis más completo debido a que no existe un corpus destinado para oraciones comparativas. No se realiza un análisis a profundidad de los resultados, por ejemplo, determinando posibles competidores o tendencias de productos. El modelo no se ha implementado en un caso de estudio real.
Mining reviews for product comparison and recommendation Sun et al. (2009).	Propone un sistema automatizado basado en gramática de dependencia y árbol de evolución. Su sistema podría comparar y recomendar productos a los clientes desde perspectivas subjetivas y objetivas.	NLP con grafos gramaticales de dependencia	Presenta información de productos y sus características de una manera entendible. Realiza recomendaciones de productos en base a los resultados obtenidos.	No se realiza una evaluación para determinar si los resultados tienen un valor agregado. El modelo no se ha implementado en un caso de estudio real.
Mining comparative opinions from customer reviews for competitive intelligence Xu et al. (2011).	Propone un modelo gráfico novedoso para extraer y visualizar relaciones comparativas de reviews de Amazon.	NLP con grafos gramaticales de dependencia, caminos sintácticos y aprendizaje automático con SVM.	El rendimiento de la extracción de opiniones comparativas es prometedor.	No se realiza una evaluación para determinar si los resultados tienen un valor agregado. El modelo no se ha implementado en un caso de estudio real.

3.3 NER para detectar Posibles competidores.

NER sirve para realizar tareas cruciales en la gestión de la información, como la anotación semántica, la respuesta a preguntas, la población de ontologías, la minería de opiniones y la minería de opiniones comparativas (Marrero et al., 2013). En el campo de la Inteligencia Competitiva, NER es importante debido a que con este algoritmo se puede entrenar para detectar entidades como las organizaciones de texto generado por el usuario en plataformas digitales y posteriormente después de un análisis poder determinar si existen algunas empresas que pueden ser competidores de una empresa específica.

En este sentido, NER se puede utilizar para la detección de competidores en texto generado por el usuario. Por ejemplo, autores como Wu et al. (2012) presentaron un estudio para el reconocimiento preciso de nombre de productos a partir de contenido generado por el usuario. En dicha investigación se implementaron modelos como Standard Match model, Rule Templates model and Conditional Random Field model, el mejor resultado se obtuvo al realizar una combinación de los mismos, obteniendo una mejora significativa en su rendimiento con respecto a los modelos individuales, en un caso específico incluso mejora hasta un 11-12%. Estos resultados demuestran que dependiendo del contexto que se analiza en la detección de entidades para determinar productos en contenido generado por el usuario, la combinación de varios modelos puede ser una buena herramienta para tener los mejores resultados.

Otro estudio realizado por Feng et al. (2018) menciona los problemas de hacer NER mediante redes neuronales para el idioma español y el holandés que debido a la limitación de recursos y falta de datos anotados tienden a tener rendimientos en NER más bajos. Por lo tanto, se presenta una investigación del conocimiento entre idiomas para enriquecer las representaciones semánticas de los idiomas de bajos recursos. Los lenguajes de bajos recursos son aquellos que tienen relativamente menos datos disponibles para entrenar sistemas de Inteligencia Artificial conversacionales. Para ello primero desarrolla redes neuronales para para mejorar la representación de las palabras de bajos recursos a través de la transferencia de conocimientos desde un lenguaje de altos recursos utilizando léxicos bilingües, luego diseñan una estrategia de extensión del léxico para abordar el problema fuera del léxico aprendiendo automáticamente las proyecciones semánticas. Por último, consideran las características de distribución de tipo de entidad a nivel de palabra como un conocimiento independiente del lenguaje externo e incorporan la arquitectura neuronal. En esta investigación se realizan análisis en dos idiomas de bajos recursos (holandés y español) y demuestran la efectividad de las representaciones semánticas que se han realizado (mejora promedio del 4.8%). También, en el conjunto de datos chino OntoNotes 4.0 (Weischedel et al., 2010), muestra un enfoque que alcanza una puntuación F del 83,07% con una ganancia absoluta del 2,91% en comparación con los sistemas de última generación.

Así también, otro estudio realizado por Loster et al. (2017) realizó el reconocimiento de nombres de empresas a partir de texto no estructurado mediante el uso de diccionarios. Este es un estudio que presenta una forma alternativa para hacer NER que ayuden a una empresa. En esta investigación, se utilizó un sistema de aprendizaje automático con Campos Aleatorios Condicionales (CRF) para reconocer organizaciones de manera confiable de texto en alemán. Construyeron y emplearon varios diccionarios, expresiones regulares, contexto de texto y otras técnicas para

mejorar los resultados. Para la evaluación se centran en analizar el impacto de utilizar un "diccionario perfecto" que es un diccionario creado por los investigadores que contiene todas las empresas anotadas manualmente del conjunto de datos de prueba y entrenamiento, y diferentes diccionarios de empresas del mundo real, así como los efectos de diferentes formas de integrar el conocimiento contenido en los diccionarios sobre el rendimiento del sistema NER. Por último, contribuyen con los siguientes aportes: i) creación de un sistema NER capaz de reconocer con éxito empresas en textos alemanes con una precisión del 91,11% y una recuperación del 78,82%, y ii) análisis del impacto de varias estrategias de funciones basadas en diccionarios sobre el rendimiento de NER.

Los estudios presentados por Loster et al. (2017) y Wu et al. (2012) realizan sus investigaciones e implementan sus algoritmos en textos del idioma inglés. Un estudio adicional presentado por Molina et al. (2015) realiza una investigación con datos del idioma español con el objetivo de hacer un prototipo para reconocimiento de entidades. En este estudio se utiliza CRF para hacer NER en el idioma español y dos corpus anotados, ANCORA (Taulé, M., M.A. Martí, 2008) y CoNLL2002 (Tjong Kim Sang & de Meulder, 2003) que definen diferentes categorías de entidades nombradas. Dado que el CoNLL2002 tiene un conjunto de prueba, se realizaron las pruebas sobre este conjunto y se obtuvieron resultados de precisión entre 79 % y 86%, Recall entre 64% y 79% y F Score entre 73% y 80%, dependiendo de las características utilizadas en la evaluación. Por otro lado, ANCORA no dispone de un conjunto de prueba especificado, por esto para medir el rendimiento del sistema se utilizó el método de cross-validation y se obtuvieron resultados de precisión entre 77% y 78%, Recall entre 47% y 49% y F-Score, en promedio 59%. En este estudio al igual que el presentado por Feng et al. (2018) es únicamente para NER más no para detectar nombres de empresas y/o producto de un texto. Todos los estudios revisados (Resumen en **Tabla 9**) se centran en hacer análisis de algoritmos para NER, en algunos casos como resultado obtienen que utilizar una combinación de modelos puede mejorar las métricas finales, sin embargo, ningún estudio se especializa en hacer NER para CI. La utilización de estos algoritmos analizados puede ser muy interesantes para los objetivos planteados en esta investigación.

Tabla 9: *Trabajos relacionados de Reconocimiento de Entidades Nombradas.*

Referencia	Titulo	Descripción	Enfoque
Wu et al. (2012)	Reconocimiento preciso de nombres de productos a partir de contenido generado por el usuario.	Solución del equipo ganador en el concurso de ICDM 20121, el concurso trata se trata de reconocer automáticamente las menciones de los productos de un corpus que le dan en el concurso.	Standard Match model, Rule Templates model and Conditional Random Field model
Feng et al. (2018)	Mejora del reconocimiento de entidades con nombre de bajos recursos mediante la transferencia de conocimientos entre idiomas.	Investigación del conocimiento interlingüística para enriquecer las representaciones semánticas de idiomas de bajos recursos (español y holandés) a través de la transferencia de conocimientos desde un lenguaje de altos recursos utilizando léxicos bilingües, luego diseñan una estrategia de extensión del léxico para abordar el problema fuera del léxico aprendiendo automáticamente las proyecciones semánticas. Por último, consideran las características de distribución de tipo de entidad a nivel de palabra como un conocimiento independiente del lenguaje externo e incorporan la arquitectura neuronal.	NER con redes neuronales.
Loster et al. (2017)	Mejora del reconocimiento de la empresa a partir de texto no estructurado mediante el uso de diccionarios	Presentan un sistema de aprendizaje automático con CRF para reconocer organizaciones de manera confiable de texto en alemán. Construyen y emplean varios diccionarios, expresiones regulares, contexto de texto y otras técnicas para mejorar los resultados.	NER con CRFs utilizando diccionarios.
(Molina et al., 2015)	Prototipo para el reconocimiento de entidades nombradas en el idioma español	Presentan un sistema de aprendizaje automático con CRF para reconocer organizaciones de manera confiable de texto en español.	NER con CRFs utilizando dos conjuntos de datos ANCORA y CoNLL2002.

CAPÍTULO 4: DISEÑO E IMPLEMENTACIÓN

El presente capítulo muestra la metodología para la creación del corpus, así como para su evaluación (**Figura 11**). En primer lugar, se seleccionaron los corpus disponibles en el idioma español con contenido generado por el usuario orientados hacia plataformas de redes sociales. En segundo lugar, se realizó un modelo para la detección de entidades y un filtrado por características comparativas para generar un ranking de los corpus que son de utilidad para la detección de competidores. Con los corpus rankeados se realizó un análisis de detección de texto comparativo donde con el modelo entrenado se etiquetó a todos los textos como comparativos y no comparativos. Posteriormente, con los corpus etiquetados se generó el corpus final, el cual tiene textos de todos los corpus existentes y con las nuevas etiquetas que son las entidades encontradas y si el texto es comparativo o no. Por último, se realiza una extracción de datos de las principales plataformas de redes sociales del país para realizar la evaluación de los algoritmos realizados, para esto se toma en consideración tres fuerzas de Michael Porter, muy reconocido investigador sobre la competitividad de empresas, donde se evalúa la utilidad del corpus, por lo tanto, se realiza un Dashboard para mostrar todos los resultados encontrados en el caso de estudio, el sector textil.

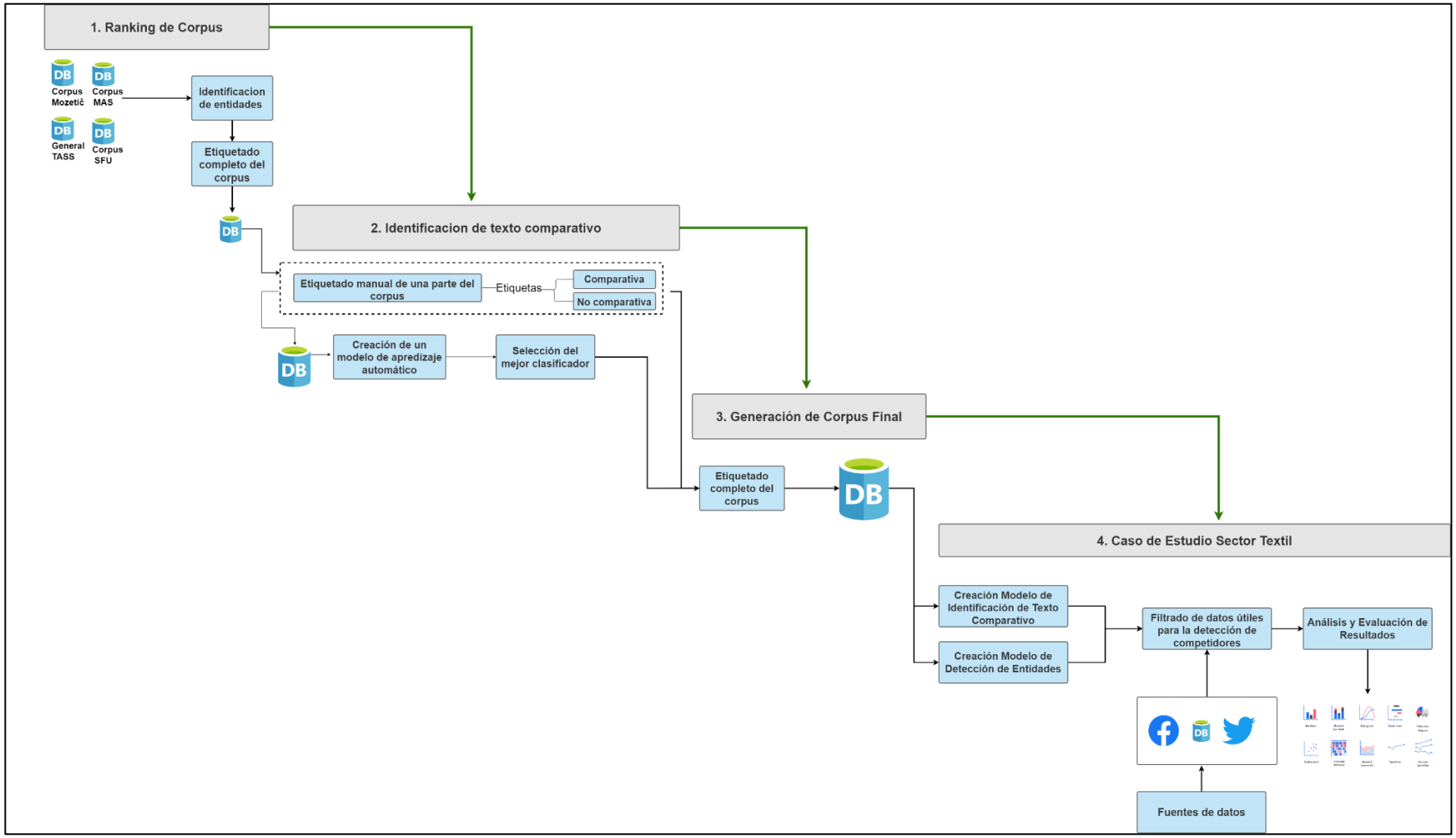


Figura 11: *Metodología para la creación y evaluación del corpus.*
Fuente: Construcción Propia

4.1 Ranking de Corpus

El proceso de CI, comprende de 5 pasos, en el paso de extracción de datos y análisis es donde mayormente existen problemas (Xu et al., 2011). Con respecto a la extracción de datos, los problemas están en la integración de varias fuentes, muchos estudios como en Gao et al., 2018, He et al. (2016) y Y. Xue et al. (2018) realizan un análisis de inteligencia competitiva pero se basan en una fuente de datos de una plataforma única porque generalmente sus estudios se basan en empresas internacionales como Apple, Microsoft, etc., los cuales tienen una cantidad de datos muy grande para recolectar, pero eso no sucede con empresas pequeñas como las PYMEs. En las PYMEs existe una cantidad muy pequeña de datos que son insuficientes para entrenar algoritmos de aprendizaje automático y crear un análisis robusto. Otro problema en la extracción de datos es conseguir información suficiente en el idioma español, como menciona Cedeno-Moreno & Vargas (2020) existe una escasez de contenido en español para hacer un análisis en comparación con corpus con contenido en inglés.

Por lo tanto, en esta subsección se creó una metodología para priorizar los corpus existentes (**Figura 12**) que sean de utilidad para el objetivo de esta investigación. Como primer punto se hizo una búsqueda de corpus existentes en el idioma español que sean de años actuales y contengan información de plataformas digitales, especialmente plataformas como las redes sociales, después se desarrolló un modelo para realizar NER.

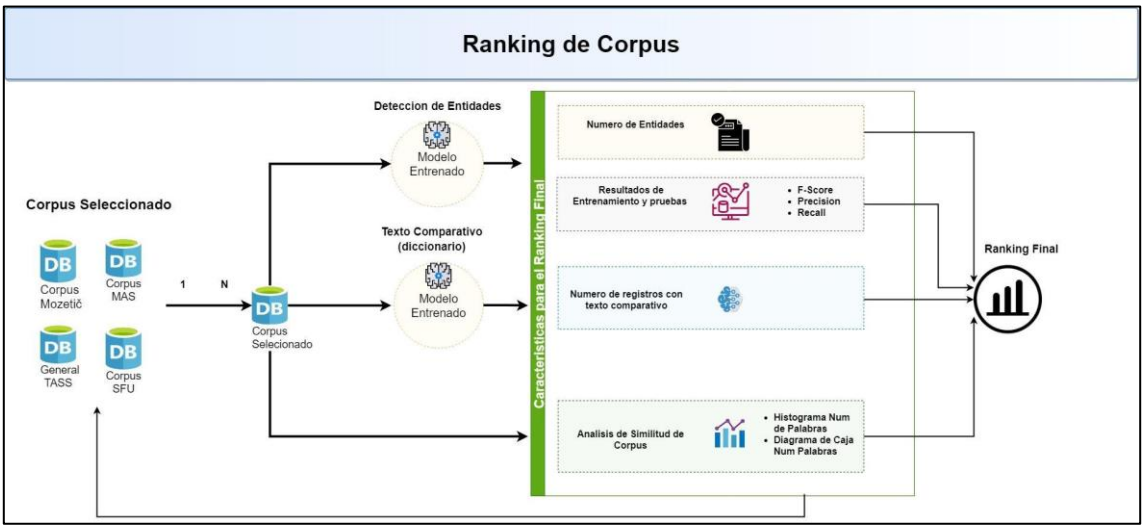


Figura 12: *Proceso general para ranking de corpus existentes.*
Fuente: Construcción Propia

Para el NER se utiliza las entidades más estudiadas que son: nombres propios de personas, lugares y de organizaciones. Para el entendimiento y prueba de este modelo se etiquetaron las entidades existentes de forma manual tomando solo un subconjunto de datos de cada corpus. Posteriormente se entrenó el modelo y se realizaron las respectivas pruebas con las métricas de evaluación de modelos recall, f-score y precisión, las cuales ayudan a determinar si los modelos tienen buenos resultados de entrenamiento y prueba para NER. También se realiza un algoritmo que permite

determinar si un comentario/tweet tiene texto comparativo o no, para el cual se utiliza un diccionario con todas las palabras que pueden determinar un texto comparativo. Estos resultados tanto del NER y la detección de texto comparativo permiten obtener un primer ranking de los corpus. Luego se realiza un análisis de similitud entre los corpus tomando en consideración el número de palabras debido a que existen corpus que tienen el texto del comentario/tweet con muchas palabras y otros con muy pocas palabras, este proceso se realiza para descartar corpus con muchos datos atípicos con respecto al número de palabras en cada registro. Por último, se unen los datos de los corpus seleccionados y se crea un único corpus general para la detección de competidores. A continuación, se brinda una explicación más detallada del proceso.

4.1.1 Corpus para detección de competidores

Para el ranking del corpus como primer paso se investigó y encontró varios corpus que según la descripción de los mismos han sido construidos generalmente con la intención realizar minería de opiniones, la mayoría de los datos de cada corpus fueron extraídos de redes sociales como Twitter, pero también hay comentarios de páginas de comercio electrónico de otros países. Después de realizar una búsqueda de corpus que estén exclusivamente en el idioma español y hacer un análisis exploratorio donde se revisa la descripción y que sean datos de redes sociales o plataformas de comercio electrónico, se realiza una indagación previa de los datos para comprobar que tengan etiquetas como el sentimiento. Se han seleccionado 4 corpus como se muestra en la **Tabla 10**, sin embargo, al momento de realizar este trabajo no se encontró un corpus que esté orientado para la detección de competidores. Todos los corpus presentados en la **Tabla 10** tienen licencia Creative Commons (CC). Después de esta selección se realiza un ranking tomando en consideración algunas características que se describirán en la sección siguiente y por último se crea un corpus general y se procede a hacer un modelo para detectar competidores.

Tabla 10: *Corpus con contenido en el idioma español*

Nombre	Tamaño	Etiquetado	Descripción
General Corpus/ General TASS (TASS Team, 2012)	14529 tweets	Sí (Sentimientos)	Este conjunto de datos es un corpus de texto principalmente con tweets etiquetados para tareas relacionados con el análisis de sentimientos (minería de opiniones).
Mozetič, Grčar & Smailovič corpus (Mozetic et al., 2016)	275588 tweets	Sí (Sentimientos)	Este conjunto de datos contiene 1.6 millones de tweets etiquetados con sentimientos por anotadores humanos de 15 lenguas europeas, para español existen 275588 tweets. Los datos se pueden utilizar para entrenar y evaluar clasificadores de opiniones de Twitter, para calcular el acuerdo de anotador o para estudiar las diferencias entre el uso del lenguaje en Twitter.
SFU Spanish review corpus (Taboada, 2017)	400 reviews	Sí (Sentimientos)	Este conjunto de datos contiene 400 reseñas sobre diferentes temas: automóviles, libros, hoteles, teléfonos celulares, música, computadores y películas. Los datos son extraídos del sitio web Ciao.es. Cada categoría contiene 25 reseñas positivas y 25 negativas, definidas como positivas o negativas en función del número de estrellas otorgadas por el revisor (1-2 = negativo; 4-5 = positivo; las reseñas de 3 estrellas no están incluidas).
MASS Corpus (Corpus for Marketing Analysis in Spanish) (Navas-Loro et al., 2018)	3765 tweets	Sí (Sentimientos)	MASS Corpus contiene un conjunto de tweets etiquetados manualmente en español para fines de marketing. Para cada publicación de Twitter, se proporcionan etiquetas para describir tres aspectos diferentes del texto: los sentimientos; si hace una mención a un elemento de la mezcla de marketing; y, la posición del autor del tweet con respecto al embudo de compra. Cada etiqueta está relacionada con una sola marca, que también se especifica para cada tweet.

4.1.2 Etiquetado Manual para NER

Como se puede observar en la descripción de cada corpus (**Tabla 10**), estos están construidos por sus autores para objetivos diferentes a la detección de competidores como, por ejemplo, la determinación de sentimientos de contenido generado por el usuario. Por lo tanto, algunos de estos corpus tienen ya etiquetas para entrenar modelos relacionados, aunque también hay corpus que tienen solo el texto y no tienen ninguna etiqueta.

El objetivo de este primer paso para el ranking de los corpus es encontrar las organizaciones y/o empresas en los corpus. Una técnica para este propósito es un algoritmo de aprendizaje automático denominada NER. Para aplicar el algoritmo y poder evaluar los resultados se necesita tener suficientes datos etiquetados. Para crear un etiquetado de datos para NER se puede abordar algunas opciones, un método bastante usado es el etiquetado BILOU debido a que considera entidades que contienen más de una palabra.

El etiquetado BILOU se realiza anotando el inicio de la entidad (Beginning), dentro de la entidad (Inside), fin de la entidad (Last) y fuera de la entidad (Outside). También cuando la entidad se compone de un solo elemento (Unique) (Ver **Figura 13**) (Molina et al., 2015).

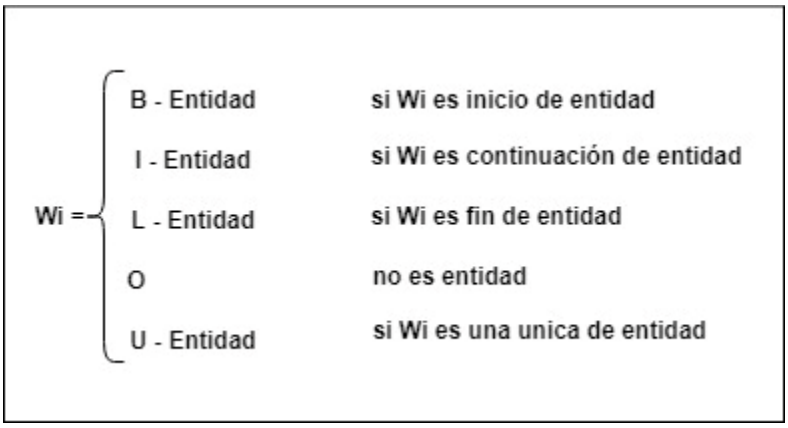


***Figura 13**: Tipos para etiquetado BILOU*
Fuente: (Molina et al., 2015).

El efecto del uso de algún estilo de anotación en particular se basa en dos aspectos. Primero, en el número de clases que se maneja, ya que, al distinguir entre el inicio de una entidad habrá que hacerlo para todas las entidades que contenga el corpus, de la misma forma con los demás prefijos (B, I, O, L, U). Segundo, en las métricas de evaluación, ya que para considerar que un clasificador ha etiquetado correctamente se toma en cuenta que toda la entidad esté correctamente reconocida, en caso de que alguno de sus elementos no sea correcto, tal anotación es considerada como incorrecta; impactando directamente en la Precisión, Recall y F-score.

***Tabla 11:** Ejemplo de etiquetado manual de un corpus*

Texto	Etiquetado BELOU
FC Barcelona 3 vs 0 Athletic Club BBVA 4 2 2017 via YouTube	['B-ORG', 'L-ORG', 'O', 'O', 'O', 'B-ORG', 'L-ORG', 'U-ORG', 'O', 'O', 'O', 'O', 'U-ORG']
El Banco Santander es reconocido como mejor banco privado en Portugal y Chile	['O', 'B-ORG', 'L-ORG', 'O', 'O', 'O', 'O', 'B-MISC', 'L-MISC', 'O', 'U-LOC', 'O', 'U-LOC']
WKAQ580 diganle a Luis Pabon Roca que los Alfa Romeo vienen de Italia pero son parte del grupo Fiat Chrysler	['O', 'O', 'O', 'B-PER', 'I-PER', 'L-PER', 'O', 'O', 'B-ORG', 'L-ORG', 'O', 'O', 'U-LOC', 'O', 'O', 'O', 'O', 'O', 'U-ORG', 'U-ORG']

Entonces, un ejemplo del etiquetado de un corpus se presenta en la **Tabla 11**, en cada palabra se etiqueta con cualquier prefijo de BILOU y para se considera varias entidades las cuales son: Persona (PER), Organización (ORG), Locación (LOC), Cualquier otra entidad (MISC).

Como se puede evidenciar en la descripción de cada corpus en la **Tabla 10**, cada uno tiene una gran cantidad de registros y etiquetar todos los datos conllevan un gran trabajo, motivo por el cual se ha realizado el etiquetado de aproximadamente mil registros de cada corpus. En la **Figura 14** se muestra una representación de cómo se subdivide cada conjunto de datos para realizar el etiquetado de los mismos y luego proceder a implementar la detección de entidades. La parte 1 del corpus se etiqueta manualmente, de estos datos etiquetados se dividen en tres partes, datos para entrenar (80%), datos de prueba (10%) y datos de validación (10%), con estos datos se generan las métricas de evaluación que fueron consideradas como una de las características para el ranking. Luego con el modelo entrenado, se procede a etiquetar los datos de la parte 2 y así obtener las entidades detectadas, las cuales también sirvieron como una característica adicional para el ranking general.

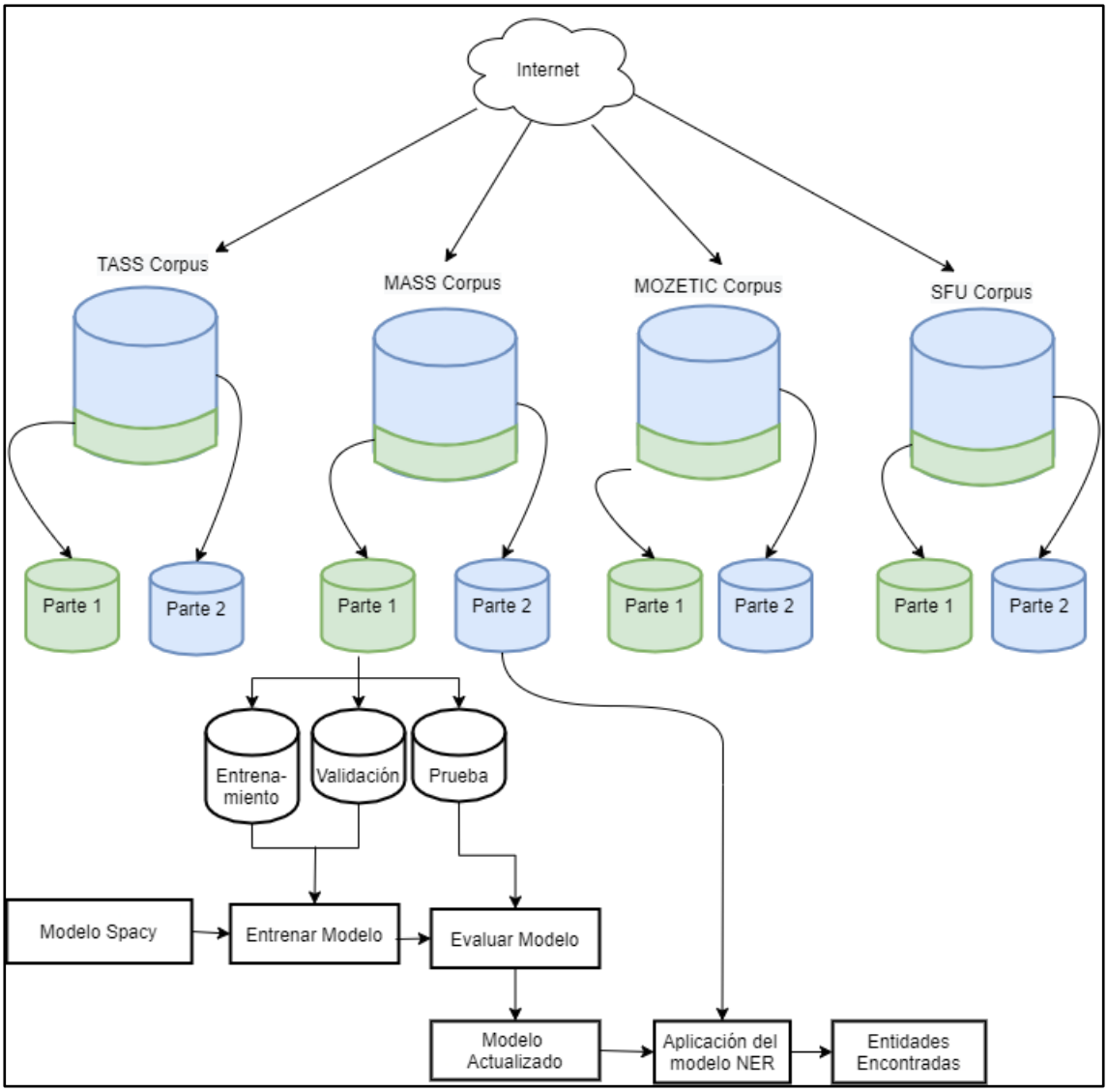


Figura 14: *Metodología para la detección de entidades de cada corpus.*
Fuente: Construcción propia.

4.1.3 Modelo para Identificar Entidades

En la actualidad para implementar un algoritmo NER se puede realizar desde cero, pero también se puede utilizar herramientas especializadas en estos algoritmos, una ventaja de utilizar estas últimas es que tienen modelos ya pre-entrenados con una gran cantidad de datos, lo cual ayuda a mejorar los resultados en la detección de entidades. De acuerdo a autores como Herreros (2021), SpaCy (Explosion AI, 2017) es una de las herramientas que mejor rendimiento tiene en cuanto al NER, motivo por el cual se eligió esta herramienta. Spacy dispone de varios modelos pre-entrenados, los cuales muestran diferentes valores de recall, precisión y f-score de entrenamiento y prueba con sus datos. En esta investigación se utiliza el modelo denominado es_core_news_lg que tiene un 90% de recall, precision y f-score. Este modelo pre-entrenado no se debe utilizar como modelo final, debido

a que es muy general, por lo tanto, se debe hacer una especialización agregando más datos etiquetados al modelo. Debido a que los corpus no tienen etiquetas para este ámbito, se realiza un etiquetado manual de todas las entidades (personas, lugares y organizaciones) encontradas en cada comentario/tweet. Luego se reentrena el modelo de SpaCy con los nuevos datos y se ajusta parámetros propios de la herramienta hasta obtener resultados con métricas aceptables de recall, precisión y f-score para una investigación (métricas mayores al 70%). Para obtener las métricas se divide el corpus etiquetado en datos de entrenamiento y datos de prueba. El proceso que se sigue para la detección de entidades se presenta en la **Figura 15**.

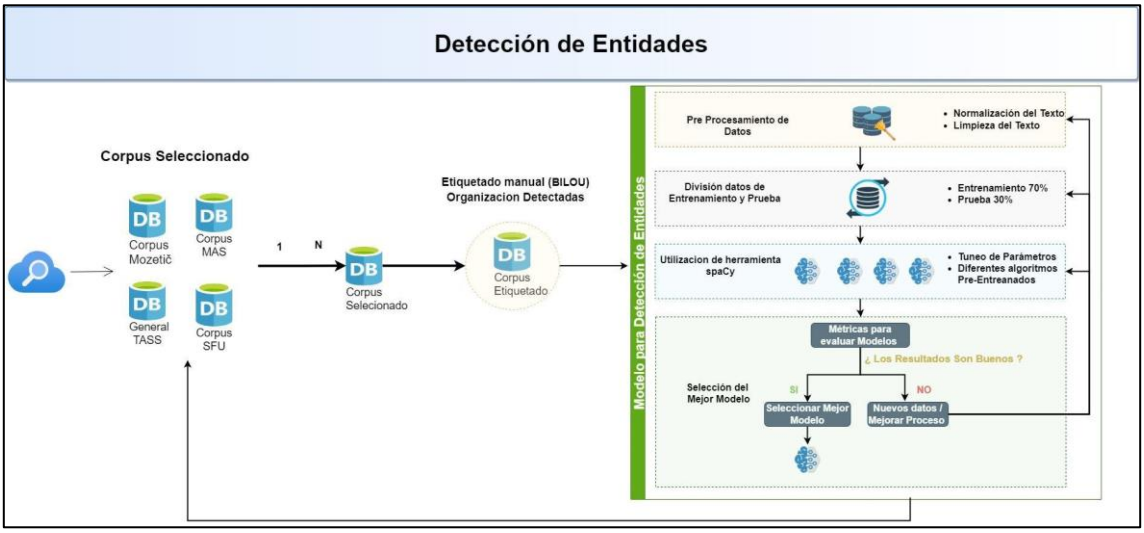


Figura 15: *Proceso general para la detección de entidades con corpus existentes.*
Fuente: Construcción Propia

4.2 Identificación de Características Comparativas

Para realizar la identificación de características comparativas se crea un diccionario con palabras o términos claves que indican si en un texto puede estar describiendo algún objeto y haciendo una comparación con otros objetos, luego se implementa una búsqueda del diccionario en cada registro y se etiqueta al corpus con SÍ, sí tiene texto comparativo, caso contrario con NO. Este proceso permite tener la cantidad de registros de un corpus que tienen características comparativas. Esto es un indicador importante para el ranking del corpus.

Existe un gran cantidad de palabras que pueden determinar sí en un oración se está realizando una comparación, en esta investigación se han utilizado el siguiente conjunto de palabras y sinónimos de las palabras propuestas por Gao et al. (2018), Jindal & Liu (2006a), Varathan et al. (2017), Wang et al. (2015) en el idioma inglés: "mejor", "igual", "supera", "superar", "vence", "vencer", "peor", "sobrepasa", "superior", "preferible", "distinto", "destacar", "destacado", "sobresaliente", "deficiente", "inferior", "pesimo", "identico", "equivalente", "semejante", "similar", "más que", "menos que", "mas largo", "comparado", "comparamos", "comparando", "contrastado", "comparar", "diferenciado", "diferenciar" "contrapuesto", "mas util", "barato", "caro", "mas

pequeño", "mas pequeña", "mas grande", "mas rapido", "menos grande", "menos rapido", "igual", "iguales", "menos util", "mayor que", "menor que".

4.3 Ranking de Corpus en español para CI

Como se indicó en el capítulo 1, la presente investigación tiene varios objetivos específicos y uno de ellos es crear un corpus para la detección de competidores. Esta creación tiene como base el estudio e integración de varios corpus existentes, para ello se realiza un análisis cada uno de ellos y se determina mediante un ranking cuales están más orientados hacia la detección de competidores. Para realizar el ranking de los corpus seleccionados se consideran las siguientes características:

I. NER en cada corpus. - Permite encontrar organizaciones en los datos y con qué frecuencia se mencionan en cada corpus, con ello se puede observar sí el corpus analizado está orientado hacia la inteligencia competitiva debido a que en los tweets y/o comentarios que tienen los corpus posiblemente el usuario está hablando de empresas y/o productos. Dado que se entrena un modelo en cada corpus, se encontró un score de NER que posteriormente sirvió para realizar el ranking general.
II. Número de entidades detectadas. - Permite observar con qué frecuencia y que tan variado las empresas son mencionadas en cada corpus. Esta característica se obtiene con el modelo NER previamente entrenado. Con el fin de obtener el valor cuantitativo para el ranking final, se otorga al corpus con mayor número de entidades detectadas un valor de 10 puntos, el número de entidades de los siguientes corpus se asigna un número proporcional.
III. Número de registros que tienen texto comparativo. - Permite determinar sí un corpus está orientado hacia la inteligencia competitiva debido a que, al tener registros con textos comparativos, existe una posibilidad de se esté comparando productos, artículos o incluso empresas, que son la parte clave de esta investigación.
IV. Utilidad y similitud de los corpus. - Se crea un diagrama de caja considerando el número de palabras que tiene cada registro de cada corpus, si la mediana del número de palabras de un corpus es muy superior en comparación con los otros corpus, entonces dicho corpus será penalizado. A los corpus que son semejantes en la mediana del número de palabras se le asigna un valor de 10, mientras que a los corpus que tienen una mediana superior se le asigna un valor proporcional, sí es muy superior incluso puede llegar a ser cero.

Como último paso para generar el ranking, con un grupo de expertos se procede a dar pesos a cada característica que se ha descrito, luego se realiza una operación matemática donde se suma el producto del valor calculado para cada característica con el peso asignado obteniendo un resultado sobre 100, este es el score final.

4.4 Identificación de Texto Comparativo

En esta subsección se creó un modelo para identificar texto comparativo. Como primer punto se generó el corpus en base al ranking con el cual se creó el modelo, después se etiquetó manualmente

un porcentaje de los textos como comparativos y no comparativos para realizar el análisis. Con estos datos etiquetados se realizó un análisis exploratorio, para tareas de limpieza y normalización; después se dividió los datos en entrenamiento y prueba. Posteriormente se crearon varios modelos de aprendizaje automático, donde se aplicaron métricas de evaluación para seleccionar el mejor modelo. Hay que tener en cuenta que este es un proceso cíclico, donde en cada iteración se buscó mejorar los resultados, como se indica en la **Figura 16**. A continuación, se brinda una explicación más detallada del proceso.

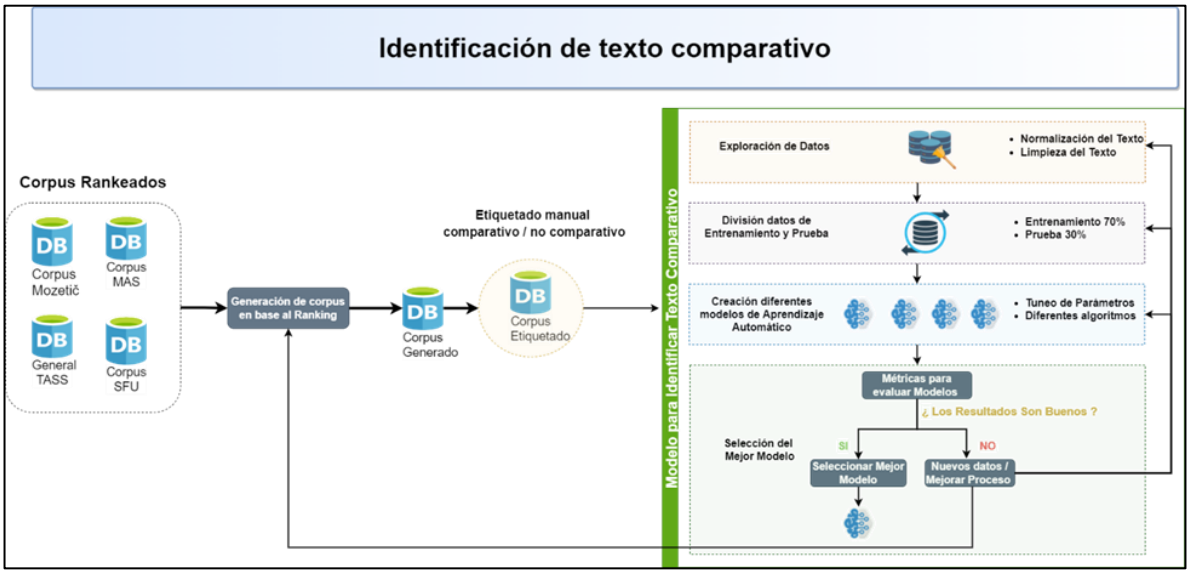


***Figura 16**: Proceso general para la identificación de texto comparativo.*
Fuente: Construcción Propia

4.5 Generación del corpus y etiquetado manual de los datos

Para la generación del corpus, se priorizaron los dos corpus mejores rankeados, de los cuales se etiquetaron 700 datos para realizar un primer análisis y evaluar la necesidad de etiquetar más datos para obtener mejores resultados en la creación del modelo. En el proceso de etiquetado (**Figura 17**) se siguió una metodología basada en la de Kessler & Kuhn (2014). Cada texto se lo etiquetó por los dos autores de este trabajo; en el caso de que las etiquetas coincidan, se agrega el texto al corpus. Por otra parte, los textos donde las etiquetas no coinciden se realizó un segundo análisis y discusiones en conjunto para determinar si es comparativo o no y se los agregó al corpus.

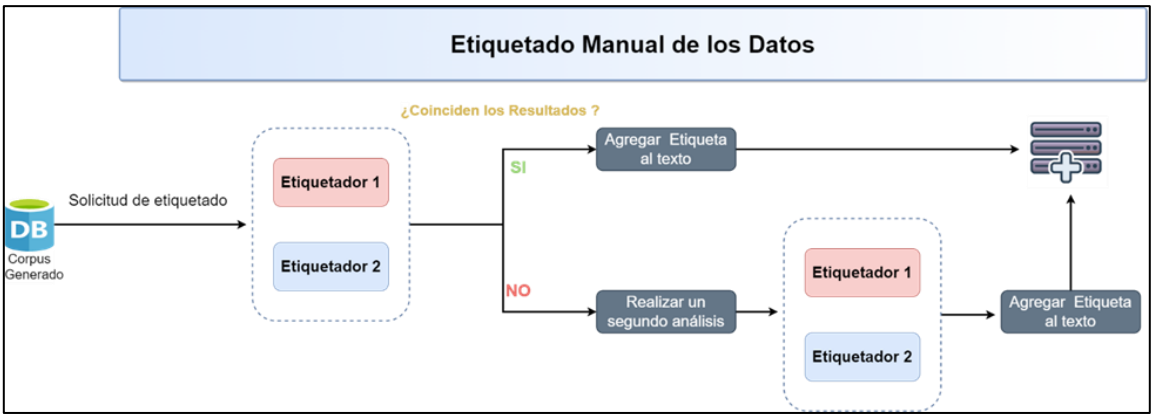


Figura 17: *Proceso que se siguió para realizar el etiquetado manual de los datos.*
Fuente: Construcción Propia

Según Kessler & Kuhn (2014) y Wang et al. (2015), en el mejor de los casos únicamente uno de cada diez textos creados por usuarios en internet es comparativo; respaldado por Khan et al. (2016), donde el 5% del de los textos fueron comparativos, y por Y. Li et al. (2017), donde únicamente el 2.1% fueron comparativos. Teniendo en cuenta esto, en el caso de que no se presenten buenos resultados en el contexto de las métricas de evaluación de modelos de aprendizaje automático presentados en la sección 2, se creó otra etapa, donde se utilizan todos los corpus que se definieron son de utilidad en el ranking para tratar de extraer más textos comparativos utilizando un filtrado para mejorar la probabilidad de encontrar estos textos **(Figura 18**). Este filtrado es importante debido a que etiquetar los textos manualmente es una labor que, si bien no presenta dificultad, demanda mucho tiempo (Khan et al., 2016).

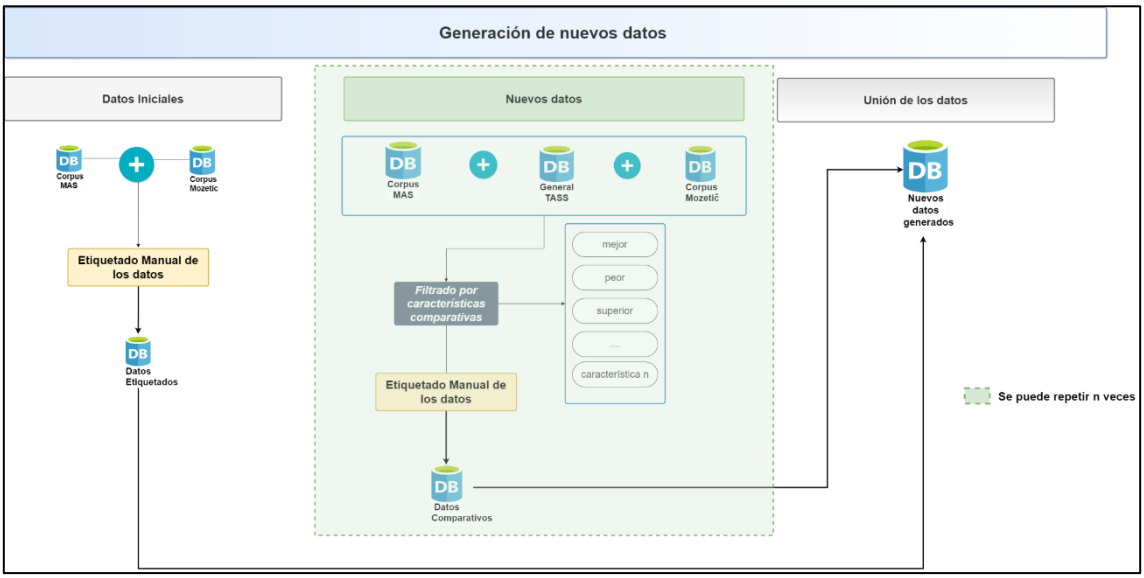


Figura 18*: Proceso para la generación de datos.*
Fuente: Construcción Propia

El corpus generado en esta fase sirve para la creación del modelo para identificar texto comparativo y tiene la estructura detallada en la **Tabla 12**.

Tabla 12: *Estructura del corpus para la creación del modelo.*

Atributos del corpus generado	Descripción	Tipo de Atributo
id	Identificador único para poder distinguir a qué texto y de que corpus original se está haciendo referencia.	String
texto	Variable que contiene el texto del corpus.	String
es_comparativo	Valor que indica si el texto es comparativo o no 0 = No es comparativo 1 = Es comparativo	Binario

4.6 Modelo para Identificar Texto Comparativo

Se presenta a detalle el proceso que se siguió para la creación del modelo con el fin de identificar texto comparativo. Como se puede observar en la **Figura 16**, este proceso puede realizarse en varios ciclos y regresar de una fase a otra, dependiendo de la aceptabilidad de los resultados.

4.6.1 Preprocesamiento de datos

La tarea de minería de opiniones comparativa requiere una fase de preprocesamiento que incluye herramientas de NPL como: stemming/lemmatization, POS, entre otras. De acuerdo a Kotsiantis et al. (2006) esta etapa puede tener un gran impacto dentro de modelos de aprendizaje supervisado, mejorando la capacidad de generalización. La **Tabla 13** muestra texto con y sin preprocesamiento. Aunque esas herramientas son muy importantes, en muchos casos no se mencionan explícitamente en las investigaciones (Varathan et al., 2017).

Tabla 13: *Ejemplo de textos antes del preprocesamiento y después del preprocesamiento.*

Antes del preprocesamiento	Después del preprocesamiento
@egolaxista Está en carrera faltan por llegar Astana, Sky y Movistar, BMC acaba de meter el mejor tiempo.	egolaxistar este carrera faltar llegar astana sky movistar bmc acabar meter mejor tiempo
@MariaOrtegon Qué buen video María. Siempre he escrito, desde mi primer libro, que Hichtcock es el artista más grande del mundo moderno.	mariaortegon que buen video maria siempre escrito primero libro que hichtcock artista mas grande mundo moderno
Recomiendo a @fumarellion. Tienda de venta de cigarrillos electrónicos al mejor precio. Visita su web PymesUnidas.	recomiendo fumarellion tienda venta cigarrillo electronico mejor precio visita web pymesunidas
#Economia Toyota y Suzuki inician negociaciones para forjar una asociación https://t.co/yi46EgdrsO	economia toyota suzuki iniciar negociación forjar asociacion

Una vivienda para refugiados creada por Ikea, el mejor proyecto de arquitectura de 2016 https://t.co/331kraoksU vía @A3Noticias	uno vivienda refugiado creado ikea mejor proyecto arquitecturar 2016 via a3noticias

A continuación, se presentan los procesos de limpieza y normalización que se aplicaron a los textos mostrados en la **Tabla 13** y a cada uno de los textos.

- **Eliminación de textos nulos y duplicados**. Se eliminó textos que no existan o que estén duplicados.
- **Manejo de tildes**. Se reemplazaron las vocales con tildes por vocales en su forma normal.
- **Eliminación de Uniform Resource Locators (URLs)**. A través de expresiones regulares se eliminan las URLs que se encuentren dentro de los textos.
- **Eliminación de StopWords**. Se eliminaron las palabras en español denominadas como StopWords. Aquí se agregaron excepciones a los StopWords: i) mas, ii) menos, y iii) igual; las cuales son importantes dentro de la minería de opiniones comparativas.
- **Stemming y Lemmmatization**. Se aplica stemming y lemmatization para aumentar la relevancia de los textos y obtener mejores resultados. En el proceso se usó stemming, o lemmatization, o incluso las dos juntas para comparar resultados.

4.6.2 Generación de datos de entrenamiento y prueba

Se usó el método de división aleatoria para dividir el conjunto de datos en entrenamiento y prueba. En la división aleatoria, el conjunto de datos se divide en entrenamiento y prueba en una proporción predefinida (Younis et al., 2020). El método de división aleatoria es más sólido en comparación con otros métodos, ya que el conjunto de datos se divide con mayor precisión (Reitermanov´a, 2010). En aprendizaje automático generalmente se utiliza una distribución de 70 – 90 % para entrenamiento y 10 -30% para prueba. En este caso al tratarse de un corpus con pocos textos comparativos en comparación de los no comparativos, se usó una distribución 70% para entrenamiento y 30% para prueba con el fin de facilitar la evaluación de los modelos.

4.6.3 Creación de los modelos (Entrenamiento)

Posteriormente se realizó la creación de modelos de aprendizaje supervisado usando los datos de entrenamiento generados. Estos modelos tienen el objetivo de clasificar un texto como comparativo o no comparativo. En esta sección, primero se crea una representación numérica de los textos y después se crean y entrenan los modelos usando los siguientes algoritmos: i) Naive Bayes, ii) SVM, iii) Random Forest, iv) Regresión Logística, y v) RNA.

a. Representación numérica de los textos

Los algoritmos de aprendizaje automático funcionan con números, por lo que es necesario encontrar una representación adecuada de los textos en números. Existen algunas técnicas que se han desarrollado para cumplir esta tarea. En este trabajo se utilizaron las siguientes técnicas:

- **Vector de características**: Transforma un texto en una matriz de recuentos de tokens (Effrosynidis et al., 2018)

- **Frecuencia de término (tf):** Cuenta una cantidad de ocurrencias de un término en relación al número de textos (Effrosynidis et al., 2018).

- **Frecuencia de término – Frecuencia de documento inversa(tf-idf):** El cálculo de tf-idf muestra la importancia de una palabra en un documento(texto) o conjunto de datos determinado (Effrosynidis et al., 2018)

b. Creación y entrenamiento de los diferentes algoritmos

Cuando se crean modelos de aprendizaje automático, la herramienta (en este caso Python con la librería scikit-learn) permite crear modelos con valores por defecto, pero para obtener modelos robustos y que se puedan generalizar, esto no es suficiente, por lo que es necesario realizar un proceso de tuneo de hiperparámetros. Los hiperparámetros son diferentes para cada algoritmo, varía la cantidad dependiendo de la complejidad del mismo y se los debe establecer antes de ejecutarse. A continuación, la **Tabla 14** muestra los hiperparámetros que se tunearon y los valores que se asignaron para cada algoritmo

Tabla 14: *Tuneado de hiperparámetros aplicado a los diferentes algoritmos.*

Algoritmo	Hiperparámetro	Descripción	Valores
Naive Bayes	alpha	Parámetro de suavizado aditivo (0 para ningún suavizado)	[1,0,10]
	fit_prior	Si aprender las probabilidades previas de la clase o no. Si es falso, se utilizará un anterior uniforme.	[True, False]
Regresión Logística	C	Inverso de la fuerza de regularización; debe ser un número positivo. Al igual que en SVM, los valores más pequeños especifican una regularización más fuerte.	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
	penalty	Especifica la norma de la sanción.	['l1','l2']
Random	n_estimators	El número de árboles	[10, 50, 100,150,200]

Forest	criterion	La función para medir la calidad de una división. Los criterios admitidos son "gini" para la impureza de Gini y "entropía" para la ganancia de información.	["gini", "entropy"]
	max_depth	La profundidad máxima del árbol.	[10,25, 50,75, 100]
	min_samples_split	El número mínimo de muestras requeridas para dividir un nodo interno.	[2, 3, 5,7,9,10]
	max_leaf_nodes	Se definen los mejores nodos como una reducción relativa de la impureza.	[10, 50, 100,200]
SVM	C	Parámetro de regularización. La fuerza de la regularización es inversamente proporcional a C. Debe ser estrictamente positiva. La penalización es una penalización de l2 al cuadrado.	[0.1,1,0.01,0.001, 10, 100,1000]
	kernel	Especifica el tipo de kernel que se utilizará en el algoritmo. Si no se proporciona ninguno, se utilizará 'rbf'.	['rbf','sigmoid','linear' ,'poly']
	gamma	Coeficiente kernel para 'rbf', 'poly' y 'sigmoid'.	[1,0.1,0.01,0.001,10,1 00]
Red Neuronal	hidden_layer_sizes	El i-ésimo elemento representa el número de neuronas en la i-ésima capa oculta.	[(10,10,10), (5,10,5), (10,)]
	activation	Función de activación de la capa oculta.	['tanh', 'relu','logistic']
	solver	El solucionador para la optimización del peso.	['sgd', 'adam']
	alpha	Parámetro de penalización L2 (plazo de regularización).	[0.0001, 0.05]
	learning_rate	Horario de tasa de aprendizaje para actualizaciones de peso.	['constant','adaptive']

Una vez se obtuvieron los mejores hiper parámetros para cada algoritmo, se procedió a entrenar los modelos usando los datos de entrenamiento generados en el punto anterior.

4.6.4 Selección del mejor modelo (Evaluación)

Con los modelos entrenados, se procedió a realizar su evaluación utilizando las métricas comúnmente utilizadas para evaluar modelos de aprendizaje supervisado cuando se trata de un problema de clasificación. Estas métricas incluyen: i) accuracy, ii) precisión, iii) recall, iv) f1 score, y v) ROC AUC ponderado. Todas estas métricas tienen un rango de 0 a 1, donde mientras más cerca el valor esté de 1, el modelo clasifica mejor los datos de prueba.

El f1 score, precisión y recall, se obtienen como un promedio o como un valor específico para cada clase (etiqueta). Existen dos tipos de promedios: i) micro y ii) macro. La elección de una métrica depende de cómo clasifica la importancia de las clases. Por ejemplo:

Si se tiene un conjunto de datos con una distribución de clases del 90 % al 10 %, un clasificador de referencia puede lograr una precisión del 90 % al asignar la etiqueta de clase mayoritaria. Si este es el objetivo, se elegiría un micro promedio como métrica, sin embargo, si se valora más la clase minoritaria, se deberían usar métricas macro promediadas (Manning et al., 2008), donde sólo se obtendría una puntuación del 50 %. Esta métrica es insensible al desequilibrio de las clases y las trata a todas por igual. En este trabajo se utilizaron métricas macro promediadas y de la clase comparativa, ya que se tiene un conjunto de datos desbalanceado.

Otro factor importante dentro de la evaluación es que se otorgó mayor importancia al f1 score y al ROC AUC ponderado al momento de decidir qué modelo es mejor ya que como se menciona en la sección **2.10**, estas se adaptan mejor al caso de aprendizaje de datos desbalanceados. Finalmente, con los dos mejores modelos se realizó un test de McNemar para determinar si existe una diferencia estadísticamente significativa entre los mejores modelos. Si el test demuestra que los modelos no son diferentes estadísticamente, entonces se selecciona como mejor modelo al que fue creado con un algoritmo menos complejo, por otra parte, si el test demuestra que los modelos son diferentes estadísticamente, se selecciona el modelo que tenga mejores resultados con las métricas utilizadas.

Todo este proceso como ya se ha mencionado anteriormente es cíclico y se puede regresar a cualquiera de sus fases según sea conveniente o según la investigación lo demande. El resultado final de esta sección es un modelo que puede etiquetar un texto como comparativo y no comparativo.

4.7 Generación del Corpus Final

Con los modelos creados para identificar entidades de cada corpus (sección **4.1.3**) se realiza la unión de datos (**Figura 19**) y con el modelo creado para identificar texto comparativo (sección 4.4) se procedió a generar el corpus final (**Figura 20**). Primero se tomaron los corpus que se encontraban sin clasificar y se realizó un filtrado por características comparativas para tener más posibilidades de tener un corpus balanceado. Después se utilizaron los modelos creados anteriormente para

etiquetar a los corpus como comparativos y no comparativos, además de indicar las entidades encontradas en cada texto. Por último, estos textos etiquetados se los unió con los textos que se etiquetaron en las secciones **4.1.2** para obtener el corpus final. El corpus final que se aporta en esta investigación contiene los atributos mostrados en la **Tabla 15**.

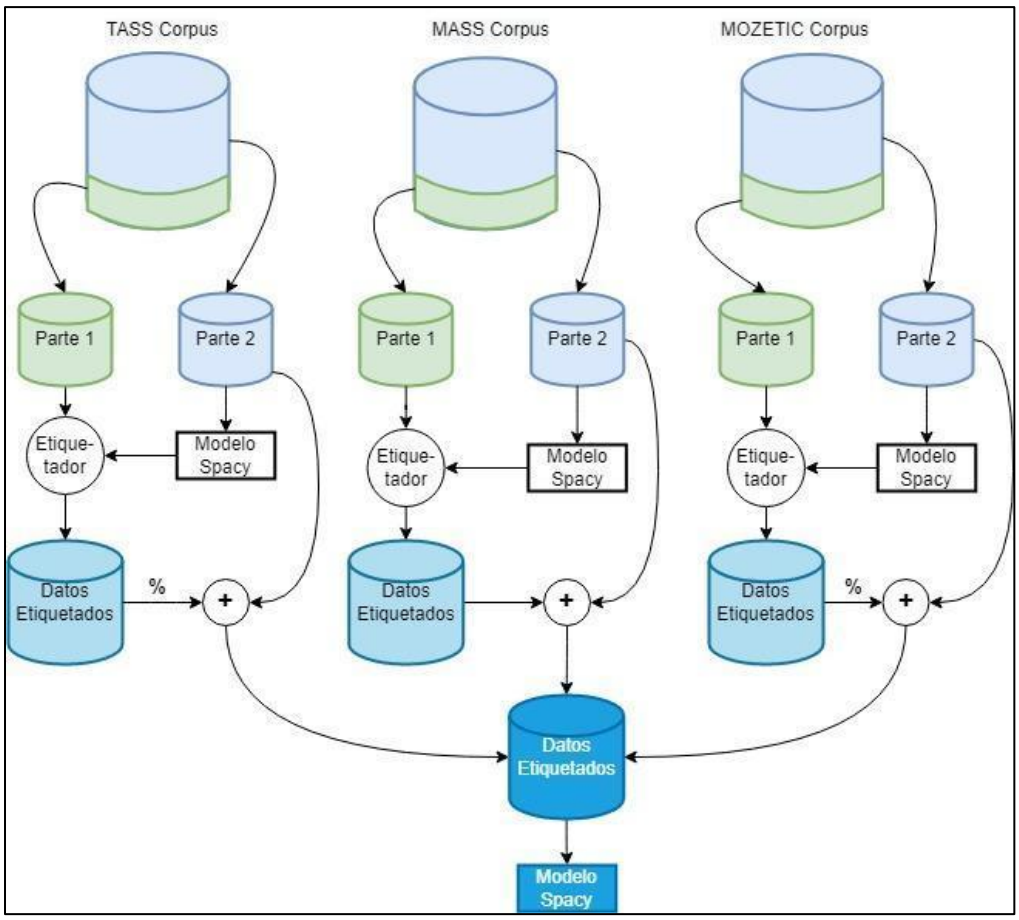


***Figura 19**: Metodología para Modelo Final de Detección de Entidades.*
Fuente: Construcción propia

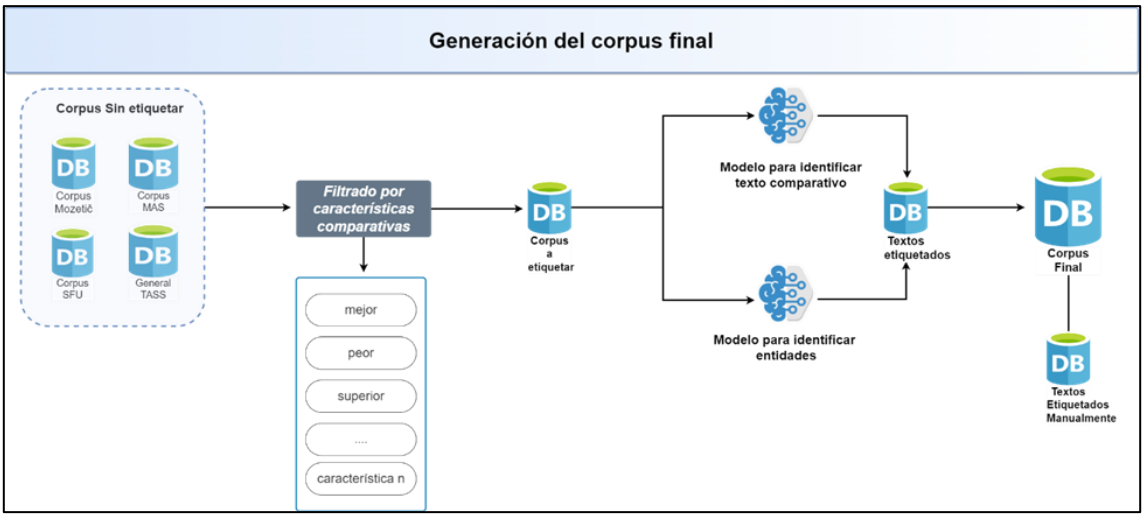


Figura 20: *Proceso para la generación del corpus final.*
Fuente: Construcción Propia

Tabla 15: *Atributos del corpus final generado*

Atributo	Descripción
Id	Identificador único de cada elemento del corpus. (Numérico)
corpusName	Valor que identifica a qué corpus original pertenece el texto (Categórico)
texto	Variable que contiene el texto del corpus (String)
entidades	Entidades que se han reconocido en el texto
es_comparativo	Valor que indica si el texto es comparativo o no (Binario)
sentimiento	Valor que indica sí el texto tiene sentimiento positivo, negativo o neutral

Por último, es necesario mencionar que con este corpus que tiene el etiquetado manual de entidades (como se presenta en la **Tabla 11**) todavía no es posible implementar un modelo NER con SpaCy. Para entrenar un modelo NER con esta librería es necesario etiquetas tipo JSON como se presenta en la **Figura 21**, por lo tanto, se desarrolló un script en Python para convertir los datos etiquetados manualmente en datos tipo JSON para un modelo NER de SpaCy. Este script se utilizó para desarrollar los modelos NER finales utilizados para en el Dashboard.

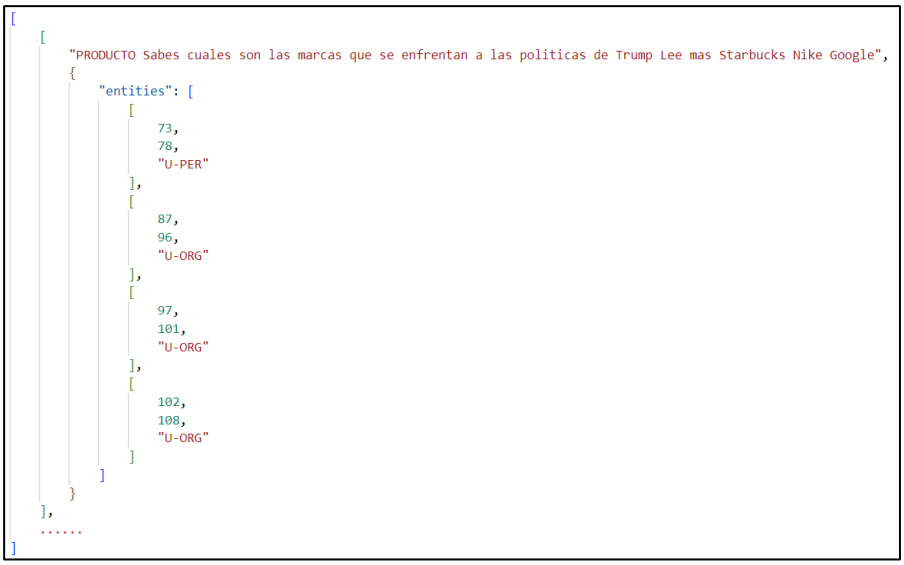


Figura 21: *JSON para un entrenar un modelo NER con SpaCy.*
Fuente: Construcción propia

4.8 Evaluación

Una parte importante de la investigación es la comprobación de la utilidad de los modelos creados en el sector de análisis. Para este proceso se busca datos generados por posibles clientes y que sean propios del sector emplear los modelos creados para posteriormente observar el comportamiento de los mismos y generar nuevas conclusiones que sean de interés para el sector (**Figura 22**).

En la actualidad, los datos se encuentran en: páginas web propias de la empresa, páginas de comercio electrónico, pero también una parte importante de dichos datos se encuentran en páginas de redes sociales (Tripathi, 2015). En la investigación presentada en CEDIA TIC.EC 2021 (Fajardo Cárdenas et al., 2021) se muestra un ranking de redes sociales para estudios de mercado, en esta investigación se presentan las tres plataformas mejor rankeadas para el sector textil, Facebook es la mejor rankeada ya que es una red social que contiene datos que pueden ser de mucha importancia para las empresas, pero debido a sus limitaciones en su API se debe buscar nuevas alternativas de extracción de datos. Luego está Twitter y YouTube como las siguientes plataformas mejor rankeadas en vista de que también tienen datos de mucha importancia para el análisis de mercado y además su extracción es menos compleja debido a que sus APIs no están muy restringidas aún.

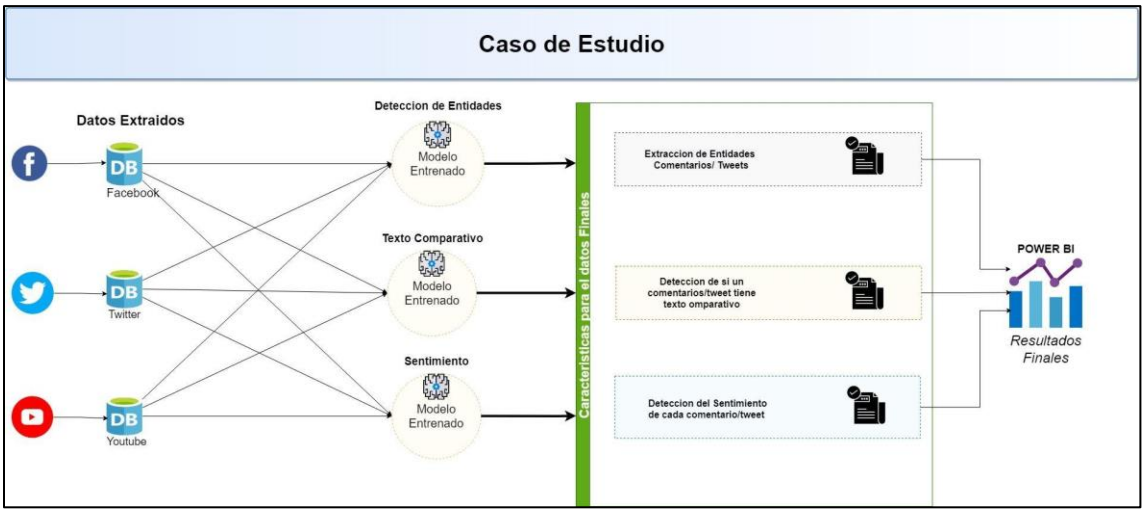


Figura 22: *Implementación de modelo con datos del caso de estudio*
Fuente: Construcción propia

4.8.1 Extracción de datos para Evaluación

En esta sección se procede con la extracción de datos de las diferentes plataformas digitales seleccionadas para utilizar los algoritmos desarrollados y hacer una evaluación de la utilidad del corpus generado. Para este proceso de extracción de datos existen dos alternativas, Web Scraping y mediante las propias APIs de cada plataforma. debido a la limitante de su API para extraer los datos de Facebook se utilizó como alternativa el Web Scraping. Se utilizó una librería llamada facebook-scrapper desarrollada con el lenguaje de programación Python (Hellriegel, 2021), mientras que para Twitter y YouTube se utilizó sus respectivas APIs.

4.8.2 Análisis de Sentimientos

En esta investigación también se realiza un análisis de sentimientos (negativo, neutro o positivo) expresado en cada comentario/tweet con el fin de determinar posturas de los clientes hacia algunos posibles competidores detectados. Debido a que el análisis de sentimientos no es un campo nuevo y ha tenido una gran popularidad, se hacen uso de algoritmos ya desarrollados, ya que esa gran popularidad ha propiciado que se investiguen y publiquen muchos estudios y algoritmos al respecto. Generalmente hacen uso de algoritmos de aprendizaje automático y en la habilidad que tienen este tipo de sistemas para la clasificación de textos a partir de las palabras y de las relaciones que se establecen entre ellas (Pauli, 2019). Para esta investigación se ha utilizado la librería pysentimiento desarrollado por Pérez et al. (2021) en su investigación “A python toolkit for sentiment analysis and social nlp tasks”.

4.8.3 Detección de Adjetivos

Una parte importante de esta investigación también es hacer detección de todos los adjetivos que expresan un usuario en un comentario, con esto se permite al usuario final del Dashboard hacer un análisis subjetivo sobre los posibles competidores detectados, ya que, si un posible competidor en el Dashboard tiene adjetivos como: “caro”, “costoso”, “bueno”, etc., se puede decir que la empresa en análisis comercializa productos de alto costo. Entonces para este fin se utiliza la librería SpaCy (Explosion AI, 2017) , la cual además de NER, también sus modelos pre-entrenados permiten hacer la detección de adjetivos. Este modelo de SpaCy utiliza el algoritmo POS.

4.8.4 Creación del Dashboard

En esta subsección se desarrolla un Dashboard en base al corpus y modelos generados utilizando los datos extraídos de las diferentes plataformas en el sector textil. Para este proceso se realiza una limpieza de datos extraídos de cada plataforma y se utiliza los modelos desarrollados en secciones anteriores para extraer todas las entidades detectadas en cada comentario/tweet, el sentimiento de cada comentario/tweet y una etiqueta (sí/no) sobre sí el comentario/tweet es un texto comparativo. Luego se crean diferentes gráficas en Power BI como diagramas de barras, nube de palabras, histogramas y también se implementan predicciones de series de tiempo, todas estas gráficas permiten presentar los resultados de una manera objetiva hacia las empresas involucradas. Para la predicción de las series de tiempo que permiten hacer estimaciones o pronósticos de los productos o empresas detectadas en el futuro, se utilizó el algoritmo exponential smooth, el cual Power BI tiene implementado internamente. Power BI tiene dos versiones de exponential smooth, una para datos estacionales (ETS AAA) y otra para datos no estacionales (ETS AAN). Power View usa el modelo apropiado automáticamente cuando inicia un pronóstico para su gráfico de líneas, basado en un análisis de los datos históricos (Pablo Moreno, 2018). El conjunto de datos final que se utiliza para el Dashboard en Power BI tiene los campos presentados en la **Tabla 16**.

Tabla 16: *Campos del Corpus utilizado para el Dashboard*

Atributo	Descripción
Id_publicacion	Identificación del comentario
Id_comentario	Identificación de las réplicas del comentario (Puede ser nulo)
comentario/tweet	Variable que contiene el texto del comentario o replica (String)
entidades	Entidades detectadas con el modelo NER generado con SpaCy
es_comparativo	Valor que indica si el texto es comparativo o no (Binario)
adjetivos	Adjetivos detectados con SpaCy en el comentario/tweet
Sentimiento	Valor que indica si el texto tiene un sentimiento positivo, negativo o neutral.

4.8.5 Objetivos de la Evaluación

Al finalizar la tarea de la recolección de datos, limpieza de los datos e implementación de los modelos desarrollados se realizó la evaluación de los resultados obtenidos. Esta evaluación se lleva a cabo mediante la utilización de Method Evaluation Model (MEM) tomando en cuenta únicamente la variable utilidad percibida (PU). La evaluación realizada se basa en las fuerzas de Michael Porter, el cual hace un análisis de varios elementos dentro del mercado que se deben tener en consideración para saber si una empresa puede ser competitiva o no, esto se realiza con el fin de estimar la utilidad del corpus desarrollado mediante la utilización del Dashboard. Michael Porter es profesor de la Universidad de Harvard en la escuela de negocios y un reconocido investigador sobre entornos competitivos entre empresas. En uno de sus artículos presenta 5 fuerzas competitivas que influyen en la rentabilidad de una empresa y argumenta que el objetivo del estratega competitivo es reconocer y manejar un entorno competitivo mirando directamente a los competidores, o contemplar una perspectiva más amplia que compita contra la organización (Bruijl, 2018; Porter, 2008). Porter sugiere la siguientes Fuerzas:

1. **Fuerza 1 - Amenaza de competidores potenciales**: Esta primera fuerza ayuda a analizar sí la posición actual puede verse afectada por la capacidad de las personas para ingresar a su mercado. Permite observar las barreras de entradas que tiene su mercado.
2. **Fuerza 2** - **Amenaza de nuevos productos**: Esta fuerza ayuda a analizar la probabilidad de que sus clientes encuentren una forma diferente de hacer lo que la empresa hace.
3. **Fuerza 3** - **Poder de negociación de proveedores**: Esta fuerza ayuda a analizar la facilidad que los proveedores tienen para aumentar sus precios. Se pueden plantear preguntas como ¿Cuántos proveedores potenciales tenemos? ¿Qué tan único es el producto o servicio que ofrecen? ¿Y qué tan caro sería cambiar de un proveedor a otro?
4. **Fuerza 4** - **Poder de negociación de clientes**: Esta fuerza ayuda a analizar cuando el mercado cuenta con pocos clientes, ya que corre el riesgo de que sí los consumidores estén organizados y se ponga de acuerdo pueden hacer que los precios bajen y la empresa contar con menos margen de actuación.
5. **Fuerza 5** - **La rivalidad entre competidores actuales**: Esta fuerza ayuda a analizar el número y el poder de los competidores como: ¿Cuántos rivales tienes? ¿Quiénes son y cómo se compara la calidad de sus productos y servicios con la suya?

Luego de tener en cuenta cada una de las fuerzas presentadas por Michael Porter se realiza un análisis de cada fuerza conjuntamente con los tutores de esta investigación para seleccionar las más relacionadas para este trabajo de investigación, tomando en cuenta que parte del objetivo de esta investigación es hacer detección de competidores. Con respecto a la fuerza 1, Porter menciona que es importante hacer un análisis del mercado, donde se observe las nuevas empresas que están incursionando en el sector, además, examinar que barreras existen para que un nuevo competidor

ingrese al mercado, debido a que es una fuerza que analiza la competencia se toma en consideración para evaluar en esta investigación.

En la fuerza 2, Porter menciona la necesidad de hacer un análisis de amenazas de productos nuevos que pueden sustituir a los productos actuales de la empresa y estén ingresando al mercado, lo cual esto afectará la rentabilidad de la empresa. Pero, debido a que es un análisis específico de productos y no es con respecto a competidores se descarta esta fuerza para la evaluación en esta investigación.

En la fuerza 3 Porter menciona que se debe analizar a los proveedores sobre la calidad de sus productos o servicios donde se observe como afectará el precio si se cambia de proveedor o cual es el costo de cambiarse de proveedor. Pero, debido a que en esta fuerza se realiza un análisis a proveedores y no a competidores entonces también se descarta esta fuerza para la evaluación en esta investigación.

En la fuerza 4, Porter menciona que se debe realizar un análisis del mercado considerando a los clientes como también los productos. Este análisis debe permitir observar de los clientes la aceptación o desacuerdo de los productos que adquieren y así tomar decisiones para que el cliente este complacido con el producto ofrecido y de esta manera en lo posible ayudar a obtener la fidelidad del cliente hacia la empresa. Debido a que se realiza un análisis a los competidores considerando sus posturas hacia una empresa entonces esta fuerza se toma en consideración para la evaluación en esta investigación.

Por último, en la fuerza 5, Porter menciona que se debe realizar un análisis a la competencia donde permita determinar el tamaño de la misma, crecimiento que está teniendo y número de competidores, pero también señala que se debe realizar análisis de tendencias de mercado para luego tomar decisiones al respecto y obtener una ventaja competitiva. Debido a que esta fuerza también realiza un análisis hacia la competencia, entonces se toma en consideración para la evaluación. En resumen, se seleccionaron tres fuerzas de Porter para la evaluación y se descartaron 2 fuerzas, las fuerzas seleccionadas son:

- **Fuerza 1**: Amenaza de competidores potenciales
- **Fuerza 4**: Poder de negociación de clientes
- **Fuerza 5**: Rivalidad entre competidores actuales

4.8.6 Selección de Evaluadores

Las personas que participaron como evaluadores fueron escogidos de una muestra por conveniencia, según la disponibilidad de recursos. Consisten en personas que estén relacionados al sector empresarial los cuales pueden ser personas a cargo de empresas o también estudiantes de último año de la Carrera ciencias económicas de la Universidad de Cuenca. Como lo indican Kitchenham et al. (2001) los evaluadores con estas características de nivel académico y que estén relacionados con el sector y ámbito de análisis poseen lo necesario para realizar tareas

profesionales y con eso se puede garantizar de que todos los participantes en el estudio tengan una experiencia comparable.

4.8.7 Proceso de Evaluación

Cada evaluador fue expuesto al Dashboard con todos los resultados obtenidos en esta investigación, el Dashboard presentado es interactivo y los investigadores dieron una sesión de capacitación referente a la utilización de esta herramienta. Luego de la presentación realizada, cada evaluador tuvo que llenar un cuestionario (Anexo 1) donde expresaron si lo que observaron respondía afirmativamente o negativamente a cada pregunta relacionado con las fuerzas de Michael Porter, con esto se evaluó la utilidad del corpus en cada fuerza de Porter

El proceso que se realizó en la evaluación se presenta a continuación (**Figura 23**). Dentro del cuestionario a desarrollar, cada atributo a evaluar se cuantifica a través de preguntas de opción múltiple de acuerdo a la escala de Likert. La escala de Likert es una escala de respuesta psicométrica utilizada principalmente en cuestionarios para obtener las preferencias de los participantes o el grado de acuerdo de una declaración o conjunto de declaraciones. Las escalas de Likert son una técnica de escala no comparativa y son unidimensionales (solo miden un rasgo único) por naturaleza. Para utilizar este método se pide a los encuestados que indiquen su nivel de acuerdo con una declaración dada por medio de una escala ordinal (Joshi et al., 2015; Southerton, 2014).

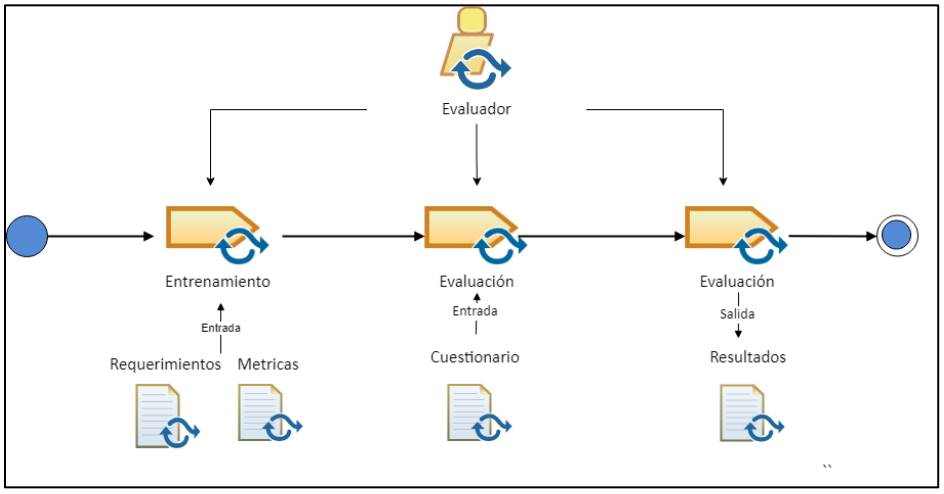


Figura 23: *Metodología para la evaluación de resultados de la Investigación*
Fuente: Construcción propia

CAPÍTULO 5: RESULTADOS Y DISCUSIÓN

En el presente capítulo se muestran los resultados obtenidos con la finalidad de cumplir cada uno de los objetivos planteados al inicio del presente trabajo.

5.1 Generación del Corpus

En esta subsección se muestran los resultados del proceso que se desarrolló para la generación del corpus. Primero se detallan los resultados más relevantes en el modelo para detectar entidades, la realización del ranking de los corpus existentes, la creación del modelo para identificar texto comparativo y un análisis estadístico del corpus final generado basado en estos resultados.

5.2 Modelo para detectar entidades

El primer paso para la detección de entidades es el etiquetado de datos de cada corpus, el etiquetado de datos se realizó manualmente por los investigadores siguiendo la metodología presentada en la **Figura 14** y en la **Tabla 17** se presenta el número de datos etiquetados de cada parte subdividida. MASS corpus tiene una cantidad de 2720 registros y se divide en dos partes, 816 datos etiquetados manualmente y 1904 datos que no se etiquetan y se utilizarán después de entrenar cada modelo. De la misma manera se procede con MOZETIC corpus, TASS corpus y SFU corpus, todos tienen un número similar de datos etiquetados a excepción de SFU corpus que tiene solo 150 datos etiquetados. El SFU corpus al tener comentarios de una página web no hay límite de palabras, motivo por el cual hay registros que solo tienen una oración, pero también hay registros que tienen un párrafo completo o incluso más. Además, este corpus tiene una columna para clasificar los textos en varios tipos, se eligió los principales tipos que se puede poseer más entidades de interés para este trabajo, por lo tanto, se seleccionaron los 150 registros que son relacionados a coches, móviles y ordenadores, ya que son los tipos que generalmente discuten de empresa y/o productos.

Tabla 17: *Número de datos de cada corpus*

Corpus	Número de Datos	Número Datos Parte 1	Número Datos Parte 2
MAS corpus	2720	816	1904
SFU corpus	400	150	250
MOZETIC corpus	131388	1106	130282
TASS corpus	14529	998	13531

Después de etiquetar la parte 1 de cada corpus se procede a implementar el modelo con la herramienta SpaCy, en esta herramienta se realiza el tuneo de parámetros y se obtienen los siguientes resultados (**Tabla 18**). Las métricas presentadas en la **Tabla 18** están relacionadas a las entidades que conforman una sola palabra, ya que las entidades que están conformadas por varias palabras se obtienen unas métricas relativamente muy bajas dado que no existen datos suficientes con esas características para que sean entrenados.

Tabla 18: *Resultado de la Detección de Entidades de Cada Corpus*

Corpus	Precisión	Recall	F-score	# Entidades encontradas	Tiempo de entrenamiento con GPU (hh:mm:ss)
MAS corpus	0.821	0.8727	0.8458	562	0:20:01.84
SFU corpus	0.8	0.6857	0.7385	459	1:01:26.48
MOZETIC corpus	0.5	0.3429	0.4068	10535	0:54:36.79
TASS corpus	0.4	0.2857	0.3333	461	0:19:05.17

En las siguientes figuras (**Figura 24**, **Figura 25**, **Figura 26**, **Figura 27**) se presentan las principales entidades encontradas en cada uno de los corpus. Es importante mencionar que estos resultados son la suma de las entidades detectadas en el etiquetado manual más las entidades detectadas utilizando el modelo NER desarrollado.

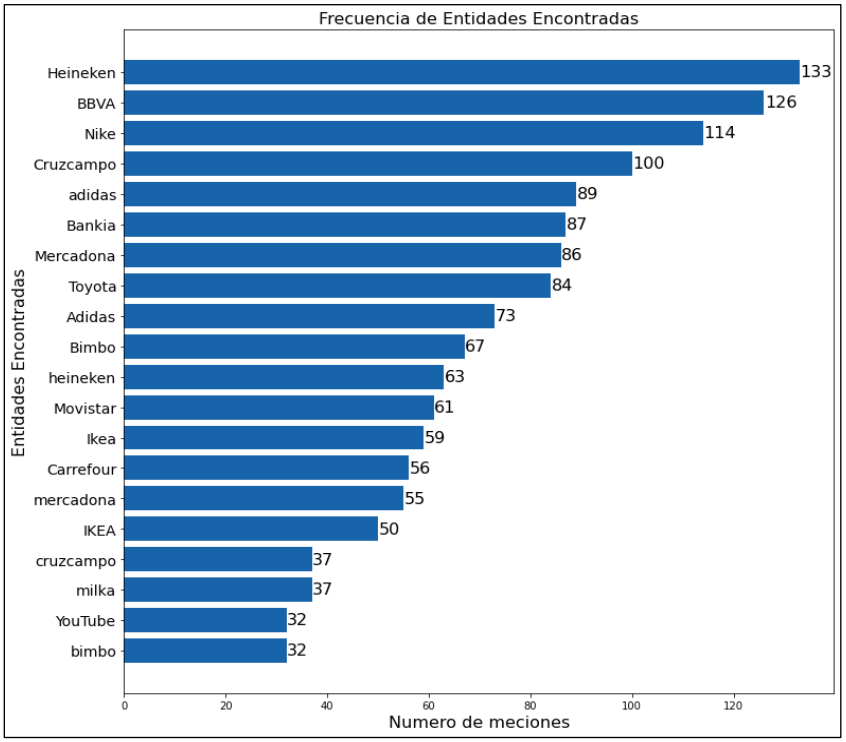


Figura 24: *Entidades de MASS Corpus*

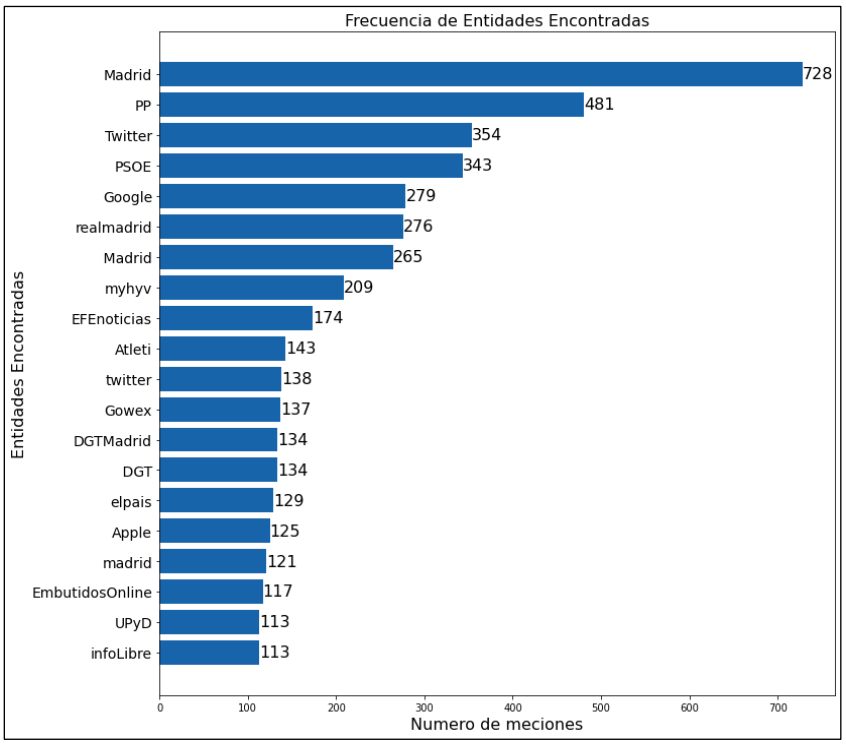


Figura 25: *Entidades de MOZETIC Corpus*

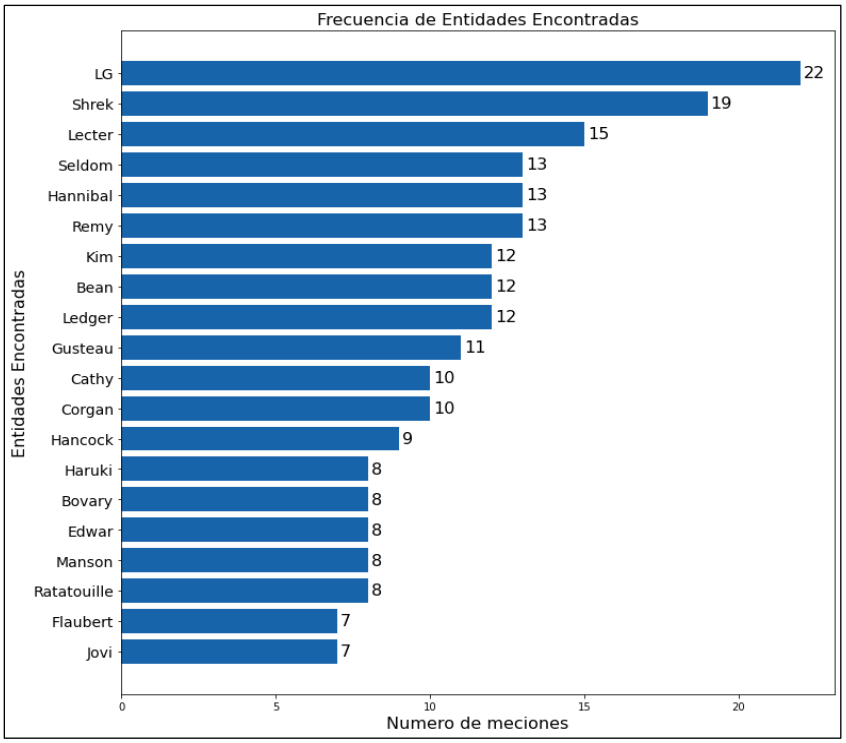


Figura 26*: Entidades de SFU corpus*

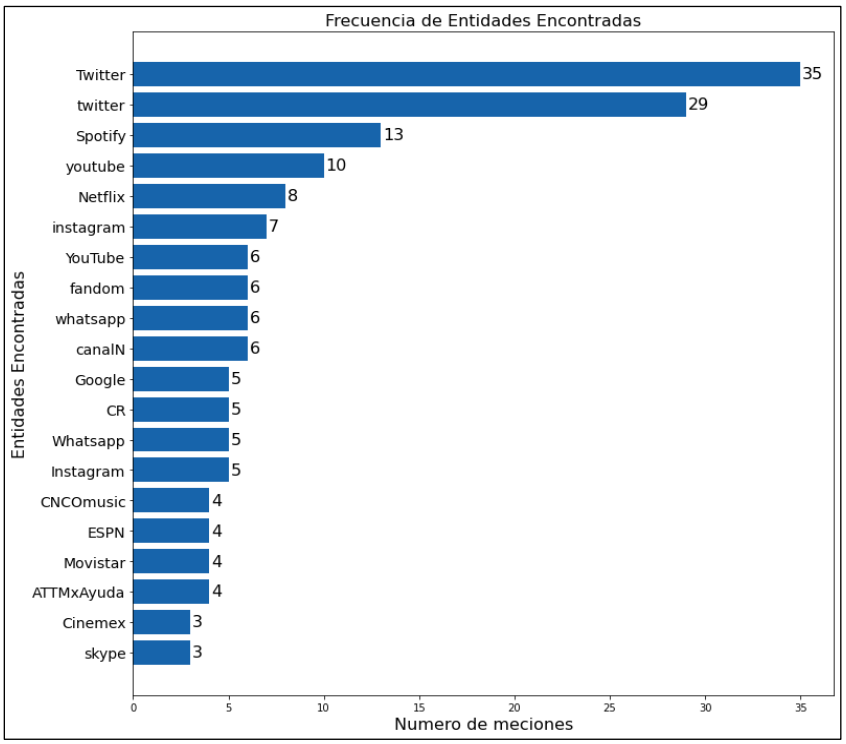


***Figura 27**: Entidades de TASS corpus*

Después de observar los resultados(**Figura 24**, **Figura 25**, **Figura 26**, **Figura 27**), se determina que el corpus que más entidades detecta es MOZETIC corpus, luego está MASS corpus, en tercer lugar se encuentra SFU corpus y el que menos entidades detecta es TASS corpus, a pesar que es el segundo con mayor número de registros. MOZETIC corpus está primero en cuanto al número de registros. Por otra parte, al observar la **Tabla 18** se puede determinar que MASS corpus tiene las mejores métricas de evaluación en el modelo NER con respecto los otros corpus, y eso también se puede visualizar en las entidades que presenta **Figura 24**, donde hay muy pocas palabras que no tengan sentido (entidades mal detectadas), mientras que en las gráficas de los otros corpus existen más entidades mal detectadas.

5.3 Características Comparativas

El siguiente paso es determinar sí en el texto de cada registro de un corpus los usuarios realizan comparaciones de productos, empresas, entre otros. Esto ayuda a determinar algunas empresas o productos que son competidores. Para esta característica se realizó una búsqueda por diccionario a cada corpus basándose en adjetivos y palabras clave utilizados presentadas en la metodología. Con estos adjetivos y palabras clave se buscaron sinónimos para extender la lista y generalizar el procedimiento. Al final se obtuvieron 45 palabras clave, con las cuales se realizó la búsqueda en cada uno de los corpus y en la **Tabla 19** se presentan los resultados obtenidos.

***Tabla 19**: Resultados en la detección de características comparativas*

Corpus	# Datos	# Texto con características comparativas	Relación con el número de datos	# Texto Características comparativas (10)	Relación con el número de datos (10)
MAS corpus	2714	160	5.89%	4	6
SFU corpus	400	312	78.0 %	6	10
MOZETIC corpus	131388	7064	5.19%	10	4
TASS corpus	14529	1463	10.06%	8	8

5.4 Análisis de Similitud de Corpus

El SFU corpus tiene datos de comentarios recolectados de una página de comercio electrónico, al hacer un análisis exploratorio y al trabajar con dicho corpus en las características descritas anteriormente, se encuentra que tiene registros con cantidades de palabras muy diferentes a otros corpus, por lo tanto, se procede a realizar un análisis de similitud de corpus para obtener datos cuantitativos sobre la diferencia de datos que existe. Por otra parte, también se realiza una investigación sobre la similitud de datos en páginas de comercio electrónicos semejantes en nuestra región (Ecuador y países vecinos como Colombia y Perú), es decir, se investiga sí en páginas de comercio electrónico semejantes existen comentarios con una gran cantidad de palabras. Luego de dicha investigación se encontraron las siguientes plataformas: Mercado Libre y OLX como plataformas similares. Pero estas plataformas no permiten a sus usuarios hacer comentarios de sus productos, solo tienen una opción para hacer preguntas las cuales generalmente consultan por los precios y plazos de entrega. Además, sus APIs son bastantes limitadas, Mercado Libre solo permite obtener enlaces de búsquedas que están en tendencia en la plataforma y en OLX sólo se permite administrar los propios productos.

Después del primer análisis investigativo se desarrolló un diagrama de caja con la cantidad de palabras que tiene cada registro de cada corpus. Con esto se puede observar que los textos del SFU corpus varían con respecto a los otros corpus (**Figura 28**), se eliminaron valores atípicos, con oraciones de hasta cerca de 5000 palabras.

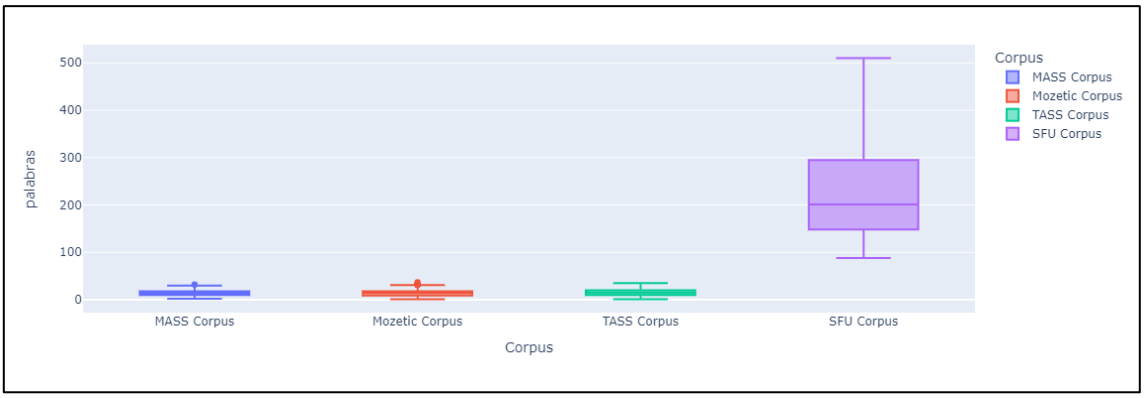


***Figura 28**: Diagrama de Caja de Número de Palabras en cada texto en los diferentes corpus*

En la **Figura 29** se puede observar con más claridad la diferencia de las distribuciones que tiene el SFU corpus con respecto a los otros corpus. Se puede ver que la media de palabras del SFU corpus es de 271 palabras por oración mientras que en los otros corpus son de 14 a 15 palabras. Esta gran desigualdad de datos con respecto al número de palabras por registro y la no semejanza de datos con plataforma de la región hace que el SFU corpus tenga una calificación de cero en la característica Utilidad y Similitud de corpus

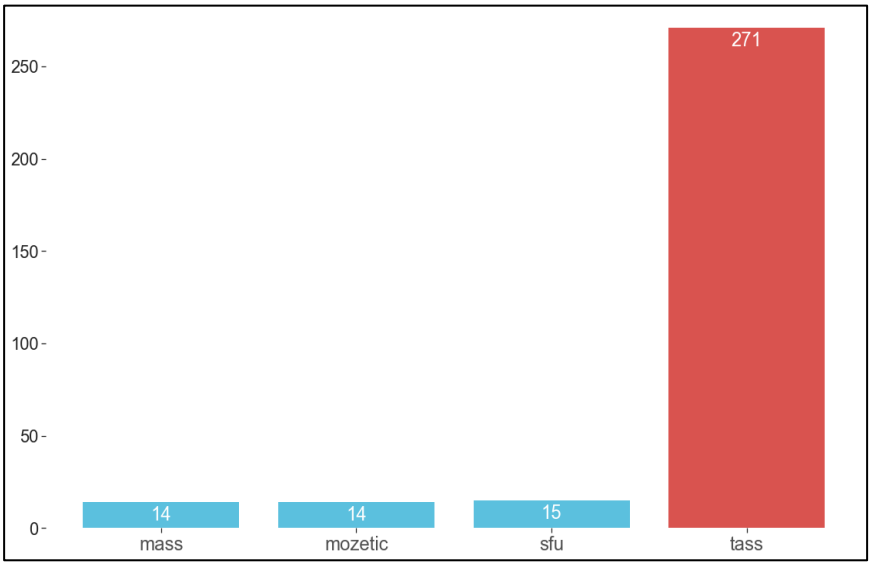


***Figura 29**: Medianas del Número de Palabras de Cada Texto en los Diferentes Corpus*

5.5 Ranking de los Corpus Existentes

Tomando en cuenta todas las características antes mencionadas, se realizó un ranking de los corpus, para ello se asigna a cada característica un peso de acuerdo a la importancia con el objetivo de esta investigación (**Tabla 20**). Como resultado final se obtiene que MAS Corpus y MOZETIC corpus son los mejores rankeados respectivamente.

***Tabla 20**: Valoración de características para el ranking de corpus*

	Detección de entidades (sobre 50)		Detección de características comparativas (sobre 40)		Número de entidades detectadas (sobre 10)		Utilidad y Similitud de los Corpus		Score Final (Sobre 100)
Plataforma	**Calificación**	**Peso**	**Calificación**	**Peso**	**Calificación**	**Peso**	**Calificación**	**Peso**	
MAS corpus	8,45	5	5	4	0,533	1	10	2	69,3334
MOZETIC corpus	4,06	4	7	3	10	1	10	2	67,24
TASS corpus	3,33	4	8	3	0,438	1	10	2	57,76
SFU corpus	7,38	4	8	3	0,436	1	0	2	53,96

En la **Figura 30** se presenta de manera gráfica el score final del ranking realizado a los diferentes corpus.

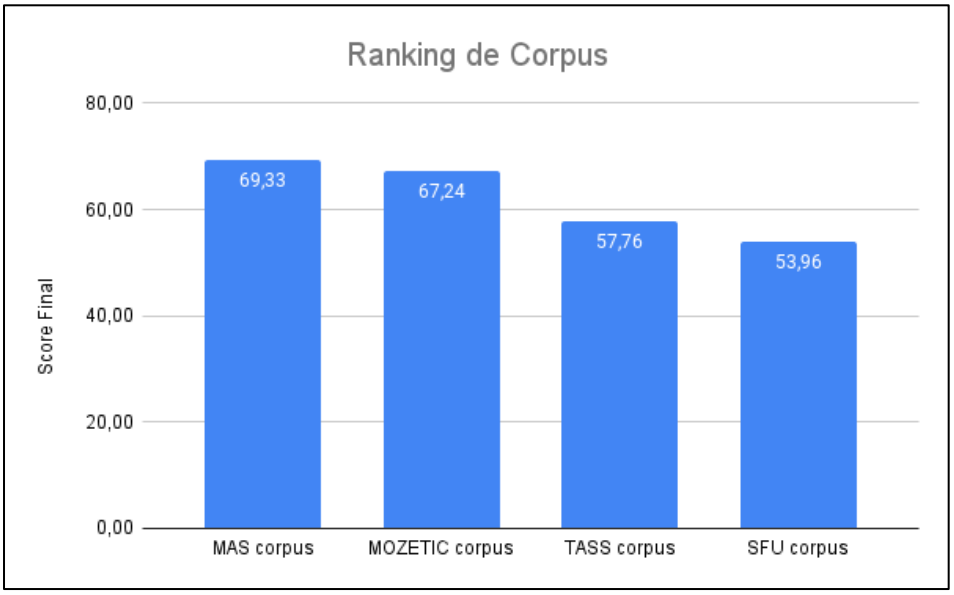


***Figura 30**: Ranking de Corpus*

Después de este análisis se observa que los corpus en las tres primeras posiciones del ranking realizado son: MÁS corpus, Mozetic corpus y TASS corpus. Un corpus al estar en las primeras posiciones significa que es más probable que se pueda utilizar para la inteligencia competitiva debido a que tiene varias menciones de posibles competidores, además también tiene texto comparativo lo cual indica que puede existir comentarios/tweets que hagan comparaciones entre posibles competidores. Por último, se realiza un análisis de similitud con respecto al número de palabras de cada registro en cada corpus, esto realiza con el objetivo de tener corpus balanceados

al momento de implementar algoritmo de aprendizaje automático y así conseguir métricas de evaluación aceptables para una investigación (recall, precisión y f-score con al menos 70%).

5.6 Modelo para identificar texto comparativo

La creación del modelo, como se explicó es un proceso cíclico y de varias etapas (sección **4.6**), por lo que a continuación se detalla cada análisis realizado para obtener los mejores resultados.

5.6.1 Primer análisis

Este análisis se lleva a cabo con una parte de los tres corpus mejor ubicados en el ranking, los cuales fueron etiquetados manualmente como comparativos y no comparativos. El corpus en total consta de 693 textos. Del total de textos, únicamente 71 (10.24%) fueron etiquetados como comparativos. Así también, 485 textos fueron usados para entrenamiento y 208 para prueba.

Resultados del análisis

La **Tabla 21** muestra los resultados obtenidos al aplicar diferentes clasificadores de aprendizaje automático con su respectivo tuneo de hiperparámetros: Naive Bayes, Regresión Logística, SVM, Random Forest, y RNA, en términos de diferentes medidas de evaluación como accuracy, precisión, recall, f1 score y ROC (AUC) ponderado. En la ***Figura 31*** se presenta los resultados de cada algoritmo implementado.

Tabla 21: *Resultados comparativos de diferentes clasificadores de aprendizaje automático (Análisis 1).*

Clasificador	Accuracy	F1 Score		Precisión		Recall		Roc (AUC)
		Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	
Naive Bayes	0.76	0.55	0.24	0.55	0.20	0.57	0.31	0.56
Random Forest	0.86	0.63	0.32	0.72	0.55	0.60	0.23	0.60
Regresión Logística	0.87	0.50	0.07	0.69	0.50	0.52	0.04	0.51
SVM	0.88	0.66	0.39	0.72	0.53	0.63	0.31	0.63
Red Neuronal	0.87	0.62	0.31	0.68	0.46	0.60	0.23	0.59

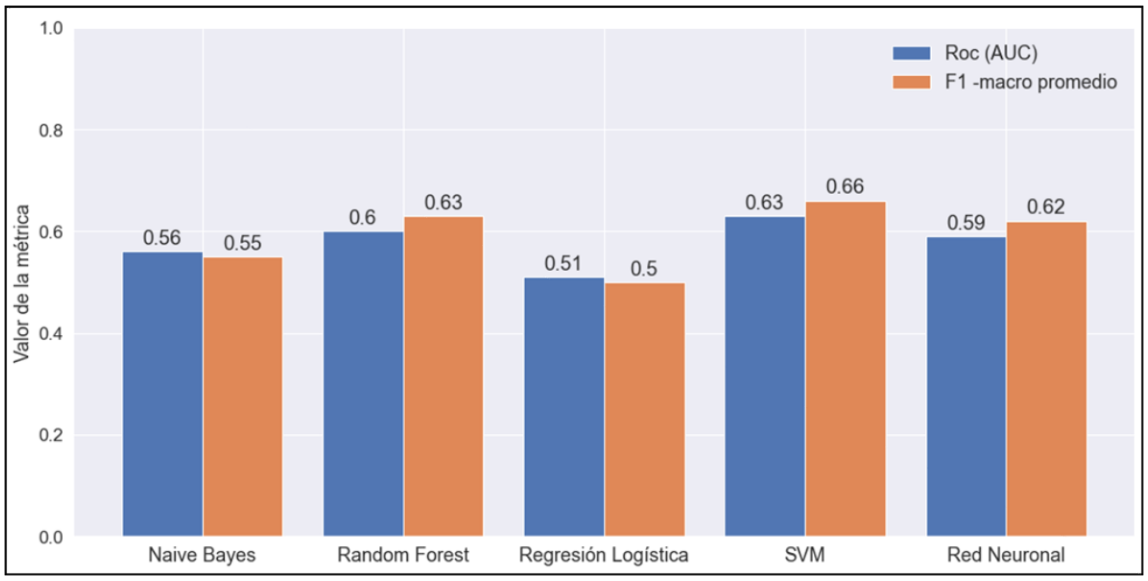


Figura 31: *Gráfico de barras del F1-macro-promedio y Roc (AUC) ponderado de los diferentes clasificadores.*

Análisis y discusión del Resultado

- Al tener datos desbalanceados, el accuracy no es una métrica confiable para determinar qué modelo es mejor, es por esto que se le asignó mayor relevancia al f1 score y al Roc (AUC) ponderado.
- Los dos mejores clasificadores son SVM y Random Forest con un valor de ROC (AUC) ponderado igual a 0.63 y 0.60 respectivamente (**Tabla 21** y Figura 31).
- Los valores obtenidos por las métricas de evaluación son bajos, especialmente si se analiza el f1 de la clase comparativa. Esto se debe a que el conjunto de datos se encuentra muy desbalanceado. Por este motivo los clasificadores no generalizan correctamente y no son confiables para detectar texto comparativo.
- Los resultados podrían mejorar agregando nuevos datos comparativos para tratar de balancear las clases.
- Se puede evidenciar que incluso el corpus mejor rankeados por sí solos no podrían ser de total utilidad para la detección de competidores, lo que genera mayor interés en poder generar un corpus para esta finalidad.

5.6.2 Segundo análisis

Este análisis se lleva a cabo con los textos etiquetados manualmente de los tres corpus mejor ubicados en el ranking y con los nuevos textos etiquetados manualmente como comparativos provenientes de todos los corpus disponibles, a excepción de SFU corpus, el cual se demostró en la sección **5.4** que no es de utilidad para este estudio. En total se tiene 823 textos. Del total de textos, 204(24.79%) fueron etiquetados como comparativos.

Resultados del análisis

La **Tabla 22** y la **Figura 32**, **Figura 33** muestra los resultados obtenidos al aplicar diferentes clasificadores de aprendizaje automático con su respectivo tuneo de hiperparámetros: Naive Bayes, Regresión Logística, SVM, Random Forest, y RNA, en términos de diferentes medidas de evaluación como accuracy, precisión, recall, f1 score y ROC (AUC) ponderado.

***Tabla 22**: Resultados comparativos de diferentes clasificadores de aprendizaje automático (Análisis 2).*

Clasificador	Accuracy	F1 Score		Precisión		Recall		Roc (AUC)
		Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	
Naive Bayes	0.81	0.69	0.50	0.74	0.65	0.67	0.41	0.66
Random Forest	0.87	0.82	0.73	0.83	0.76	0.81	0.69	0.81
Regresión Logística	0.88	0.84	0.75	0.84	0.76	0.84	0.75	0.83
SVM	0.89	0.83	0.74	0.86	0.82	0.82	0.68	0.815
Red Neuronal	0.87	0.80	0.69	0.79	0.68	0.80	0.69	0.79

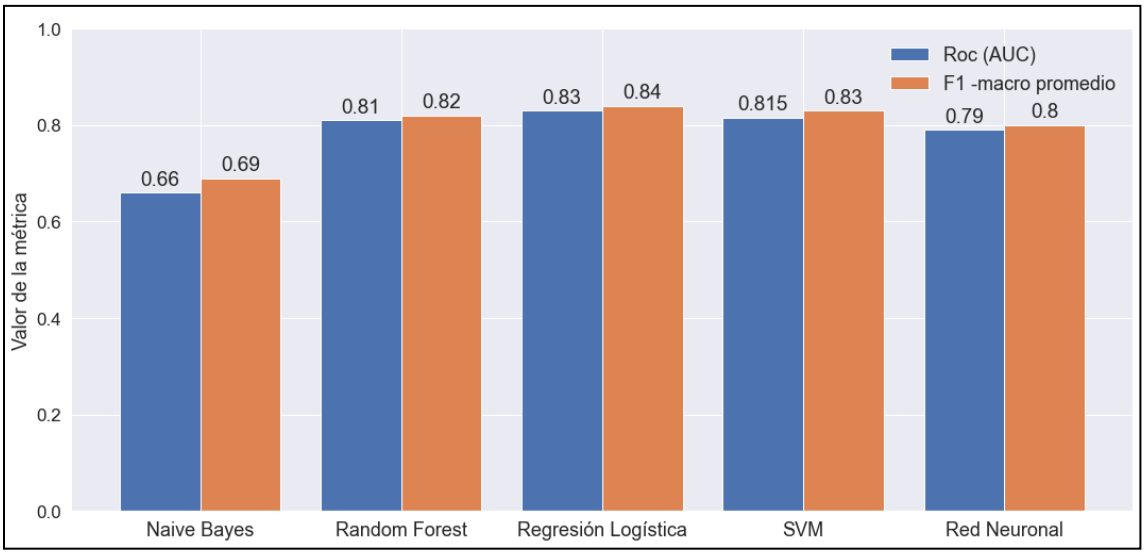


***Figura 32**: Gráfico de barras del F1-macro-promedio y Roc (AUC) ponderado de los diferentes clasificadores del análisis 2.*

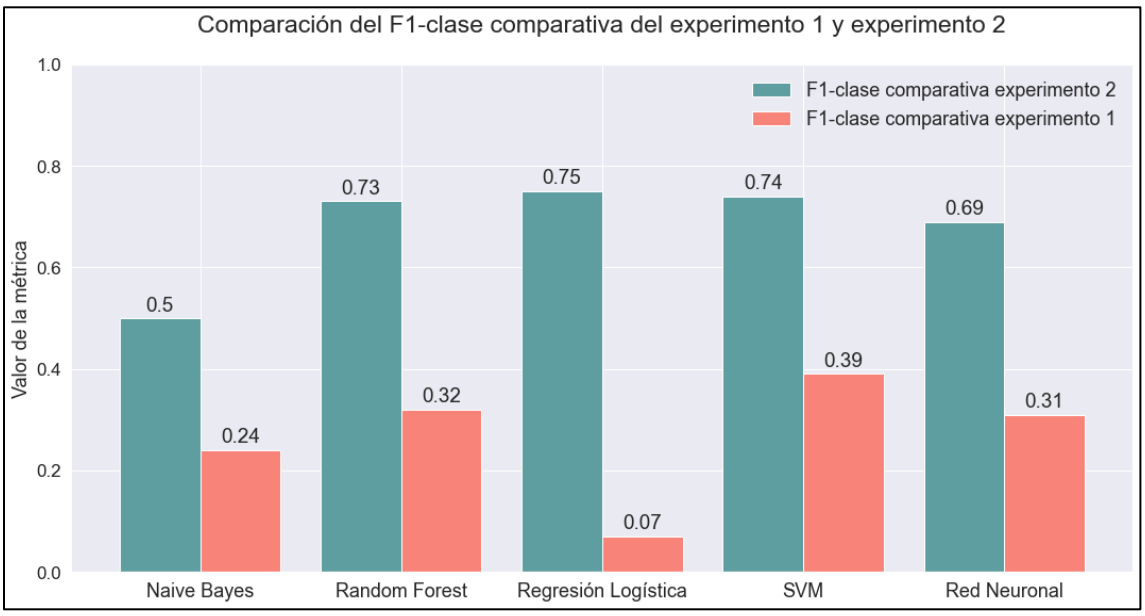


Figura 33: *Gráfico de barras de la comparación del F1 score-clase comparativa del análisis 1 y análisis 2.*

Análisis y discusión del Resultado

- Los dos mejores clasificadores son Regresión Logística y SVM con un valor de Roc (AUC) ponderado igual a 0.84 y 0.83 respectivamente (**Tabla 23** y **Figura 32**).
- Se puede observar que existe una notable mejoría en los resultados comparando con los del análisis 1, esto queda evidenciado en la **Figura 33**, donde se muestra una gran diferencia en los clasificadores cuando se trata de identificar el texto comparativo.
- Los resultados mejoraron notablemente al agregar nuevos datos comparativos, lo que genera un conjunto de datos más balanceado y una mejora en los modelos.

5.6.3 Tercer Análisis

Este análisis se lleva a cabo con los textos etiquetados manualmente de los tres corpus mejor ubicados en el ranking y con los nuevos textos etiquetados manualmente como comparativos provenientes de todos los corpus disponibles, a excepción de SFU corpus, el cual se demostró en la sección 5.4 que no es de utilidad para este estudio. En total se tiene 857 textos. Del total de textos, 238(27.77%) fueron etiquetados como comparativos. También vale la pena mencionar que 599 textos fueron usados para entrenamiento y 258 para prueba.

Resultados del Análisis

La **Tabla 23** y la **Figura 34**, **Figura 35** muestra los resultados obtenidos al aplicar diferentes clasificadores de aprendizaje automático con su respectivo tuneo de hiperparámetros: Naive Bayes, Regresión Logística, SVM, Random Forest, y RNA, en términos de diferentes medidas de evaluación como accuracy, precisión, recall, f1 score y ROC (AUC) ponderado.

***Tabla 23**: Resultados comparativos de diferentes clasificadores de aprendizaje automático (Análisis 3).*

Clasificador	Accuracy	F1 Score		Precisión		Recall		Roc (AUC)
		Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	
Naive Bayes	0.85	0.75	0.59	0.84	0.82	0.71	0.46	0.71
Random Forest	0.85	0.80	0.69	0.79	0.66	0.81	0.73	0.81
Regresión Logística	0.89	0.83	0.74	0.85	0.79	0.82	0.69	0.82
SVM	0.90	0.85	0.77	0.86	0.80	0.85	0.75	0.85
Red Neuronal	0.88	0.82	0.72	0.83	0.74	0.82	0.71	0.818

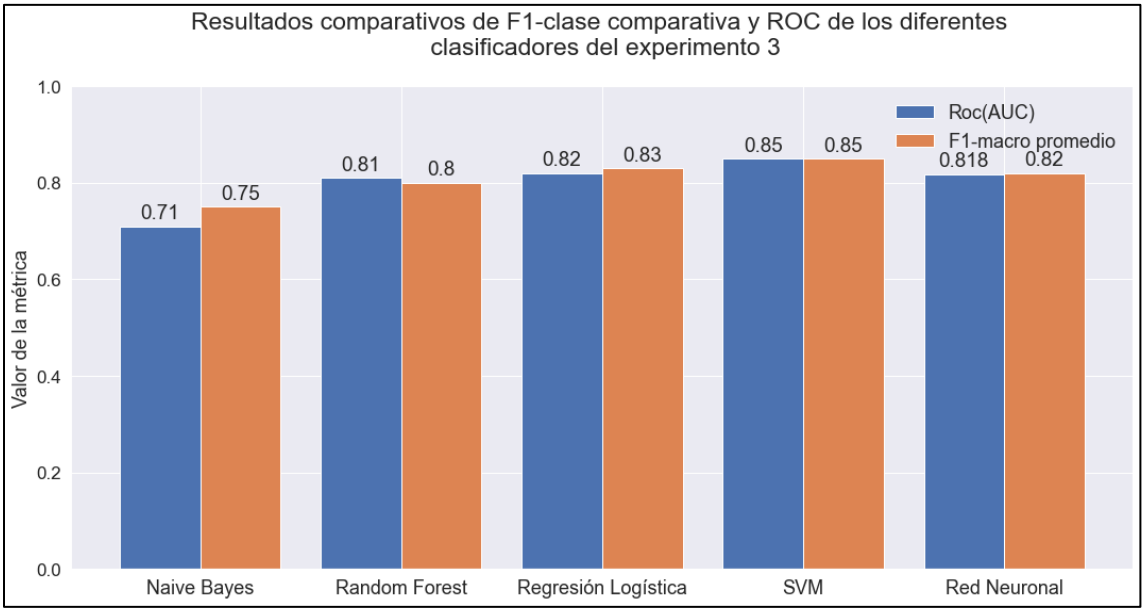


***Figura 34:** Gráfico de barras del F1-macro-promedio y Roc (AUC) ponderado de los diferentes clasificadores del análisis 3.*

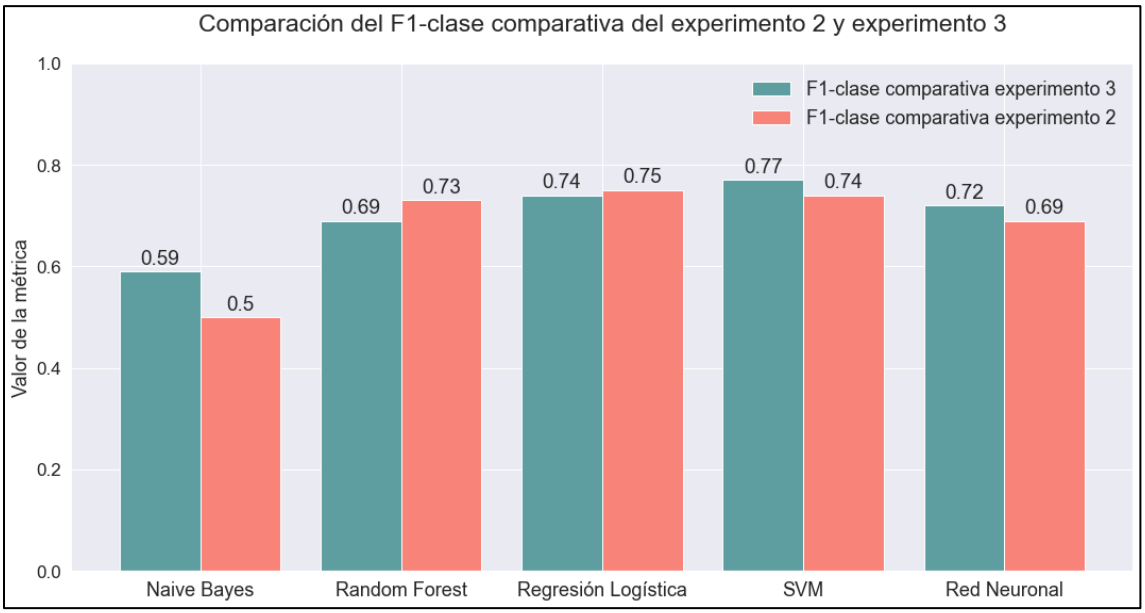


***Figura 35**: Gráfico de barras de la comparación del F1 score-clase comparativa del análisis 2 y análisis 3.*

Análisis y discusión del Resultado

1. Los dos mejores clasificadores son SVM y Regresión Logística con un valor de Roc (AUC) ponderado igual a 0.85 y 0.82 respectivamente (**Tabla 23** y **Figura 34**).
2. En la **Figura 34** se puede observar que los resultados tienen cierta mejora comparando con los del análisis 2, en los clasificadores Naive Bayes, SVM y Red Neuronal.
3. También en la **Figura 35** puede observar que en el caso de Random Forest y Regresión Logística disminuye el rendimiento en la clasificación de la clase comparativa comparando con los del análisis 2.
4. En general no se ha tenido una gran mejora en los resultados comparando con el análisis 2.

5.6.4 Cuarto Análisis

Último análisis realizado, y se lleva a cabo con los textos etiquetados manualmente de los tres corpus mejor ubicados en el ranking y con los nuevos textos etiquetados manualmente como comparativos provenientes de todos los corpus disponibles, a excepción de SFU corpus, el cual se demostró en la sección 5.4 que no es de utilidad para este estudio. En total se tiene 921 textos. Del total de textos, 302 (32.79%) fueron etiquetados como comparativos.

Resultados del Análisis

La **Tabla 24** y **Figura 36**, **Figura 37** muestra los resultados obtenidos al aplicar diferentes clasificadores de aprendizaje automático con su respectivo tuneo de hiperparámetros: Naive Bayes, Regresión Logística, SVM, Random Forest, y RNA, en términos de diferentes medidas de evaluación como accuracy, precisión, recall, f1 score y ROC (AUC) ponderado.

***Tabla 24**: Resultados comparativos de diferentes clasificadores de aprendizaje automático (Análisis 4).*

Clasificador	Accuracy	F1 Score		Precisión		Recall		Roc (AUC)
		Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	Macro Promedio	Clase comparativa	
Naive Bayes	0.85	0.83	0.78	0.83	0.73	0.85	0.84	0.848
Random Forest	0.87	0.84	0.78	0.86	0.83	0.83	0.74	0.836
Regresión Logística	0.88	0.85	0.79	0.87	0.84	0.84	0.75	0.841
SVM	0.88	0.86	0.81	0.87	0.85	0.85	0.77	0.853
Red Neuronal	0.85	0.83	0.76	0.83	0.77	0.82	0.75	0.823

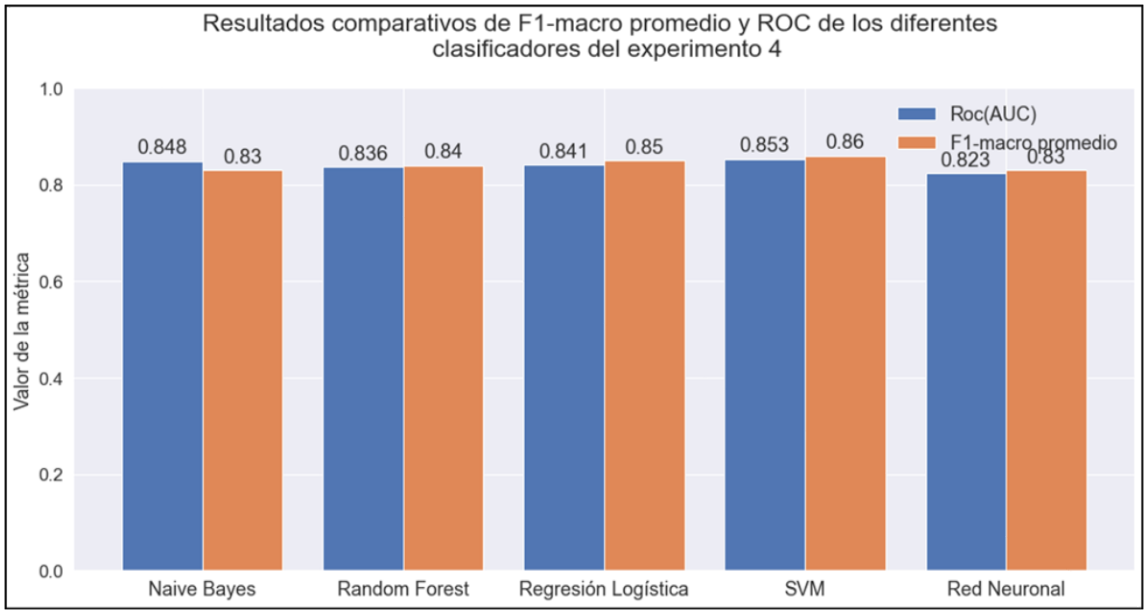


***Figura 36**: Gráfico de barras del F1-macro-promedio y Roc (AUC) ponderado de los diferentes clasificadores del análisis 4.*

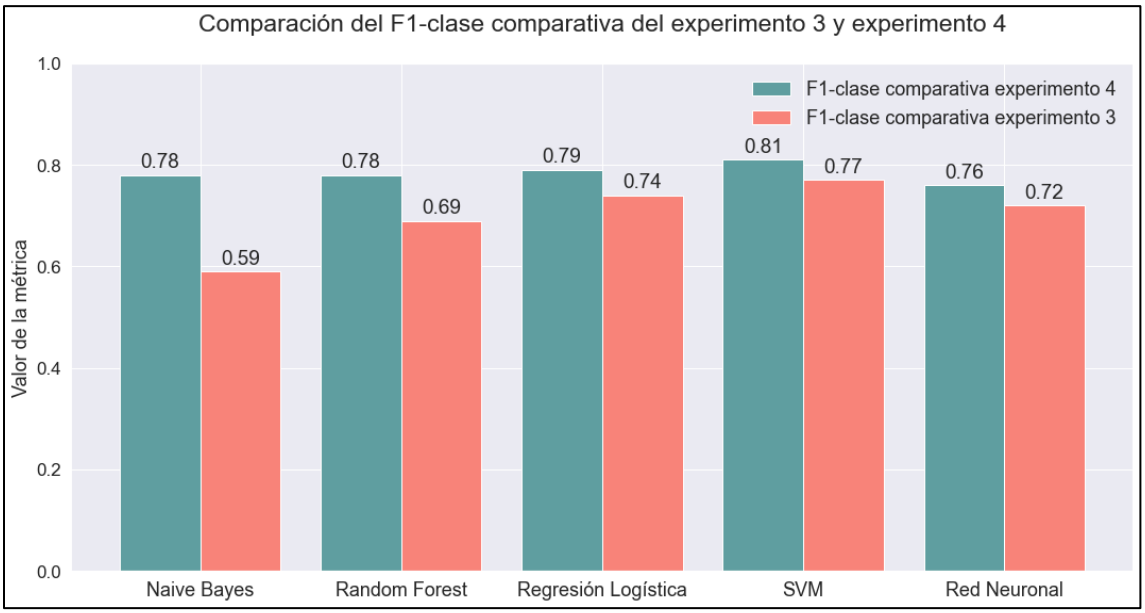


Figura 37: *Gráfico de barras de la comparación del F1 score-clase comparativa del análisis 3 y análisis 4.*

En este análisis se aplicó el test de McNemar a los dos mejores clasificadores (SVM y Naive Bayes) según el ROC(AUC) ponderado para determinar si son estadísticamente diferentes. La **Tabla 25** muestra el resultado de la tabla de contingencia del test, y a continuación se detalla los resultados del mismo.

Tabla 25*: Tabla de contingencia del test de McNemar.*

	Naive Bayes (clasificados correctamente)	**Naive Bayes (incorrectos)**
SVM (clasificados correctamente)	227	18
SVM (incorrectos)	9	23

Para el cálculo del test de McNemar se partió con un valor de alpha igual a 0.05 y la siguiente hipótesis nula:

H0: No existe una diferencia estadísticamente significativa entre los clasificadores SVM y Naive Bayes (ninguno de los dos clasificadores rinde mejor que el otro).

Una vez realizado el test, se obtuvo un p-value = 0.124, por lo tanto, se tiene la misma proporción de errores (No se rechaza H0), es decir: No existe una diferencia estadísticamente significativa entre los clasificadores SVM y Naive Bayes.

Análisis y discusión del Resultado

1. Los dos mejores clasificadores son SVM y Naive Bayes con un valor de Roc (AUC) ponderado igual a 0.853 y 0.848 respectivamente (**Tabla 24** y **Figura 36**).
2. En la **Figura 37** se puede observar que los resultados tienen cierta mejora comparando con los del análisis 3, en función de la clase comparativa, especialmente en Naive Bayes.
3. En este punto al contar con resultados muy buenos en relación al Roc (AUC) ponderado, y al no tener una gran diferencia entre los análisis 2,3 y 4; se decidió que este análisis sea el final y comparar los dos mejores modelos para decidir el mejor.
4. El test de McNemar demostró que no existe una diferencia estadísticamente significativa entre los modelos, es decir el rendimiento de los dos clasificadores es similar. Por lo tanto, se seleccionó al modelo creado con Naive Bayes ya que su formulación es menos compleja que SVM.
5. Naive Bayes es usado con frecuencia por parte de los investigadores en la minería de opiniones comparativas (Varathan, K. et al., 2017) al ser un modelo con menos Bias(diferencia entre el valor del modelo y el real) que SVM tiene más probabilidades de generalizar nuevas observaciones. En este trabajo de titulación fue seleccionado como mejor modelo.

5.7 Estadísticas del corpus final

Con los modelos para detectar entidades y texto comparativo, se etiquetó todos los textos pendientes y se generó el corpus final (sección **4.7**). A continuación, en la **Figura 38** y **Figura 39** se detalla las estadísticas del corpus.

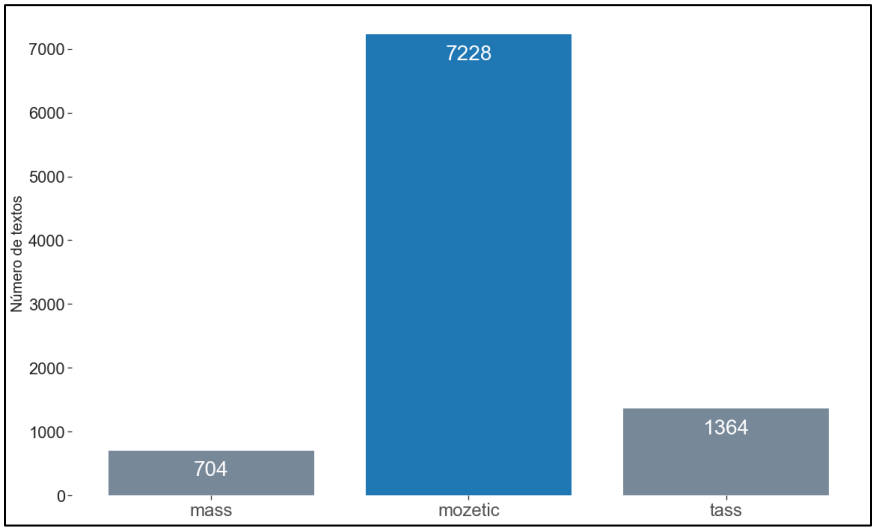


***Figura 38**: Gráfico de barras con el número de textos que aportó cada corpus original al corpus generado en este trabajo de titulación.*

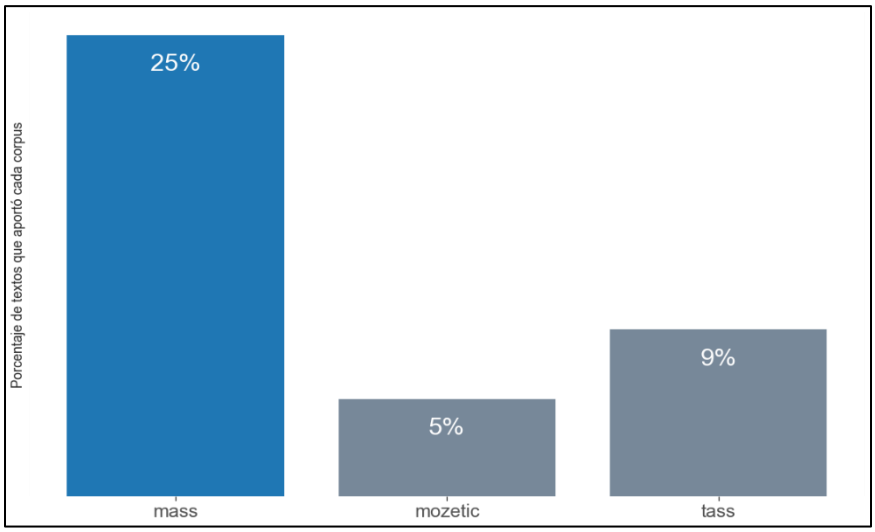


***Figura 39**: Gráfico de barras con el porcentaje que cada corpus aportó con relación al total de sus textos.*

El corpus original que más textos aportó al corpus final (**Figura 38**) fue Mozetic Corpus con 7228, seguido de TASS Corpus con 1364 y por último Mass Corpus con 704 textos. En cambio, la **Figura 39** muestra el porcentaje que cada corpus aportó con relación al total de sus textos, donde Mass Corpus aporta con el 25% del total de sus textos, TASS Corpus el 9% y Mozetic únicamente el 5%.

Comparando las **Figura 38** y **Figura 39** se puede observar que cada corpus individualmente tiene sus limitaciones ya sea de tamaño o de relevancia, por lo que, la identificación de los textos más importantes para CI de cada corpus y la unión de los mismos en uno solo, genera un valor agregado para futuras investigaciones.

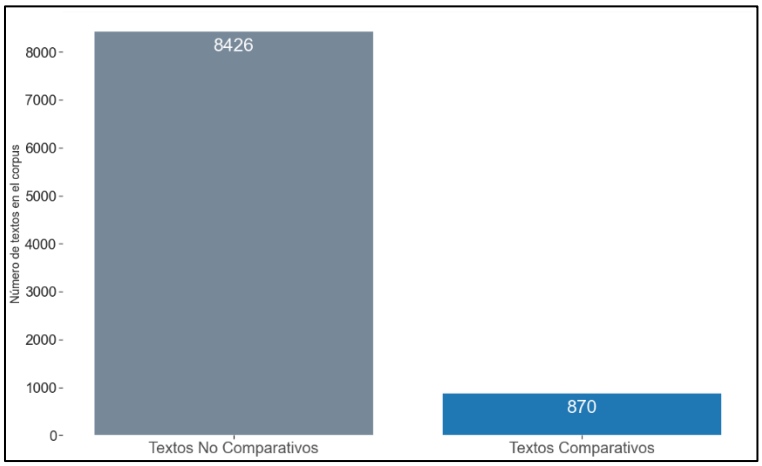


***Figura 40**: Gráfico de barras de la diferencia entre los textos comparativos y no comparativos del corpus.*

La **Figura 40** muestra que en el corpus existen 8,426 (90.64%) textos no comparativos y 870 (9.36%) textos comparativos. Esto tiene sentido de acuerdo a Kessler & Kuhn (2014) y Wang et al. (2015), donde se dice que únicamente uno de cada diez textos creados por usuarios en internet es

comparativo. Esto demuestra que el modelo realiza un correcto etiquetado de textos comparativos incluso en textos que no fueron parte del entrenamiento ni de la prueba, por lo tanto, destaca su generalidad. Por último, a este corpus final integrado por los tres corpus mejores rankeados se realiza un nuevo modelo general para NER que sirve para realizar la detección de competidores en el Dashboard. En este modelo general se utilizó todos los datos etiquetados manualmente y un subconjunto de datos etiquetados con los modelos NER de cada corpus (Ver **Figura 19**). Luego se implementa el script para convertir los datos etiquetados (Columna Entidades) de este corpus final en datos para entrenamiento de NER con SpaCy (**Figura 21**).

El corpus final con datos etiquetados está compuesto por 9296 registros, luego con estos datos se procede a dividir en datos de entrenamiento (80%), datos de prueba (10%), y datos de desarrollo (10%), por último, se entrena un modelo con SpaCy y se obtiene el resultado presentado en la **Tabla 26**.

***Tabla 26**: Tabla de Métricas del Modelo Final.*

Métricas para el Modelo	Valor
Precision	0.7837
Recall	0.7800
F-score	0.7819

5.8 Caso de Estudio

En esta sección se presenta la extracción de datos de cada plataforma seleccionada para el caso de estudio, el sector textil. En cada plataforma digital utilizada se aplicó las diferentes técnicas que se explican en la metodología, como la extracción de datos mediante API y mediante Web Scraping respectivamente, las fechas de recolección de datos se realizó desde finales del mes de diciembre del 2021 hasta el mes de abril del año 2022.

Debido a que Facebook tiene una gran cantidad de datos de muchas industrias como videojuegos, viajes, comercialización de productos, entre otros., en esta investigación se extrajeron datos de grupos dedicados a la comercialización de productos, donde se verifica que tenga al menos una mención en sus publicaciones sobre productos del sector textil como ropa, tela, etc. Los grupos de Facebook fueron escogidos manualmente debido a que, al usar técnicas como búsqueda por Hashtags se obtiene información de toda una región, por ejemplo, sí se busca el hashtag 'textil' se obtiene información de Colombia, Perú, Honduras, etc., la cual no es de interés para esta investigación. Los grupos seleccionados después de un análisis realizado considerando las características antes presentadas se muestra en la **Tabla 27**. En total se seleccionaron 13 grupos en los cuales también se toma en consideración que cada grupo seleccionado tenga publicaciones actuales, que el grupo esté conformado por más de 1000 integrantes y que al menos tengan una publicación de artículos del sector textil.

Tabla 27: *Características de grupos de Facebook al 18 enero del 2022*

Grupo	Descripción del Grupo	Número Integrantes	Tipo Grupo	Fecha Creación	Ciudad	Fecha Última Publicación	Id
VENTA DE ROPA AMERICANA BARATO CUENCA ECUADOR	Grupo que publican artículos para vender o para comprar.	83.581	Público	28/09/2016	Cuenca	18/01/2022	195414414214577
Clasificado Cuenca	Grupo que publican artículos de compra y venta.	138670	Público	31/10/2018	No delimita	18/01/2022	1906353212788111
Mercadito Cuenca	Grupo donde pueden vender, comprar e intercambiar cualquier clase de producto.	112894	Público	28/01/2016	Cuenca	18/01/2022	1563674543958142
Compra y Vende Cuenca	Grupo para comprar y vender cualquier tipo de producto.	64079	Público	20/10/2016	Cuenca	18/01/2022	1788353448109972
BUENO, BONITO Y BARATO. VENTAS DE TODO AZUAY-CUENCA	Grupo de compra, venta y cambio para Cuenca-Azuay	44677		02/06/2016	Cuenca	18/01/2022	1801507486745190
Mercado Libre Cuenca-Ecuador(Oficial)	Grupo dedicado a la comercialización de distintos productos.	41649	Público	15/09/2017	Cuenca	18/01/2022	1468205923255400
El bazar de ropa usada en Medellín, el ave fénix	Grupo dedicado a la compra-venta productos.	25125	Público	28/07/2021	General	18/01/2022	154658873437860
Promociones en Cuenca	Grupo dedicado a la compra y venta de productos de toda la provincia del Azuay.	59681 personas	Público	07/10/2015	Cuenca	18/01/2022	1695016340732257
Venta de Garage Cuenca	Grupo dedicado a la compra y venta de productos nuevos o se segunda mano que tengan el mejor precio y estén en buen estado.	49161	Privado	23/09/2014	Azogues Cuenca Cañar	18/01/2022	1492871327637970
COMPRA - VENTA - RENTA CUENCA	Grupo dedicado a la compra, venta y también renta de bienes y servicios.	25563	Público	27/06/2018	Cuenca	18/01/2022	2116573765297647
LOS HINCHAS DEL CUENQUITA	Grupo dedicado a la compra y venta de productos.	6491	Público	11/12/2013	No delimita	18/01/2022	420510008076650

En el caso de Twitter se utiliza la API oficial de esta plataforma la cual permite extraer datos solo de los últimos 7 días, por lo que se realizó un script que se ejecutaba cada semana para extraer los datos. Este script se aplicó para extraer datos de todos los tweets y retweets que contengan alguna o algunas palabras del conjunto de palabras relacionados al sector textil. Para el conjunto de palabras se buscó manualmente todas las palabras relacionadas al sector, este conjunto de palabras se presenta a continuación:

"textil, pijama, dagaronights, ropa, TEXTIMALL, prendas, dagaro, toallas, licra, pantalon, medias, lycra, nylon, Nylon, Kálido, kalido, edredones, Edredones, cortinas, toalla, lenceriadehogar, decoracion, cortinasllanas, homeandcotton, mantas,poliéster, algodón, cobertores, sabanas, sábanas, chompas, chompa, camiseta,camisetas, medias, pantalón, pantaloneta, casacas, pantalonetas, casacas , casaca , cobertores , colcha , colchas , guante , guantes , sudadera, sudaderas , vestido , vestidos , mascarilla , mascarillas , terno , ternos , falda , faldas"

De la misma manera que Twitter, para la plataforma YouTube se utilizó también la API oficial y se aplicó el mismo diccionario de palabras utilizado en Twitter. Con esto se extrae todos los videos conjuntamente con sus comentarios que tengan relación con el diccionario de palabras.

Una vez terminada la extracción de datos, se realiza una limpieza de los mismos, que consiste en eliminar datos mal extraídos como registros vacíos, reemplazar caracteres especiales, entre otros. En la **Figura 41** se observa el número de comentarios que se extrajeron de cada plataforma y también el número de entidades que se detectó después de aplicar el modelo NER general. Se puede observar que el número de entidades es mucho menor al número total de comentarios extraídos, lo que significa que en muy pocos comentarios hacen menciones de entidades que podrían ser posibles competidores tanto empresas como productos.

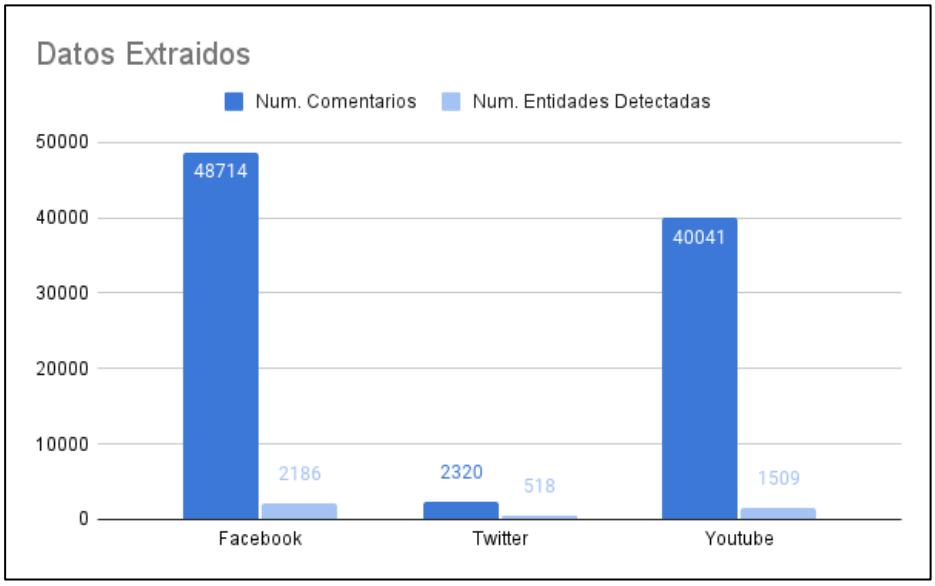


***Figura 41**: Número de datos extraídos de cada Plataforma.*
Fuente: Construcción propia

La **Figura 42** presenta un diagrama con los resultados del análisis de sentimiento, donde se visualiza el número de comentarios positivos, neutrales y negativos. Es importante mencionar que estos resultados son solo de los comentarios que tiene al menos una entidad detectada con el modelo NER. Se puede observar que Facebook es la plataforma con mayor número de datos desbalanceados, ya que, aproximadamente el 95% de los datos están con etiquetas de comentario neutral, los positivos y negativos existen muy pocos, pero con las otras plataformas no ocurre lo mismo, si existen número de datos semejante en cada grupo (positivo, neutral y negativo).

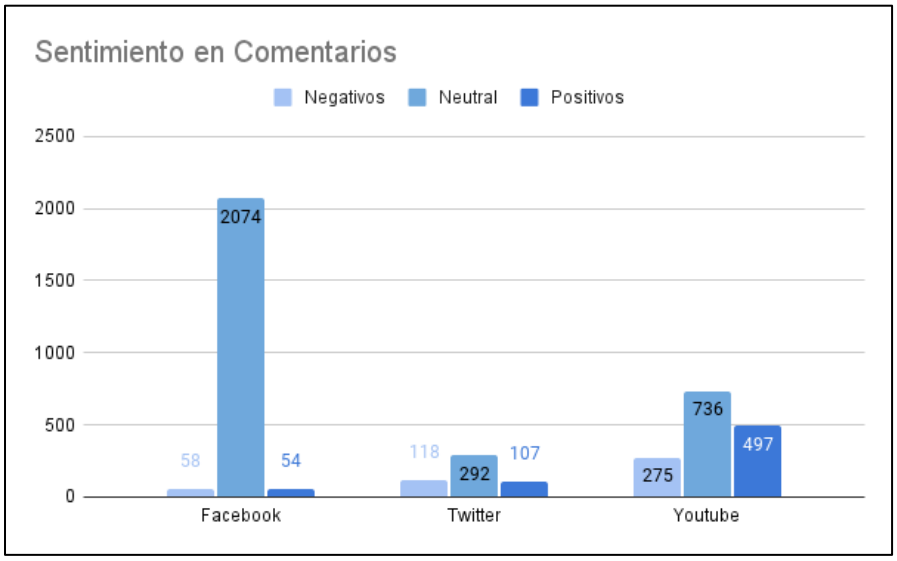


***Figura 42**: Detección de Sentimientos en comentarios.*
Fuente: Construcción propia

5.9 Desarrollo del Dashboard

El último proceso de esta investigación es el desarrollo del Dashboard para presentar los resultados y evaluar la utilidad del corpus desarrollado. Para este proceso se utilizó el corpus final desarrollado, con el cual se entrenó un modelo NER y un modelo para detección de textos comparativos que tienen las métricas de evaluación presentadas en la sección **5.7**. Con estos modelos entrenados se realiza la detección de entidades y de texto comparativo en los datos extraídos de las redes sociales. Luego se desarrolla el Dashboard en Power BI para presentar todos los resultados obtenidos. El Dashboard que se desarrolló se puede visualizar en el siguiente enlace.

https://app.powerbi.com/view?r=eyJrIjoiNjc2MzNiOTMtNjkyNC00MjA1LThmZjEtODhlMzkzOGNlODRkIiwidCI6IjhmNDY4MTZhLTcyMjAtNDg1MS04ZWYzLTY4MWI2MGM3ZmYwZiIsImMiOjR9&pageName=ReportSection32869605aed1a10ec905

A continuación, se presenta las capturas de pantalla de todas las opciones del Dashboard. En la **Figura 43** se presenta la pantalla de inicio que muestra una descripción del objetivo del Dashboard y las instrucciones para realizar la evaluación.

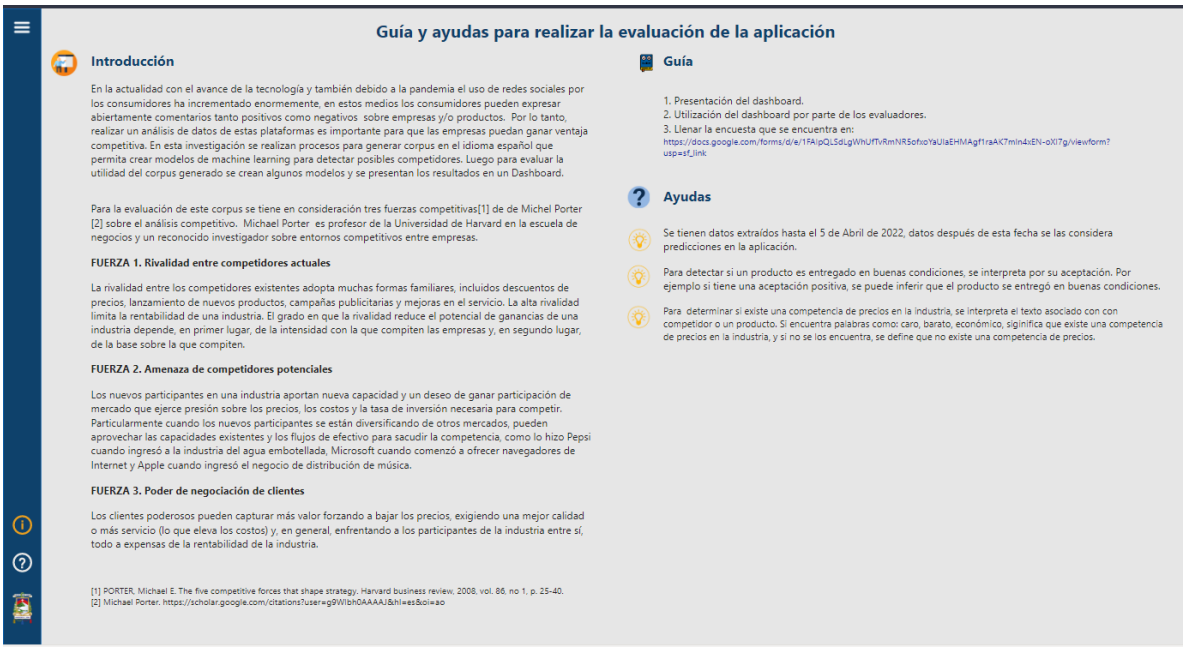


***Figura 43**: Página principal del Dashboard.*
Fuente: Construcción propia

En la **Figura 44** se presenta el menú que tiene el Dashboard, el cual cuenta con cuatro opciones que el usuario podrá navegar mientras utiliza la aplicación.

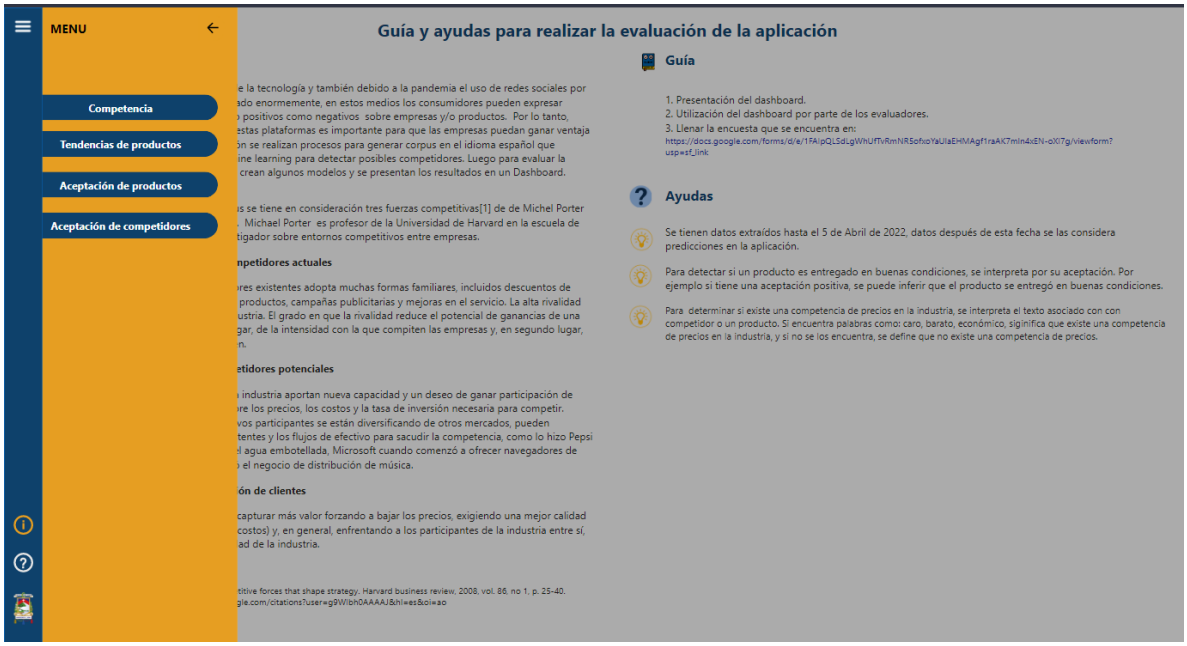


***Figura 44:** Menú disponible que tiene el Dashboard*
Fuente: Construcción propia

En la **Figura 45** se presenta las gráficas que permitirá hacer un análisis de los posibles competidores detectados con el modelo NER desarrollado en esta investigación que fue entrenado

con el corpus final y luego es aplicado a los datos de redes sociales. Esta pantalla permite también hacer un filtro para que el usuario pueda hacer un análisis de manera individual por posibles competidores o también hacer un análisis de datos de una única plataforma.

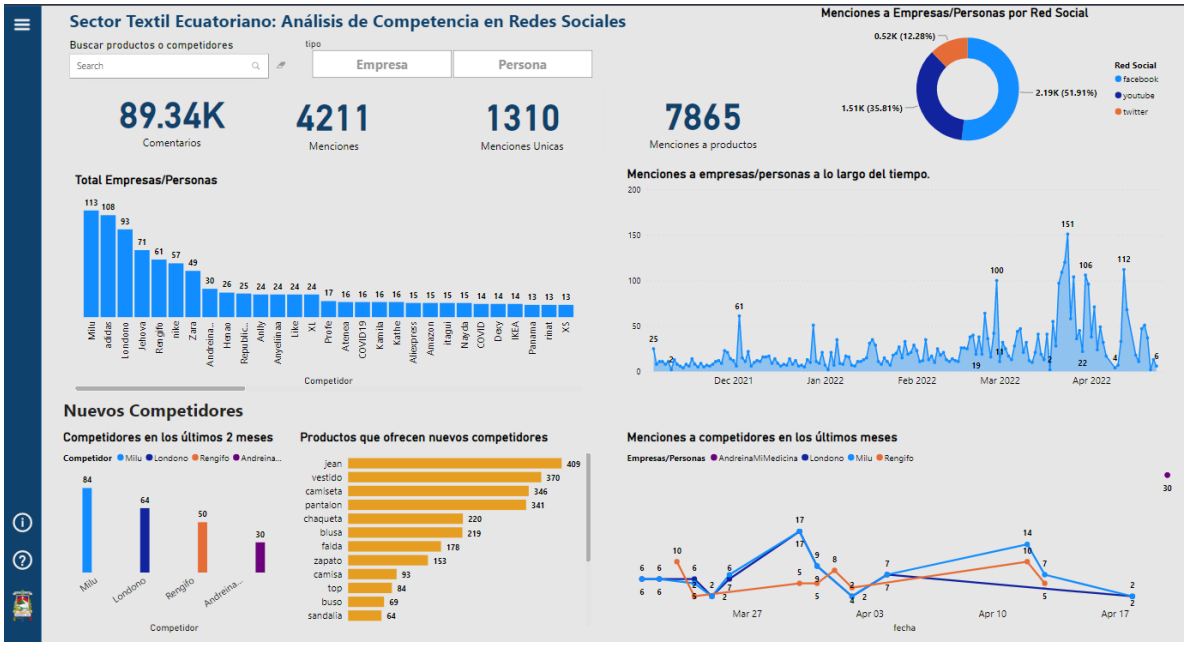


***Figura 45**: Análisis de posibles competidores.*
Fuente: Construcción propia

En la **Figura 46** se muestra la parte del Dashboard que permite hacer un análisis sobre tendencias de las menciones de un posible producto o competidor, además muestra una predicción de productos con un período de tiempo de tres meses. Es importante mencionar que si existen muy pocos datos la predicción es poco confiable ya que el algoritmo se basa en datos históricos. Esto se evidencia en el tamaño de las bandas de error que se muestra en el gráfico.

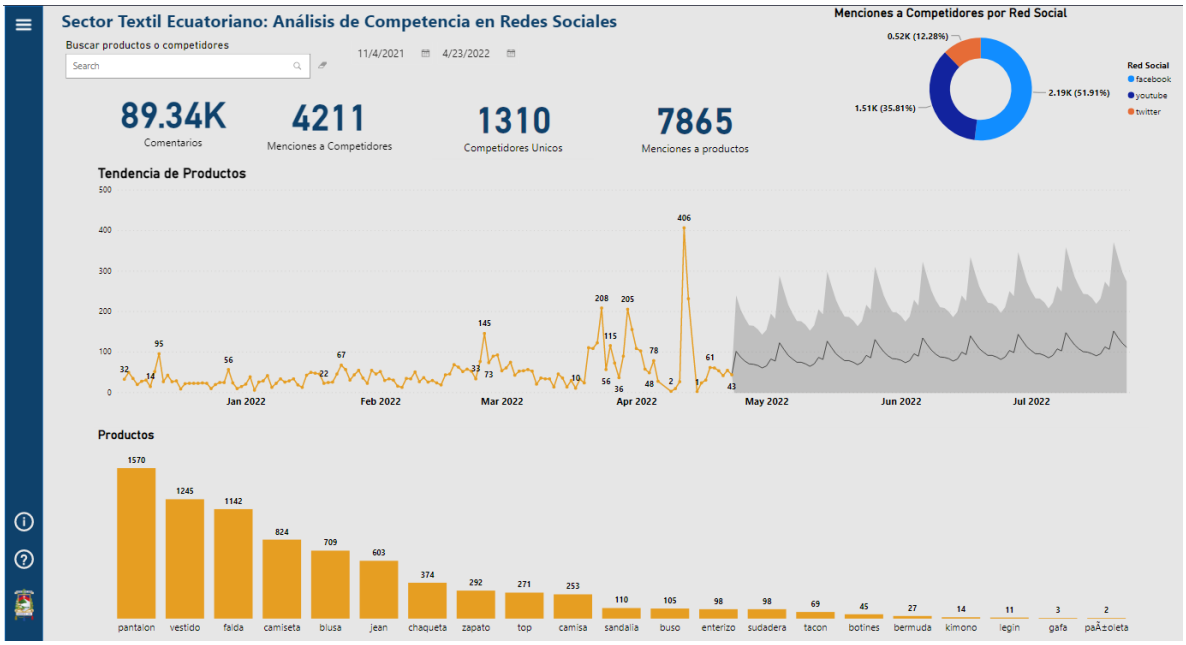


Figura 46*: Predicción de datos de las series de tiempo*.

Fuente: Construcción propia

En la **Figura 47** y **Figura 48** del Dashboard se tiene el análisis de sentimientos y la detección de adjetivos de los comentarios, esto permite observar la postura de los clientes hacia posibles competidores o productos, también tiene una nube de palabras con todos los adjetivos mencionados en los comentarios/tweets. Esto permite hacer un análisis de manera subjetiva de cada posible competidor como, por ejemplo: si un posible competidor tiene adjetivos como caro, bonito, etc., tiene más sentimientos positivos y neutrales que negativos, entonces significa que la empresa en análisis tiene una buena aceptación del cliente, pero también comercializa productos de alto valor.

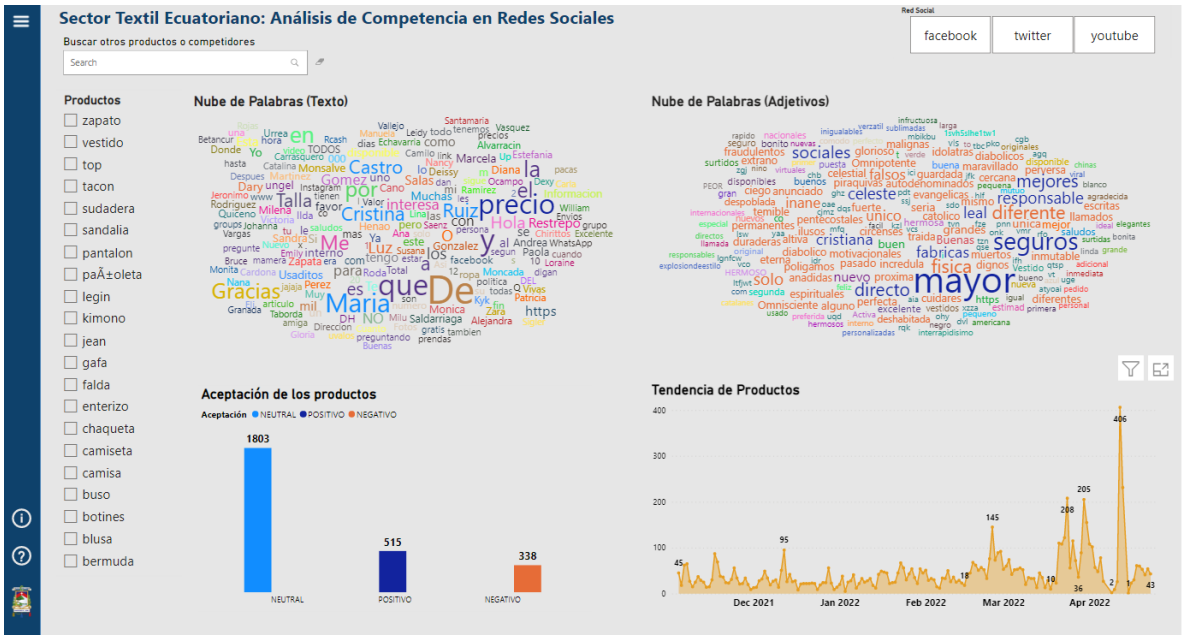


Figura 47: *Aceptación de Productos.*
Fuente: Construcción propia

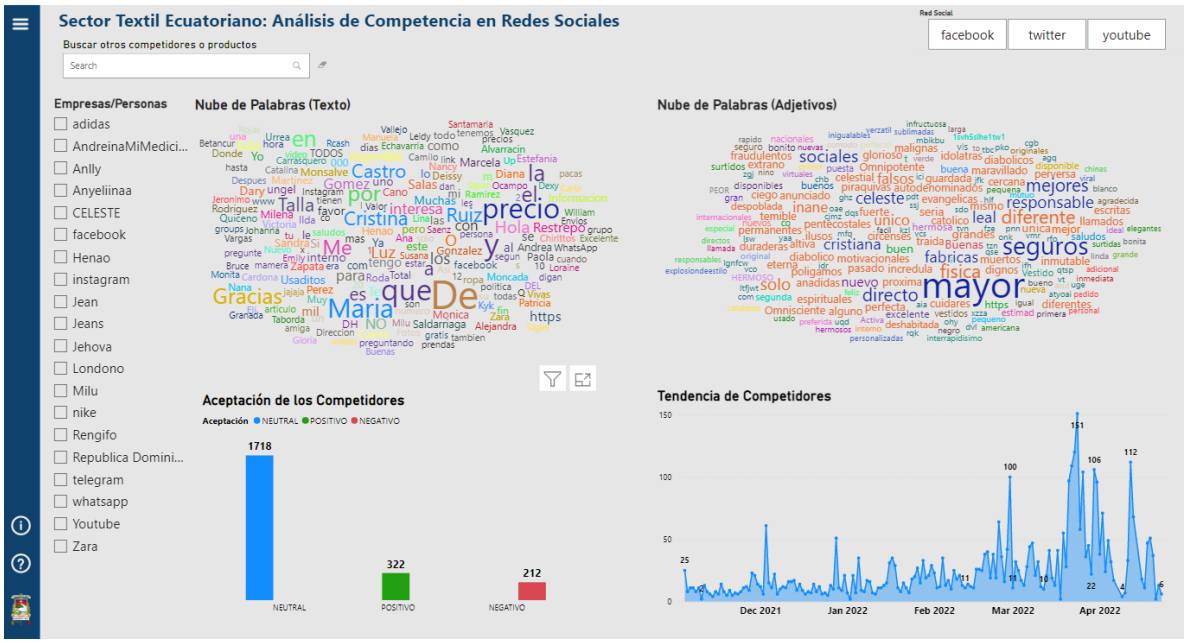


Figura 48*: Aceptación de Competidores.*
Fuente: Construcción propia

5.10 Evaluación empírica de la utilidad del Dashboard

En esta sección se realiza la evaluación empírica de la utilidad percibida de los modelos creados con el corpus generado en base a las fuerzas de Porter mediante el Dashboard desarrollado para el caso de estudio.

5.10.1 Objetivo de Evaluación

Según el paradigma Objetivo-Pregunta-Métrica (GQM – Goal Question Metric en inglés) planteado por (Basili, 1992), el objetivo de la evaluación empírica se muestra en la **Tabla 28**.

***Tabla 28**: Objetivo de la evaluación empírica*

Analizar	El Dashboard
Con el propósito de:	Evaluar la utilidad del corpus y los modelos creados.
Con respecto a:	Las fuerzas de Porter
Desde el punto de vista de:	Gerente de Pymes Textil
En el contexto de:	Usuarios utilizando el Dashboard

5.10.2 Preguntas de investigación

De acuerdo con el objetivo de evaluación, se plantean las siguientes preguntas de investigación (PI):

PI1: ¿El Dashboard creado mediante el corpus y los modelos generados es percibido con relación a la primera fuerza de Porter como útil?
PI2: ¿El Dashboard mediante el corpus y los modelos generados es percibido con relación a la cuarta fuerza de Porter como útil?
PI3: ¿El Dashboard creado mediante el corpus y los modelos generados es percibido con relación a la quinta fuerza de Porter como útil?

5.10.3 Hipótesis de investigación

Con el propósito de evaluar las preguntas de investigación se establecen las hipótesis de investigación mostradas en la **Tabla 29**. Las hipótesis nulas, que se representan por el subíndice cero, corresponden a la ausencia de un efecto de las variables independientes sobre las variables dependientes.

***Tabla 29**: Hipótesis de investigación para la evaluación*

Hipótesis	Descripción
$H1_0$	El Dashboard no es percibido como útil con relación a la primera fuerza de Porter $H1_0 \neg H1_1$
$H2_0$	El Dashboard no es percibido como útil con relación a la cuarta fuerza de Porter $H2_0 \neg H2_1$
$H3_0$	El Dashboard no es percibido como útil con relación a la quinta fuerza de Porter $H3_0 \neg H3_1$

La pregunta PI1 está soportada por la hipótesis: H1. La pregunta PI2 está soportada por la hipótesis: H2. La pregunta PI3 está soportada por la hipótesis H3.

5.10.4 Variables y métricas

Para este cuasiexperimento se considera como variable independiente el Dashboard creado mediante el corpus y los modelos generados. Por su parte se consideran como variables dependientes la primera, cuarta y quinta fuerza de Porter. Para medir estas variables dependientes se plantean una serie de preguntas definidas en la **Tabla 30**. Estas preguntas son medidas en una escala de 5 puntos de Likert.

Tabla 30: *Cuestionario para medir las variables dependientes*

Pregunta	Declaración
UF1.1	¿Puede determinar si nuevos competidores pueden ingresar a la industria?
UF1.2	¿Puede identificar los productos que ofrecen nuevos competidores?
UF1.3	¿Puede identificar el tipo de nuevos competidores que están ingresando a la industria (personas, empresas, etc.)?
UF4.1	¿Puede determinar si un competidor es aceptado por parte de los clientes?
UF4.2	¿Puede determinar si un producto es aceptado por parte de los clientes?
UF4.3	¿Puede determinar si un producto no es aceptado por parte de los clientes?
UF4.4	¿Puede determinar si un producto es entregado en buenas condiciones?
UF5.1	¿Puede determinar si existen muchos o pocos competidores en las redes sociales?
UF5.2	¿Puede determinar si existe una competencia de precios en la industria?

UF5.3	¿Puede determinar si los productos seguirán mostrando interés a los clientes en el futuro?
UF5.4	¿Puede determinar qué productos tienen una alta o baja demanda durante cierto período de tiempo?

5.10.5 Selección de la muestra

El cuasiexperimento se llevó a cabo en el mes de junio de 2022, con una muestra de 10 participantes. Los participantes son estudiantes de último semestre de la facultad de Ciencias Económicas de la Universidad de Cuenca.

5.10.6 Sesión cuasiexperimental y de capacitación

Primero se elaboró una presentación de capacitación para introducir a los participantes en el tema de este trabajo de titulación, en esta presentación se describe la inteligencia competitiva, las fuerzas de Porter en el contexto del sector textil y la utilización del Dashboard. Esta sesión se realizó de forma virtual, a través de una videoconferencia y tuvo una duración de 20 minutos.

Posteriormente se realizó la sesión experimental, en ella los participantes realizaron las tareas descritas en la **Tabla 31**. Para esto se elaboró una guía en Google Forms, la cual se dividió en las siguientes secciones:

I. Introducción de la inteligencia competitiva en el contexto de las redes sociales y las fuerzas de Porter.
II. Desarrollo del cuasiexperimento: en esta sección los participantes siguieron una serie de pasos para cumplir con las tareas planteadas.
III. Explicación de la escala de Likert y como llenar el cuestionario.
IV. Cuestionario: con el objetivo de evaluar la utilidad de los modelos creados con el corpus generado, los participantes completaron el cuestionario.

Al finalizar el cuasiexperimento se recolectó los datos necesarios para analizar las variables dependientes. Los resultados del cuestionario se recolectaron con la ayuda de la herramienta Google Forms.

Tabla 31: *Tareas del cuasiexperimento*

Tarea	Subtarea	Descripción
1		Ingrese a la Dashboard
2		Clic en esquina inferior izquierda
3		Leer la guía presentada.
4		Ingresar a Competencias
	4.1	Interactuar con la página por 3 minutos
5		Ingresar a Tendencias de Productos
	5.1	Interactuar con la página por 3 minutos
6		Ingresar a aceptación de productos
	6.1	Interactuar con la página por 3 minutos
7		Ingresar a aceptación de competidores
	7.1	Interactuar con la página por 3 minutos
8		Llenar la encuesta.

5.10.7 Validez del cuestionario

Para llevar a cabo la validación de este cuestionario realizado con los evaluadores, se implementa la técnica del coeficiente de Alfa de Cronbach. Se utiliza la fórmula **(11)** y se obtiene el siguiente valor (**Tabla 32**).

Tabla 32: *Estadística de fiabilidad.*

Nombre	Resultado
Alfa de Cronbach	0.79

El valor de alfa presentado en **Tabla 32** es aceptable según (Streiner, 2003) ya que es mayor a 0.70 qué es lo recomendable como mínimo. Este valor de 0.79 indica que el cuestionario tiene una consistencia interna alta, y como no es un valor mayor a 0.90 se puede afirmar que no hay redundancia o duplicación de preguntas.

5.10.8 Análisis e interpretación de resultados

En esta sección se efectúa el análisis de los resultados obtenidos respecto al rendimiento y las percepciones del participante con respecto a las variables de estudio. El análisis se realizó con las librerías pandas, matplotlib y statsmodel del lenguaje de programación Python.

Las preguntas del cuestionario utilizado se agrupan de acuerdo con las variables dependientes planteadas en la sección **5.10.4** y se describen en la **Tabla 33**.

Tabla 33: *Variables dependientes para la evaluación*

Variable	Descripción
UF1	Grado en el cual los participantes creen que el Dashboard creado en el caso de estudio mediante el corpus es útil al momento de realizar un análisis de la competencia en redes sociales en relación a la primera fuerza de Porter.
UF4	Grado en el cual los participantes creen que el Dashboard creado en el caso de estudio mediante el corpus es útil al momento de realizar un análisis de la competencia en redes sociales en relación a la cuarta fuerza de Porter.
UF5	Grado en el cual los participantes creen que el Dashboard creado en el caso de estudio mediante el corpus es útil al momento de realizar un análisis de la competencia en redes sociales en relación a la quinta fuerza de Porter.

La **Figura 49** presenta el diagrama de caja y bigotes para cada una de las variables dependientes, las cuales se miden con una escala de Likert de cinco puntos (1-5) donde 1 representa el valor más bajo, 3 representa el valor neutro y 5 el valor más alto.

Debido a que la muestra de este cuasiexperimento es de n<50, se utiliza la prueba de normalidad de Shapiro-Wilk. La **Tabla 34** presenta un resumen de los valores estadísticos y de la prueba de Shapiro-Wilk para cada una de las variables. Cuando la muestra tiene un valor de p<0.05, significa que la muestra no tiene una distribución normal, por lo tanto, se aplica la prueba de Wilcoxon para determinar su significancia. Caso contrario, con p>0.05, la muestra tiene una distribución normal y se aplica la prueba T-Student. Tanto para la prueba de Wilcoxon como para T-Student, cuando el valor de significancia es mayor que 0.05 se dice que la hipótesis es aceptada. Caso contrario, cuando la significancia es menor que 0.05, la hipótesis es rechazada.

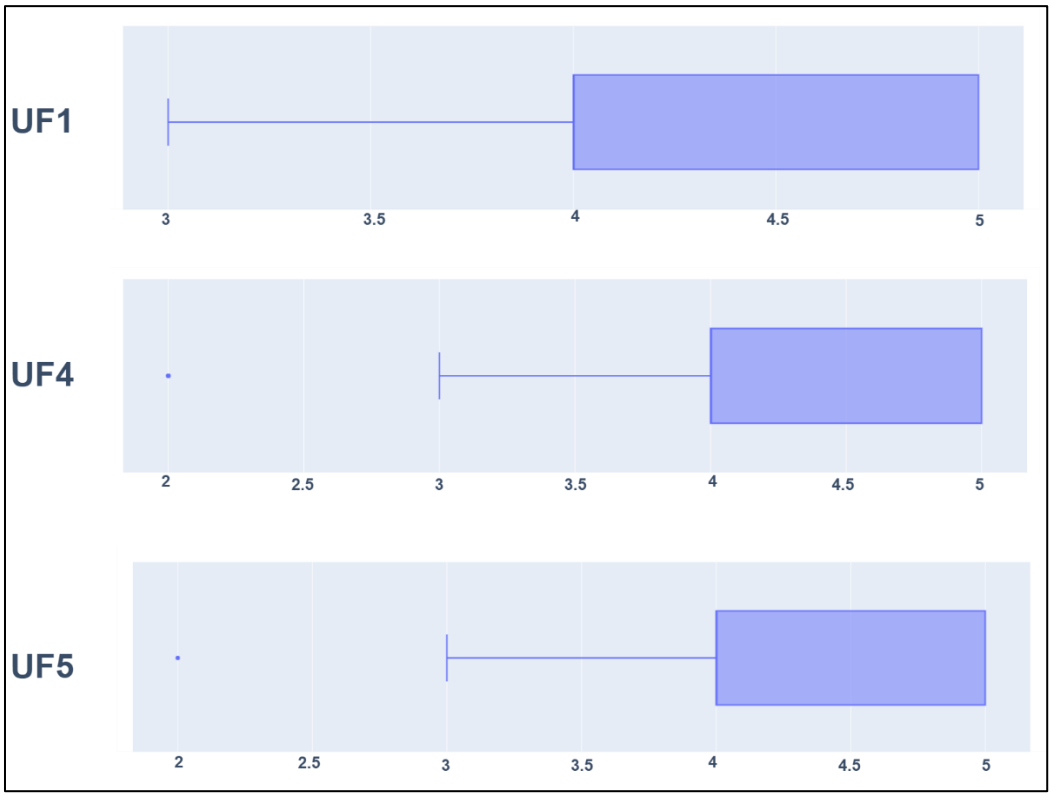


Figura 49: *Diagrama de caja y bigotes de las variables dependientes.*
Fuente: Construcción propia

Tabla 34: *Estadística descriptiva para las variables dependientes correspondientes a la percepción de los participantes*

Variable	Min	Max	Media	Mediana	Desviación estándar	Shapiro-Wilk: p
UF1	3	5	4.13	4	0.68	0.00006
UF4	3	5	4.35	4	0.63	0.000002
UF5	3	5	4.25	4	0.67	0.000004

Los valores de significancia obtenidos para cada una de las variables se muestran en la **Tabla 35**. Con base en esto, se puede concluir que las hipótesis $H1_0$, $H2_0$ y $H3_0$ son rechazadas, por lo tanto:

I. El Dashboard es percibido como útil con relación a la primera fuerza de Porter.
II. El Dashboard es percibido como útil con relación a la cuarta fuerza de Porter.
III. El Dashboard es percibido como útil con relación a la quinta fuerza de Porter.

Estos resultados indican que existe una probabilidad alta de que los modelos creados con el corpus generado sean utilizables en la práctica.

Tabla 35: *Significancias para las variables dependientes*

Variable	Prueba aplicada	Significancia
UF1	Wilcoxon	0.000001
UF4	Wilcoxon	0.00000005
UF5	Wilcoxon	0.00000002

Los resultados obtenidos en el cuasiexperimento se resumen en la **Tabla 36**. Con base en estos resultados, se responden las preguntas de investigación planteadas a inicio del cuasiexperimento:

- **PI1:** ***¿El Dashboard creado utilizando los modelos mediante el corpus generado es percibido con relación a la primera fuerza de Porter como útil?***

El Dashboard es percibido como útil con relación a la primera fuerza de Porter, esto se determina a través del rechazo de la hipótesis nula $\boldsymbol{H1_0}$. Por lo tanto, se puede determinar que los modelos creados con el corpus desarrollado en este trabajo son útiles con relación a la primera fuerza de Porter.

- **PI2:** ***¿El Dashboard creado utilizando los modelos mediante el corpus generado es percibido con relación a la cuarta fuerza de Porter como útil?***

El Dashboard es percibido como útil con relación a la cuarta fuerza de Porter, esto se determina a través del rechazo de la hipótesis nula $\boldsymbol{H2_0}$. Por lo tanto, se puede determinar que los modelos creados con el corpus desarrollado en este trabajo son útiles con relación a la cuarta fuerza de Porter.

- **PI3: ¿El Dashboard creado utilizando los modelos mediante el corpus generado es percibido con relación a la quinta fuerza de Porter como útil?**

El Dashboard es percibido como útil con relación a la quinta fuerza de Porter, esto se determina a través del rechazo de la hipótesis nula $\boldsymbol{H3_0}$. Por lo tanto, se puede determinar que los modelos creados con el corpus desarrollado en este trabajo son útiles con relación a la quinta fuerza de Porter.

Tabla 36: *Resumen de resultados del cuasiexperimento*

Pregunta de Investigación	Hipótesis	Significancia	Acción	Resultado
PI1	$H1_0$	$p < 0.001$	Rechazada	El Dashboard es percibido como útil con relación a la primera fuerza de Porter.
PI2	$H2_0$	$p < 0.001$	Rechazada	El Dashboard es percibido como útil con relación a la cuarta fuerza de Porter
PI3	$H3_0$	$p < 0.001$	Rechazada	El Dashboard es percibido como útil con relación a la quinta fuerza de Porter.

5.11 Presentación de los resultados

En esta subsección se presentan los resultados de cada una de las preguntas que conforman el cuestionario completado por los participantes al finalizar el cuasiexperimento. La primera pregunta: ¿Puede determinar si nuevos competidores pueden ingresar a la industria? La **Figura 50** indica que el 90% de los participantes están de acuerdo en que pueden determinar la cantidad de competidores utilizando el Dashboard, mientras que el 10% muestra una postura neutral.

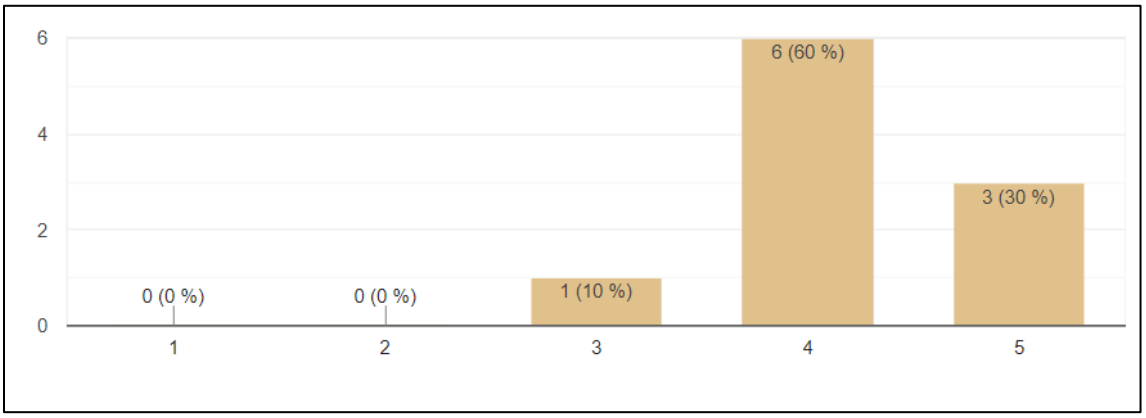


***Figura 50**: Cuestionario - pregunta 1.*
Fuente: Construcción propia

Similar a la pregunta 1, la segunda pregunta: ¿Puede identificar los productos que ofrecen nuevos competidores? La **Figura 51** indica que el 90% de los participantes están de acuerdo en que pueden identificar los productos que ofrecen nuevos competidores utilizando el Dashboard, mientras que el 10% muestra una postura neutral.

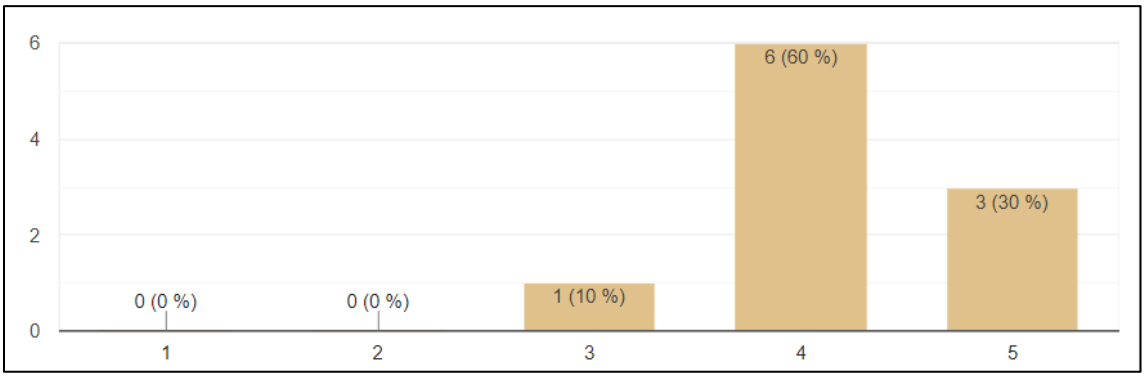


Figura 51: *Cuestionario - pregunta 2.*
Fuente: Construcción propia

La **Figura 52**, con el enunciado: ¿Puede identificar el tipo de nuevos competidores que están ingresando a la industria (personas, empresas, etc.) ?, indica que el 70% de los participantes están de acuerdo en ser capaces de identificar el tipo de nuevos competidores que están ingresando a la industria utilizando el Dashboard, mientras que el 30% muestra una postura neutral.

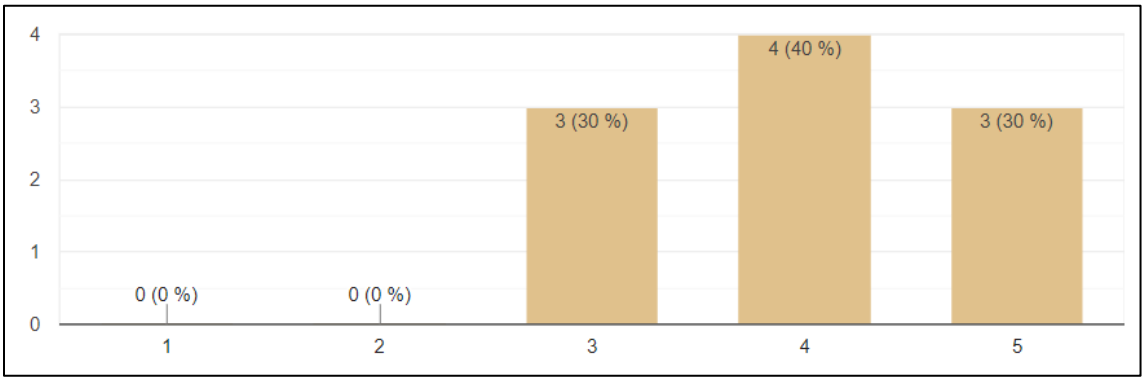


Figura 52*: Cuestionario - pregunta 3.*
Fuente: Construcción propia

La **Figura 53**, con el enunciado: ¿Puede determinar si un competidor es aceptado o no por parte de los clientes?, indica que el 90% de los participantes están de acuerdo en ser capaces de identificar si un competidor es aceptado por parte de los clientes utilizando el Dashboard, mientras que el 10% no está de acuerdo.

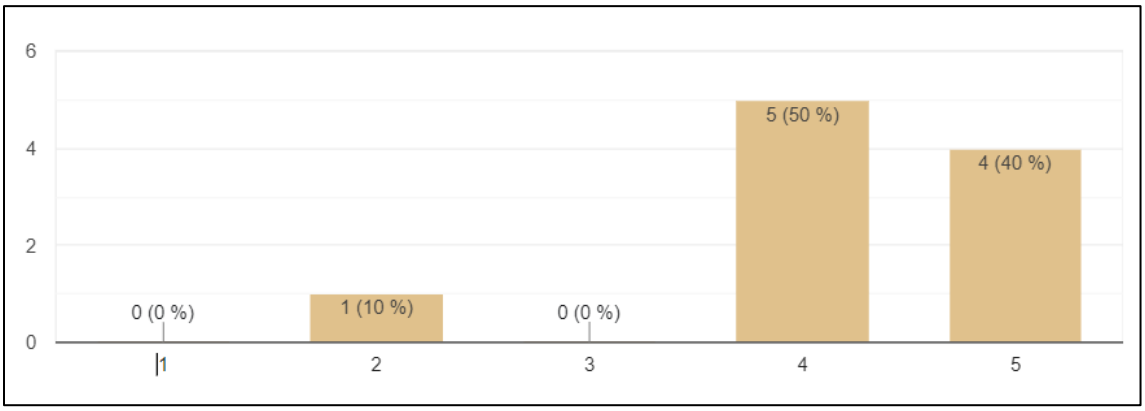


Figura 53: *Cuestionario - pregunta 4.*
Fuente: Construcción propia

La **Figura 54**, con el enunciado: ¿Puede determinar si un producto es aceptado por parte de los clientes?, indica que el 100% de los participantes están de acuerdo en ser capaces de identificar determinar si un producto es aceptado por parte de los clientes utilizando el Dashboard.

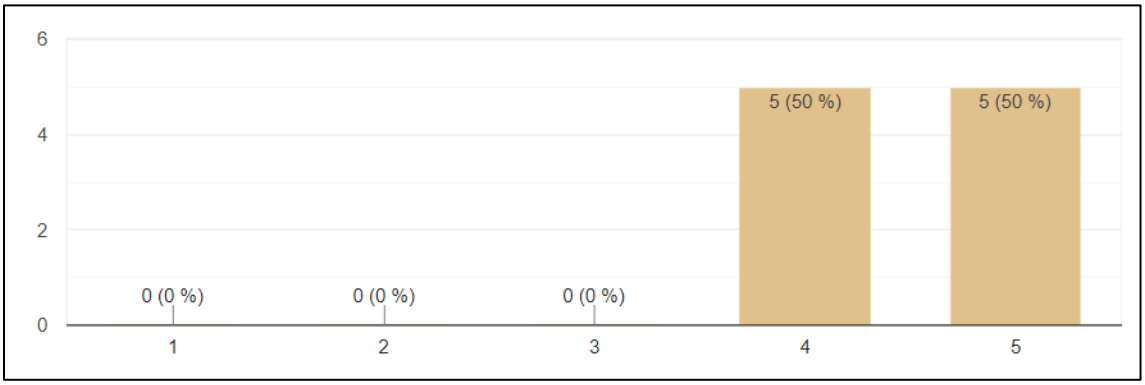


Figura 54: *Cuestionario - pregunta 5.*
Fuente: Construcción propia

La **Figura 55**, con el enunciado: ¿Puede determinar si un producto no es aceptado por parte de los clientes?, indica que el 80% de los participantes están de acuerdo en ser capaces de identificar si un producto no es aceptado por parte de los clientes utilizando el Dashboard, mientras que el 20% muestra una postura neutral.

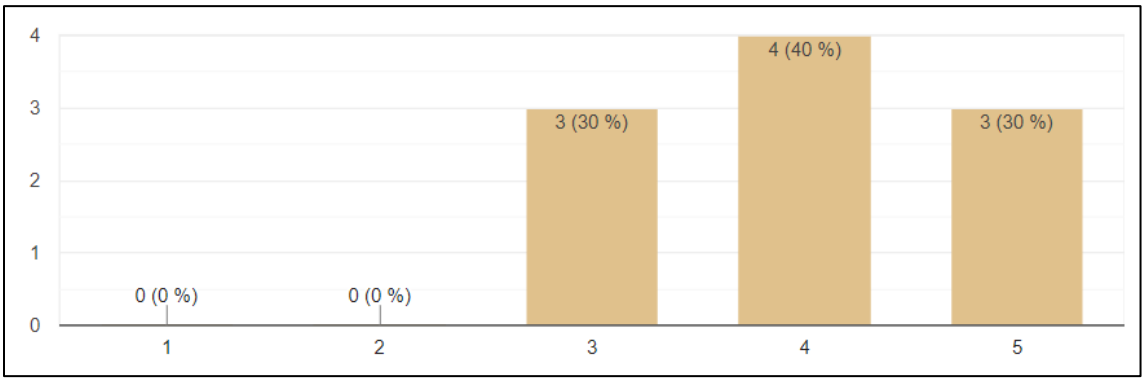


***Figura 55**: Cuestionario - pregunta 6.*
Fuente: Construcción propia

La **Figura 56**, con el enunciado: ¿Puede determinar si un producto es entregado en buenas condiciones?, indica que el 70% de los participantes están de acuerdo en ser capaces de identificar si un producto es entregado en buenas condiciones utilizando el Dashboard, mientras que el 10% muestra una postura neutral y el 20% no está de acuerdo.

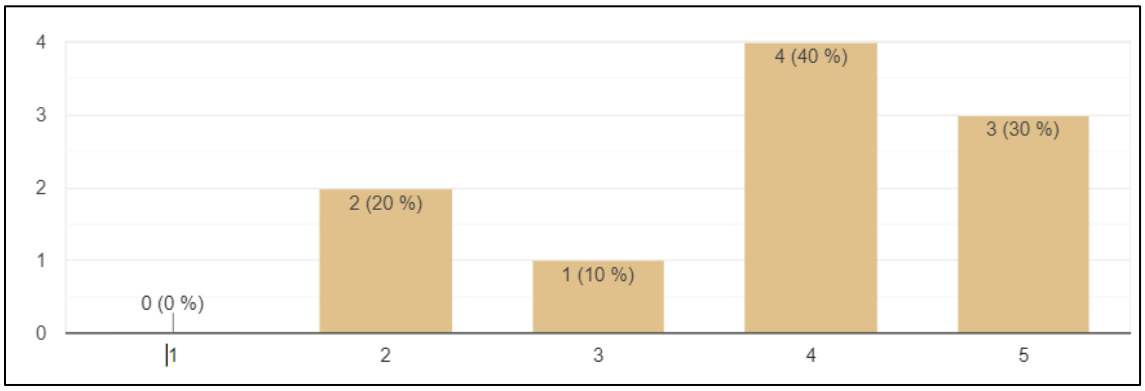


***Figura 56:** Cuestionario - pregunta 7. Elaboración propia*

La **Figura 57**, con el enunciado: ¿Puede determinar si existen muchos o pocos competidores en las redes sociales ?, indica que el 100% de los participantes están de acuerdo en ser capaces de identificar la cantidad de competidores utilizando el Dashboard.

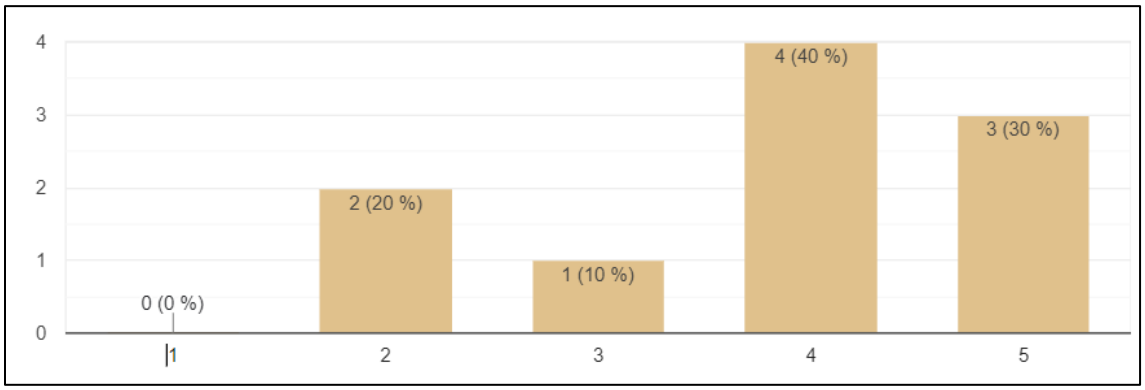


Figura 57: *Cuestionario - pregunta 8.*
Fuente: Construcción propia

La **Figura 58**, con el enunciado: ¿Puede determinar si existe una competencia de precios en la industria ?, indica que el 80% de los participantes están de acuerdo en ser capaces de determinar si existe una competencia de precios en la industria utilizando el Dashboard, mientras que el 10% muestra una postura neutral y el 10% no está de acuerdo.

.

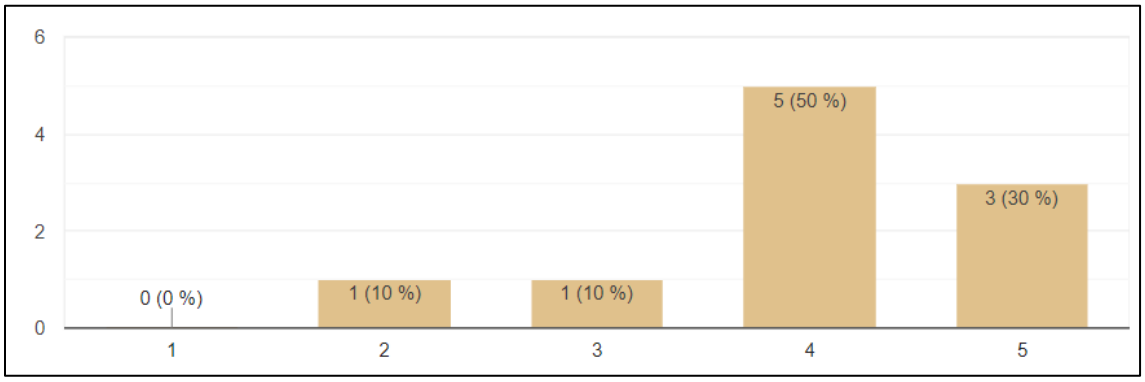


Figura 58*: Cuestionario - pregunta 9.*
Fuente: Construcción propia

La **Figura 59**, con el enunciado: ¿Puede determinar si los productos seguirán mostrando interés a los clientes en el futuro ?, indica que el 70% de los participantes están de acuerdo en ser capaces de determinar si los productos seguirán mostrando interés a los clientes en el futuro utilizando el Dashboard, mientras que el 30% muestra una postura neutral.

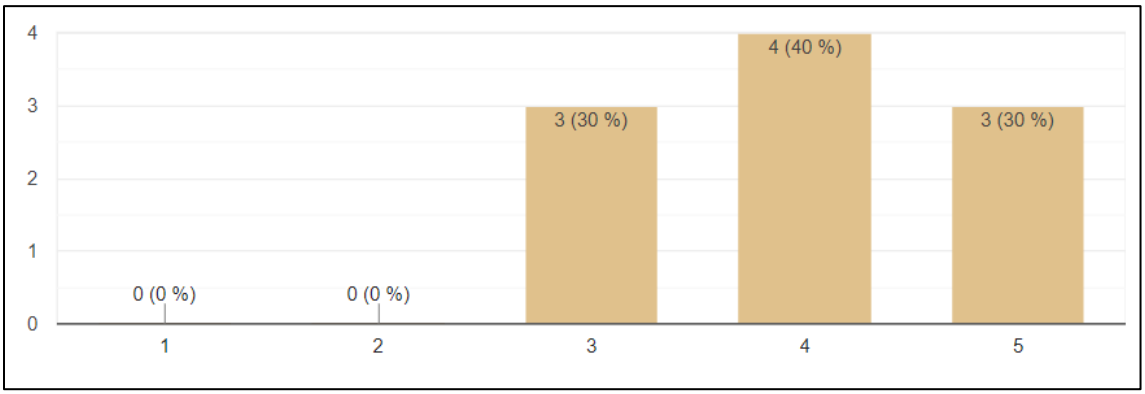


***Figura 59**: Cuestionario - pregunta 10.*
Fuente: Construcción propia

La **Figura 60**, con el enunciado: ¿Puede determinar que productos tienen una alta o baja demanda durante cierto período de tiempo ?, indica que el 90% de los participantes están de acuerdo en ser capaces de determinar qué productos tienen una alta o baja demanda durante cierto período de tiempo utilizando el Dashboard, mientras que el 10% muestra una postura neutral.

.

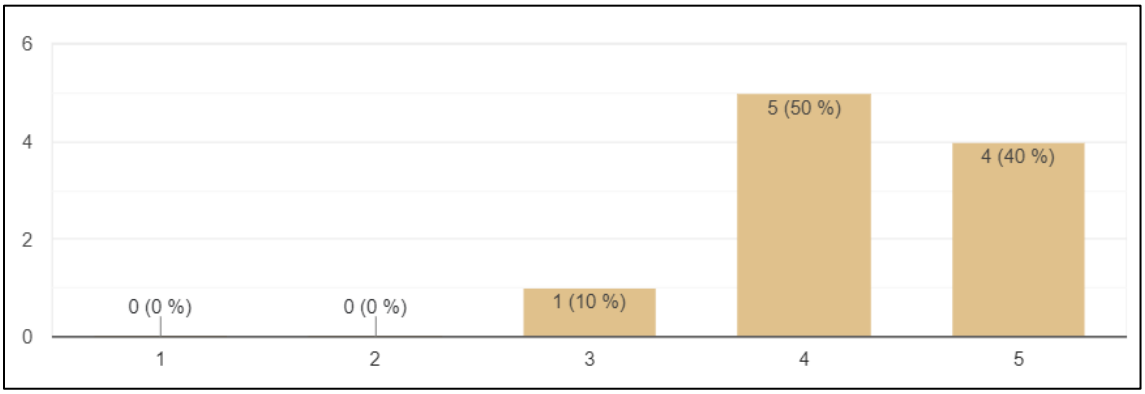


***Figura 60**: Cuestionario - pregunta 11.*
Fuente: Construcción propia

5.12 Comparación de la utilidad en relación a las Fuerzas de Porter

Con el fin de demostrar si hay una diferencia de la utilidad del corpus y modelos generados en relación a las fuerzas evaluadas, se generó un score agrupando las respuestas con cada variable que a su vez representa una de las fuerzas de Porter (**sección 5.10.4**). El score generado es sobre 100 puntos, donde 100 representa el valor en que todos los evaluadores hubiesen respondido con 5 las preguntas del cuestionario. La utilidad en relación a las fuerzas de Porter se presenta en la **tabla 37**.

Tabla 37: *Resumen de la utilidad en relación a las Fuerzas de Porter*

Fuerza de Porter	Score de utilidad
Fuerza 1: Amenaza de competidores potenciales	82.6
Fuerza 4: Poder de negociación de clientes	83.5
Fuerza 5: Rivalidad entre competidores actuales	84

La **Tabla *37*** muestra que existe una relación muy cercana del score de utilidad, donde se tiene un valor de utilidad más alto en la fuerza 5, seguido de la fuerza 4 y por último la fuerza 1. Se puede concluir que, en las tres fuerzas evaluadas, se obtuvo un valor alto en el score, lo que respalda la utilidad del corpus y modelos generados.

5.13 Amenazas a la validez

En esta sección se describen los problemas que pueden afectar la validez del cuasi-experimento. Para ello, se consideran los cuatro tipos de validez propuestos por Hyman, 92010) los cuales son: validez de la conclusión estadística, validez interna, validez de constructo y validez externa.

Validez de la conclusión estadística:

Los métodos estadísticos seleccionados para el análisis pueden afectar la validez de la conclusión. Para hacer frente a esto, se utilizó la prueba de Shapiro-Wilk con la cual se determinó que la muestra tiene una distribución normal. Posteriormente se usó la prueba Wilcoxon a través de la cual se calculó la significancia de las variables dependientes; y se determinó que las hipótesis nulas debían ser rechazadas.

Validez interna:

La validez interna puede verse afectada por el conocimiento y experiencia que tienen los participantes. Para este cuasi-experimento ninguno de los participantes había usado el Dashboard con anterioridad. Para equilibrar el conocimiento de todos los participantes, se realizó la sesión de capacitación antes de la ejecución del cuasi-experimento.

Validez del constructo:

La amenaza identificada para este tipo de validez hace referencia a la confiabilidad del cuestionario utilizado para evaluar las percepciones de los usuarios. En este sentido, se validó el cuestionario comprobando su fiabilidad y que no haya preguntas repetidas o redundantes mediante la prueba de alfa de Cronbach.

Validez externa:

Este tipo de validez hace referencia a la capacidad de generalizar los resultados del cuasiexperimento a toda la población. Para ello, se ha definido una guía sencilla que muestra paso a paso las tareas a realizar en el cuasiexperimento. Al evaluarse en un caso de estudio (sector textil), no se debe generalizar a otros contextos. Se propone como trabajo futuro replicar este trabajo para analizar contextos diferentes.

CAPITULO 6: CONCLUSIONES Y TRABAJO FUTURO

6.1 Conclusiones

El objetivo de este trabajo de investigación consistió en la creación de un corpus en el idioma español que sea de utilidad con el fin de entrenar modelos de aprendizaje automático que permita realizar detección de competidores con datos de redes sociales para empresas PYMEs del sector Textil. Para esto se desarrolló una metodología que describe detalladamente cada una de las etapas que se siguieron. Adicionalmente, para validar la utilidad de este corpus creado, se han entrenado algoritmos con el corpus desarrollado y se ha creado un Dashboard con Power BI para presentar los resultados obtenidos a un grupo de evaluadores. De la misma manera este proceso se ha descrito detalladamente en la metodología, cumpliendo todos los objetivos específicos de la investigación. También en esta investigación se planteó evaluar el corpus en un caso de estudio considerando empresas textiles pertenecientes al proyecto SUMA, sin embargo, por motivos logísticos, se evaluó considerando todo el sector textil ecuatoriano.

El proceso inicial para esta investigación radicó en la búsqueda de corpus en el idioma español que sirvan como base para el desarrollo de esta propuesta metodológica, según la literatura no existen corpus etiquetados para hacer detección de competidores en el idioma español, de manera que, se buscó algunos corpus que tengan comentarios de redes sociales y/o plataformas de comercio electrónico, donde se tuvo que agregar manualmente las etiquetas necesarias. Este proceso de etiquetado si bien no es complicado, es altamente laborioso y demanda mucho tiempo, por lo que en este trabajo las etiquetas agregadas manualmente y después automáticamente con los algoritmos generados son de gran aporte a la comunidad científica, para que en futuras investigaciones no se tenga que realizar este etiquetado.

Otro aspecto importante en esta investigación consistió en entrenar un modelo robusto para la detección de entidades nombradas, que sirve para realizar la detección de los posibles competidores. Se utilizó la librería SpaCy para este propósito, la cual dispone de modelos pre-entrenados que se pueden utilizar. El modelo que se utilizó en esta investigación (es_core_news_lg) en su descripción menciona que tanto en entrenamiento y prueba todas sus métricas de evaluación (recall, precisión, f-score) tienen el 90%, pero al utilizar con los corpus de esta investigación dichas métricas de evaluación no llegaban a más del 40%, por lo tanto, fue significativo volver a entrenar el modelo proporcionándole más datos de entrenamiento y prueba, con este proceso se logró mejorar y se obtuvo en el modelo final métricas de evaluación de 78.4 % en precisión, 78% en recall y 78.2% en f-score.

No basta con la creación de una metodología para construir un corpus de estas características, también se debe tener una alta capacidad computacional para la ejecución de los algoritmos desarrollados para que puedan realizarse en tiempos prácticos para la investigación. Características como CPU y/o GPU de alta prestaciones son necesarios, sin embargo, no es necesario tener físicamente estos dispositivos, en la actualidad existen alternativas que se pueden usar, en esta investigación se utilizó Kaggle y Google Colab que ofrecen CPU y GPU de alta prestaciones para ejecutar los algoritmos de NER, a pesar de eso, al ejecutar algunas pruebas del algoritmo NER con

GPU, como por ejemplo, en el caso de MOZETIC corpus que tiene 131388 registros se demoró más de un día.

Por otra parte, para obtener un modelo robusto que identifique opiniones comparativas, fue importante realizar varios ciclos de entrenamiento y prueba, además al ser un problema de clases desbalanceadas fue clave usar ROC ponderado y F-score para validar el modelo.

El mejor modelo que identifica opiniones comparativas fue Naive Bayes, validado por el ROC ponderado, que en clases desbalanceadas es una excelente métrica de evaluación (Weng, C. G., & Poon, J. 2008), y también validado en la literatura, ya que es usado con frecuencia por parte de los investigadores en la minería de opiniones comparativas (Varathan, K. et al., 2017).

Otro aspecto importante en esta investigación fue la evaluación del corpus desarrollado. Para ello, primero se realizó un ranking de plataformas digitales que sean de utilidad para el análisis de mercado, con lo cual se realizó un artículo científico al respecto y se lo presentó en TIC EC 2021 (Fajardo Cárdenas et al., 2021) (**Anexo 2**); luego se extrajeron datos reales de las redes sociales mejor rankeadas, siendo Facebook la plataforma más importante para el análisis debido a su popularidad y auge en el país, pero al mismo tiempo, es una de las más complicadas al momento de la extracción de datos debido a su fuerte limitación en la API, por lo tanto, se utilizó una librería de Web Scraping. En esta tarea se encontraron dos problemas cuya solución requiere de arduo trabajo y tiempo; (1) el baneo de la dirección IP, para lo cual se utilizó una VPN para solventar este problema. (2) la obtención de fecha de cada comentario, esto se arregló parcialmente cambiado el idioma del perfil del Facebook que se estaba utilizando, hasta el final se tuvo este inconveniente, pero después de implementar la solución mencionada fueron muy pocos los registros sin fecha, aproximadamente un 0.2% del total de datos. Otro aspecto importante de esta plataforma es que cinco grupos que se seleccionaron al inicio del proyecto (finales de diciembre de 2021) para extraer datos, al finalizar la extracción, ya no existían. Con respecto a Twitter y YouTube no se tuvo problemas ya que se utilizó la API oficial de cada plataforma.

Se pudo determinar que cada corpus que actualmente existe para tareas de minería de textos, individualmente tiene sus limitaciones ya sea de tamaño o de relevancia, por lo que la identificación de los textos más importantes para la Inteligencia Competitiva de cada corpus y la unión de los mismos en uno solo junto con sus etiquetas comparativas y de entidades, genera un valor agregado para futuras investigaciones ya que permite a las empresas generar modelos de aprendizaje automático para resolver tareas de Inteligencia Competitiva que todavía no se han resuelto en el lenguaje español, como por ejemplo: i) rastrear efectivamente a los competidores y su comportamiento, ii) identificar fortalezas y debilidades de competidores..

El corpus generado es útil en la práctica para realizar un análisis de competitividad en redes sociales, esto validado por una evaluación empírica aplicada en el sector textil y en base a las fuerzas de Porter, lo cual brinda un respaldo de los resultados obtenidos. Así mismo, se pudo determinar en base a un score de utilidad de 100 puntos, que la utilidad es alta para las tres fuerzas de Porter,

donde el corpus tiene un score de utilidad en relación a la fuerza 5 de 84, en relación a la fuerza 4 de Porter 83.5 y por último 82.6 en relación a la primera fuerza de Porter.

Por último, es importante mencionar que el análisis realizado en este trabajo es pionero y sirve de base para futuras investigaciones relacionadas con la inteligencia competitiva, específicamente en la detección de competidores en el lenguaje español, donde la CI estaba estrictamente restringida por la falta de un corpus, además de estar validado su utilidad con una evaluación empírica. En este trabajo se demostró que, con el corpus generado, las empresas pueden atravesar por las cinco fases del proceso de Inteligencia Competitiva, desde la planificación hasta su evaluación, especializándose en las fases de recolección de datos y análisis de datos, que es donde se presentaban los mayores problemas de la Inteligencia Competitiva.

6.2 Trabajo Futuro

Al ser un trabajo pionero, existen algunos puntos que podrían mejorarse para futuras investigaciones en el campo de la Inteligencia Competitiva en el lenguaje español. A continuación, se plantean algunas de ellas:

En este trabajo se realizó un etiquetado manual para determinar si un texto es comparativo o no. Si bien los resultados obtenidos fueron validados, se podría mejorar y validar la precisión del etiquetado de datos realizándolo con personas expertas en literatura.

En el Dashboard presentado se puede evidenciar que existen nombres que no tienen relación con nombres de posible competidores o productos como, por ejemplo: KS, XL, entre otros., esto se debe a que el modelo de detección de entidades nombradas tiene métricas de evaluación tanto para el f-score, precisión y recall de 78.19%, 78.37% y 78% respectivamente, estos resultados se puede mejorar al tener más datos de entrenamiento relacionados al sector, lo cual, sería interesante realizar en trabajos futuros.

Al momento de extraer datos de Facebook con la librería Facebook-scrapper existen varios datos que se extrajeron sin la fecha de su publicación, por lo tanto, aquellos registros se eliminaron en esta investigación debido a que la fecha es un campo importante para hacer el análisis final, entonces como trabajo futuro se debe mejorar o buscar alternativas para extraer datos de Facebook.

Este trabajo de investigación se apoyó en el proyecto SUMA, sin embargo, la evaluación del corpus desarrollado, por diferentes motivos fuera de nuestro alcance se realizó con estudiantes de último semestre de la facultad de economía de la Universidad de Cuenca. Como trabajo futuro se deberá realizar la evaluación con personas que estén directamente involucradas con empresas del sector textil como por ejemplo el proyecto SUMA y así obtener una validación más directa sobre la utilidad del corpus desarrollado.

Por último, en este trabajo realizó la evaluación de la utilidad del corpus en el sector textil, para garantizar su generalización, se propone evaluar la utilidad del corpus en otros contextos.

BIBLIOGRAFÍA

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, *20*(2), 105–123. https://doi.org/10.1007/S10502-019-09325-9/TABLES/1

Agrawal, P., & Trivedi, B. (2020, November 6). Evaluating Machine Learning Classifiers to detect Android Malware. *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*. https://doi.org/10.1109/INOCON50539.2020.9298290

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. *13*. https://doi.org/10.48550/arxiv.1707.02919

Amarouche, K., Benbrahim, H., & Kassou, I. (2015). Product Opinion Mining for Competitive Intelligence. *Procedia Computer Science*, *73*, 358–365. https://doi.org/10.1016/j.procs.2015.12.004

Anjali, Jivani, G., & Anjali, M. (2007). A Comparative Study of Stemming Algorithms. *Kenbenoit.Net*, *2*(2004), 1930–1938. https://kenbenoit.net/assets/courses/tcd2014qta/readings/Jivani_ijcta2011020632.pdf

Araujo, H., Costa, C. J., & Aparicio, M. (2017, July 11). Modelo de competitive intelligence (CI) competitive intelligence (CI) model. *Iberian Conference on Information Systems and Technologies, CISTI*. https://doi.org/10.23919/CISTI.2017.7975787

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Shlomo Levitan, L. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, *58*(6), 802–822. https://doi.org/10.1002/asi.20553

Arora, J., Agrawal, S., Goyal, P., & Pathak, S. (2017). Extracting entities of interest from comparative product reviews. *International Conference on Information and Knowledge Management, Proceedings*, *Part F1318*, 1975–1978. https://doi.org/10.1145/3132847.3133141

Arroyo, J., Muñoz, A., Roque, S., Maté, C., & And´ And´angel Sarabia, A. (2007). Exponential smoothing methods for interval time series. *In Proceedings of the 1st European Symposium on Time Series Prediction*, 231–240.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1–16. https://doi.org/10.1093/llc/7.1.1

Aurangzeb, K., Baharum, B., Hong, L. L., & Khairullah, K. (2011). Journal of Advances in Information Technology. *Open Computer Science*, *1*(1), 1.

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. *Clinical Pharmacology and Therapeutics*, *107*(4), 871–885. https://doi.org/10.1002/cpt.1796

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). *Stemming and lemmatization: A comparison of retrieval performances - UM Research Repository*. Proceedings of SCEI Seoul Conferences. https://eprints.um.edu.my/13423/

Barrientos Felipa, P. (2017). Marketing + internet = e-commerce: oportunidades y desafíos. *Revista Finanzas y Política Económica*, *9*(1), 41–56. https://doi.org/10.14718/revfinanzpolitecon.2017.9.1.3

Basili, V. (1992). *Software modeling and measurement: the Goal/Question/Metric paradigm*. https://drum.lib.umd.edu/bitstream/handle/1903/7538/?sequence=1

Baştanlar, Y., & Özuysal, M. (2014). Introduction to machine learning. *Methods in Molecular Biology*, *1107*, 105–128. https://doi.org/10.1007/978-1-62703-748-8_7/COVER/

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Baviera Puig, T. (2017). Técnicas para el Análisis de Sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. *Dígitos: Revista de Comunicación Digital*, *1*(3), 33–50. https://doi.org/10.7203/rd.v1i3.74

Bekkar, M., Kheliouane Djemaa, D., & Akrouf Alitouche, D. (2013). *Evaluation Measures for Models Assessment over Imbalanced Data Sets*. *3*(10). www.iiste.org

Berry, M. W., & Castellanos, M. (2008). Survey of Text Mining II. In *New York*. https://link.springer.com/content/pdf/10.1007/978-1-4757-4305-0.pdf

Bonaccorso, G., Fandango, A., & Shanmugamani, R. (2018). *Python : Expert Machine Learning Systems and Intelligent Agents Using Python.* https://books.google.com/books?hl=es&lr=&id=GtCBDwAAQBAJ&oi=fnd&pg=PP1&dq=Bonaccorso,+G.,+Fandango,+A.,+y+Shanmugamani,+R.+(2018).+Python:+Advanced+Guide+to+Artificial+Intelligence:+Expert+machine+learning+systems+and+intelligent+agents+using+Python.+Pack

Bose, R. (2008). Competitive intelligence process and tools for intelligence analysis. *Industrial Management and Data Systems*, *108*(4), 510–528. https://doi.org/10.1108/02635570810868362

Brath, R., & Peters, M. (2004). Dashboard Design: Why Design is Important. *Asepsis*, *16*(3), 5–8. http://cs.furman.edu/~pbatchelor/csc105/articles/TUN_DM_ONLINE.pdf

Briggs, W. M., & Zaretzki, R. (2008). The Skill Plot: A graphical technique for evaluating continuous diagnostic tests. *Biometrics*, *64*(1), 250–256. https://doi.org/10.1111/J.1541-0420.2007.00781_1.X

Bruijl, G. H. T. (2018). The Relevance of Porter's Five Forces in Today's Innovative and Changing Business Environment. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3192207

Bulley, C. A., Baku, K. F., & Allan, M. M. (2014). Competitive Intelligence Information: A Key Business Success Factor. *Journal of Management and Sustainability*, *4*(2). https://doi.org/10.5539/jms.v4n2p82

Cardoso, A., Talame, L., Amor, M., & Neil, C. (2019). *Minería de Opiniones : Análisis de Sentimientos*

en una Red Social. 1–5. http://sedici.unlp.edu.ar/handle/10915/77379

Cedeno-Moreno, D., & Vargas, M. (2020). Aprendizaje automático aplicado al análisis de sentimientos. *I+D Tecnológico*, *16*(2). https://doi.org/10.33412/idt.v16.2.2833

Chen, M. J. (1996). Competitor Analysis and Interfirm Rivalry: Toward A Theoretical Integration. *Https://Doi.Org/10.5465/Amr.1996.9602161567*, *21*(1), 100–134. https://doi.org/10.5465/AMR.1996.9602161567

Cortina, J. M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Dalianis, H. (2018). Evaluation Metrics and Evaluation. *Clinical Text Mining*, 45–53. https://doi.org/10.1007/978-3-319-78503-5_6

De Leeuw, J., Skidmore, A. K., De Leeuw{, J., Yang, L., Liu{, X., Schmidt{, K., & Skidmore{, A. K. (2006). Comparing accuracy assessments to infer superiority of image classification methods. *Taylor & Francis*, *27*(1), 223–232. https://doi.org/10.1080/01431160500275762

Del Alcázar, P. J. (2021). Ecuador Estado Digital Ene / 19. *Mentinno – Innovation & Lifetime Value Partners*, 37. https://www.academia.edu/download/65494042/Ecuador_Estado_Digital_enero_2021_Full.pdf

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

E-commerce. (2020). *DC0-Información pública E-Commerce en Ecuador*.

Effrosynidis, D., Peikos, G., Symeonidis, S., & Arampatzis, A. (2018). DUTH at SemEval-2018 Task 2: Emoji Prediction in Tweets. *Proceedings of The 12th International Workshop on Semantic Evaluation*, 466–469. https://doi.org/10.18653/V1/S18-1074

El Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In *Machine Learning in Radiation Oncology* (pp. 3–11). Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1

Elazmeh, W., Japkowicz, N., & Matwin, S. (2006). Evaluating misclassifications in imbalanced data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4212 LNAI*, 126–137. https://doi.org/10.1007/11871842_16

Explosion AI. (2017). spaCy - Indusrtial-strength Natural Language Processing in Python. In *Zenodo*. https://spacy.io/

Facebook. (2018). *Facebook Developer Docs | Facebook APIs, SDKs & Guides*. Facebook. https://developers.facebook.com/docs/

Fajardo Cárdenas, P., Bravo, A., Auquilla, A., & Vanegas, P. (2021). Plataforma para Análisis de Mercado a través de Datos de Redes Sociales. *Revista Tecnológica - ESPOL*, *33*(2), 134–146.

https://doi.org/10.37815/rte.v33n2.839

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. In *Choice Reviews Online* (Vol. 44, Issue 10). Cambridge university press. https://doi.org/10.5860/choice.44-5684

Feng, X. X., Feng, X. X., Qin, B., Feng, Z., & Liu, T. (2018). Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. *IJCAI*, *1*, 4071–4077. https://doi.org/10.24963/ijcai.2018/566

Gabes. (2012). Conceptual Model of Strategic Benefits of Competitive Intelligence Process Wadie Nasri Assistant Professor of Management in the Higher Institute of Management of. *International Journal of Business and Commerce*, *1*(6). www.ijbcnet.com

Gao, S., Tang, O., Wang, H., & Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management*, *71*, 19–32. https://doi.org/10.1016/j.ijhm.2017.09.004

Gardner, E. S. (2006). Exponential smoothing: The state of the art-Part II. *International Journal of Forecasting*, *22*(4), 637–666. https://doi.org/10.1016/J.IJFORECAST.2006.03.005

George K, S., & Joseph, S. (2014). Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature. *IOSR Journal of Computer Engineering*, *16*(1), 34–38. https://doi.org/10.9790/0661-16153438

Gilad, B., & Gilad, T. (1985). A Systems Approach to Business Intelligence. Business Horizons. *Business Horizons*, *28*, 65–70.

Gobinda, G. C. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*, 51–89.

Google Developers. (2021). *YouTube Data API | Google Developers*. https://developers.google.com/youtube/v3

Gundersen, S. (2019). *The rise of a new competitive intelligence: need of real-time competitive intelligence and the impact on decision-making*. http://hdl.handle.net/10362/106283

Havenga, J., & Botha, D. (2003). Developing competitive intelligence in the knowledge-based organization. *Southern African Online Information*, 1–22. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.3245&rep=rep1&type=pdf

He, W., Tian, X., Chen, Y., & Chong, D. (2016). Actionable Social Media Competitive Analytics For Understanding Customer Experiences. *Http://Dx.Doi.Org/10.1080/08874417.2016.1117377*, *56*(2), 145–155. https://doi.org/10.1080/08874417.2016.1117377

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*, *33*(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Hellriegel, O. T. (2021). *Scrape Facebook public pages without an API key*. GitHub.

https://github.com/oth11/facebook-scraper

Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, *52*(3), 1495–1545. https://doi.org/10.1007/S10462-017-9599-6

Hernàndez-Orallo, J., Ferri, C., Lachiche, N., & Flach, P. (2004). The 1st workshop on ROC analysis in artificial intelligence (ROCAI-2004). *ACM SIGKDD Explorations Newsletter*, *6*(2), 159–161. https://doi.org/10.1145/1046456.1046489

Herreros, D. C., & Rico, M. (2021). *Aplicación de Herramientas de Reconocimiento de Entidades Nombradas a SmartTerp*. https://oa.upm.es/id/eprint/67995

Herring, J. (1998). *"What is intelligence analysis?"* Competitive Intelligence Magazine.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. LDV Forum-GLDV J. *Comput. Linguist. Lang. Technol*, *20*(1), 19–62. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.447.4161&rep=rep1&type=pdf

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. https://doi.org/10.1145/1014052.1014073

Huang, C. R., & Yao, Y. (2015). Corpus Linguistics. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, 949–953. https://doi.org/10.1016/B978-0-08-097086-8.52004-2

Huang, K., Hussain, A., Wang, Q.-F., & Zhang, R. (Eds.). (2019). *Deep Learning: Fundamentals, Theory and Applications*. *2*. https://doi.org/10.1007/978-3-030-06073-2

Hyman, R. (2010). Quasi-Experimentation: Design and Analysis Issues for Field Settings (Book). *Http://Dx.Doi.Org/10.1207/S15327752jpa4601_16*, *46*(1), 96–97. https://doi.org/10.1207/S15327752JPA4601_16

Ismail, H., Harous, S., & B Belkhouche. (2016). A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis. *Res. Comput.* https://www.researchgate.net/profile/Heba-Ismail-2/publication/312913630_A_Comparative_Analysis_of_Machine_Learning_Classifiers_for_Twitter_Sentiment_Analysis/data/58899f2da6fdcc9a35c1a20f/A-Comparative-Analysis-of-Machine-Learning-Classifiers-for-Twitter

Jeong, B., Yoon, J., & Lee, J. M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, *48*, 280–290. https://doi.org/10.1016/j.ijinfomgt.2017.09.009

Jin, J., Ji, P., & Gu, R. (2016). Identifying comparative customer requirements from product online reviews for competitor analysis. *Engineering Applications of Artificial Intelligence*, *49*, 61–73. https://doi.org/10.1016/j.engappai.2015.12.005

Jindal, N., & Liu, B. (2006a). Identifying comparative sentences in text documents. *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *2006*, 244–251. https://doi.org/10.1145/1148170.1148215

Jindal, N., & Liu, B. (2006b). Mining comparative sentences and relations. *Proceedings of the National Conference on Artificial Intelligence*, *2*(13311336), 1331–1336.

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975

Kahaner, L. (1997). *Competitive Intelligence: How To Gather Analyze And Use Information To Move Your Business To The Top*. https://books.google.com.ec/books?hl=es&lr=&id=K3QfGoGSzmoC&oi=fnd&pg=PA7&dq=L.+Kahaner,+Competitive+intelligence:+how+to+gather,+analyze+and+use+information+to+move+your+business+to+the+top.+Simon+and+Schuster,+1997.&ots=bbsJpPFDse&sig=RJY8mqaG-T0Q3o87A3

Kessler, W., & Kuhn, J. (2014). A corpus of comparisons in product reviews. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 2242–2248. https://aclanthology.org/L14-1003/

Kessler, W., & Kuhn, J. (2013). Detection of product comparisons - How far does an out-of-the-box semantic role labeling system take you? *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1892–1897. http://www.cs.uic.edu/

Khan, A. U. R., Khan, M., & Khan, M. B. (2016). Naïve Multi-label Classification of YouTube Comments Using Comparative Opinion Mining. *Procedia Computer Science*, *82*, 57–64. https://doi.org/10.1016/j.procs.2016.04.009

Kim, Y., Dwivedi, R., Zhang, J., & Jeong, S. R. (2016). Competitive intelligence in social media Twitter: IPhone 6 vs. Galaxy S5. *Online Information Review*, *40*(1), 42–61. https://doi.org/10.1108/OIR-03-2015-0068

Kitchenham, B. A., Pickard, S. L., Jones, L. M., Hoaglin, P. W., El-Emam, & Rosenberg. (2001). Preliminary guidelines for empirical research in software engineering. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/1027796/

Koseoglu, M. A., Karayormuk, K., Parnell, J. A., & Menefee, M. L. (2011). Competitive intelligence: Evidence from Turkish SMEs. *International Journal of Entrepreneurship and Small Business*, *13*(3), 333–349. https://doi.org/10.1504/IJESB.2011.041664

Kotsiantis, S. B., & Kanellopoulos, D. (2006). Data preprocessing for supervised leaning. *International Journal of …*, *1*(2), 1–7. https://doi.org/10.1080/02331931003692557

Kramer, O. (2016). *Scikit-Learn* (pp. 45–53). Springer, Cham. https://doi.org/10.1007/978-3-319-33383-0_5

Krenker, A., Bester, J., & Kos, A. (2011). Introduction to the Artificial Neural Networks. In *Artificial Neural Networks - Methodological Advances and Biomedical Applications*.

https://doi.org/10.5772/15751

Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping. *Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*. https://www.researchgate.net/profile/Vlad-Krotov/publication/324907302_Legality_and_Ethics_of_Web_Scraping/links/5aea622345851588dd8287dc/Legality-and-Ethics-of-Web-Scraping.pdf

Li, S., Zha, Z. J., Ming, Z., Wang, M., Chua, T. S., Guo, J., & Xu, W. (2011). Product comparison using comparative relations. *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1151–1152. https://doi.org/10.1145/2009916.2010094

Li, Y., Jia, B., Guo, Y., & Chen, X. (2017). Mining User Reviews for Mobile App Comparisons. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–15. https://doi.org/10.1145/3130935

Liu, B., Zhang, L., Aggarwal, C., & Zhai, C. (2012). Mining text data. *Ch. A Survey of Opinion Mining and Sentiment Analysis. Springer, Boston*, 415–463.

Liu, Q., Huang, H., Zhang, C., Chen, Z., & Chen, J. (2013). Chinese comparative sentence identification based on the combination of rules and statistics. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8347 LNAI*(PART 2), 300–310. https://doi.org/10.1007/978-3-642-53917-6_27/COVER/

Liu, Y., Jiang, C., & Zhao, H. (2019). Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. *Decision Support Systems*, *123*, 113079. https://doi.org/10.1016/j.dss.2019.113079

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Interactive Presentation Sessions*, 69–72. https://doi.org/10.48550/arxiv.cs/0205028

Loster, M., Zuo, Z., Naumann, F., Maspfuhl, O., & Thomas, D. (2017). Improving Company Recognition from Unstructured Text by using Dictionaries. *EDBT*, 610–619.

Lu, T. J. (2015). Semi-supervised microblog sentiment analysis using social relation and text similarity. *2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015*, 194–201. https://doi.org/10.1109/35021BIGCOMP.2015.7072831

M, H., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Mahesh, B. (2018). Machine Learning Algorithms-A Review. *International Journal of Science and Research*. https://doi.org/10.21275/ART20203995

Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media and Society*, *6*(3).

https://doi.org/10.1177/2056305120940703

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval. Vol. 1*. Cambridge University.

Mark Davies. (2019). Corpus-based Studies of Lexical and Semantic Variation: The Importance of Both Corpus Size and Corpus Design. In *From Data to Evidence in English Language Research* (pp. 66–87). BRILL. https://doi.org/10.1163/9789004390652_004

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards and Interfaces*, *35*(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

Martínez Cámara, E., Martín Valdivia, M. T., Perea Ortega, J. M., & Ureña López, L. A. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en Español. In *Procesamiento de Lenguaje Natural* (Vol. 47, pp. 163–170). http://rua.ua.es/dspace/handle/10045/18524

McKinney, W., Data, P. T.-P. P., & 2015, U. (2012). Pandas-Powerful python data analysis toolkit. *Pandas.Pydata.Org*. https://pandas.pydata.org/pandas-docs/version/0.7.3/pandas.pdf

Mclean, L., & Woods, L.-A. (2014). Competitive intelligence capabilities. *Procedia - Social and Behavioral Sciences*, *110*, 669–677. https://doi.org/10.1016/j.sbspro.2018.12.911

Microsoft. (2022). *Visualización de datos | Microsoft Power BI*. https://powerbi.microsoft.com/es-es/

Mike, N. (2010). Building a corpus : What are the key considerations? *The Routledge Handbook of Corpus Linguistics*, 31–37. https://doi.org/10.4324/9780203856949-5

Miranda, C. H., Guzmán, J., & Salcedo, D. (2016). Minería de opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hotels. *Procesamiento de Lenguaje Natural*, *56*(56), 25–32. http://www.redalyc.org/articulo.oa?id=515754423002

Mohd Razali, N., & Bee Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 13–14.

Mohit, B. (2014). *Named Entity Recognition* (pp. 221–245). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45358-8_7

Mohri, R., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of Machine Learning. *Cambridge, MA: The MIT Press.*

Molina, C. A. C., Gutierrez, R. E., & Solarte, O. (2015). Prototipo para el reconocimiento de entidades nombradas en el idioma Español. *2015 10th Colombian Computing Conference, 10CCC 2015*, 364–371. https://doi.org/10.1109/COLUMBIANCC.2015.7333447

Mozetic, I., Grcar, M., & Smailovic, J. (2016). *Twitter sentiment for 15 European languages*. https://www.clarin.si/repository/xmlui/handle/11356/1054

Nasri, W. (2011). Competitive intelligence in Tunisian companies. *Journal of Enterprise Information*

Management, *24*(1), 53–67. https://doi.org/10.1108/17410391111097429/FULL/XML

Navas-Loro, M., Rodríguez-Doncel, V., Fernández-Izquierdo, A., Santana-Pérez, I., & Sánchez, A. (2018, June 20). *MAS corpus | Zenodo*. https://zenodo.org/record/1293493

Nayak, A. (2016). Comparative study of Naïve Bayes , Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds. *International Journal of Advanced Studies in Computer Science and Engineering*, *5*(1), 14–17.

Nenzhelele, T. E., & Pellissier, R. (2014). Competitive intelligence implementation challenges of small and medium-sized enterprises. *Mediterranean Journal of Social Sciences*, *5*(16), 92–99. https://doi.org/10.5901/mjss.2014.v5n16p92

Nisbet, R., Miner, G., & Yale, K. (2017). Handbook of statistical analysis and data mining applications. In *Handbook of Statistical Analysis and Data Mining Applications*. https://doi.org/10.1016/c2012-0-06451-4

Pablo Moreno. (2018, June 11). *Analisis Predictivo con Power BI*. https://www.pbiusergroup.com/blogs/pablo-moreno/2018/06/11/analisis-predictivo-con-power-bi

Parodi, G. (2008). Lingüistica de corpus: Una introduccion al ambito. *RLA*, *46*(1), 93–119. https://doi.org/10.4067/S0718-48832008000100006

Pauli, P. (2019). *Análisis de sentimiento: comparación de algoritmos predictivos y métodos utilizando un lexicon español*. https://ri.itba.edu.ar/handle/123456789/1782

Pedregosa, F., Weiss, R., Brucher, M., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Peñalver-Martínez, I., Valencia-García, R., & García-Sánchez, F. (2011). *Minería de Opiniones basada en características guiada por Ontologías*. https://rua.ua.es/dspace/bitstream/10045/16947/1/PLN_46_11.pdf

Peng, Y. S., & Liang, I. C. (2016). A dynamic framework for competitor identification: A neglecting role of dominant design. *Journal of Business Research*, *69*(5), 1898–1903. https://doi.org/10.1016/j.jbusres.2015.10.076

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. http://arxiv.org/abs/2106.09462

Perktold, J., Mckinney, W., & Seabold, S. (2011). Time Series Analysis in Python with statsmodels. *PROC. OF THE 10th PYTHON IN SCIENCE CONF*, 107. https://doi.org/10.25080/Majora-ebaa42b7-012

Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. *International Journal of Social Research Methodology*, *23*(3), 277–290. https://doi.org/10.1080/13645579.2019.1682840

Peteraf, M. A., & Bergen, M. E. (2003). Scanning dynamic competitive landscapes: a market-based and resource-based framework. *Strategic Management Journal*, *24*(10), 1027–1041. https://doi.org/10.1002/SMJ.325

Plisson, J., Lavrac, N., & Mladenić, D. D. (2004). A rule based approach to word lemmatization. *Proceedings of the 7th International Multiconference Information Society (IS'04)*, 83–86. http://eprints.pascal-network.org/archive/00000715/

Ponis, S. T., & Christou, I. T. (2013). Competitive intelligence for SMEs: A web-based decision support system. *International Journal of Business Information Systems*, *12*(3), 243–258. https://doi.org/10.1504/IJBIS.2013.052449

Porter, M. E. (2008). The five competitive forces that shape strategy. *Harvard Business Review*, *86*(1). https://www.academia.edu/download/49313875/Forces_That_Shape_Competition.pdf#page=25

Powers, D. M. W., & Ailab. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. https://doi.org/10.48550/arxiv.2010.16061

Provost, F., & Fawcett, T. (1997). *Analysis and Visualization of Classifier Performance with Nonuniform Class and Cost Distributions*. www.aaai.org

Pustejovsky, J., & Stubbs, A. (2013). Natural language annotation for machine learning. In *Vasa*. https://books.google.com/books?hl=es&lr=&id=A57TS7fs8MUC&oi=fnd&pg=PR2&dq=Pustejovsky,+J.,+%26+Stubbs,+A.+(2012).+Natural+Language+Annotation+for+Machine+Learning:+A+guide+to+corpus-building+for+applications.+%22+O%27Reilly+Media,+Inc.%22.&ots=SKfxwHZsyI&

Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017). Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. https://doi.org/10.1109/JCDL.2017.7991618

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information 2020, Vol. 11, Page 193*, *11*(4), 193. https://doi.org/10.3390/INFO11040193

Reitermanov´a, Z. (2010). DATA SPLITTING. *WDS'10 Proceedings of Contributed Paper*, *1*, 31–36. https://www.mff.cuni.cz/veda/konference/wds/proc/pdf10/WDS10_105_i1_Reitermanova.pdf

Reyes-Ortiz, J. A., Paniagua-Reyes, F., & Sánchez, L. (2017). Minería de opiniones centrada en tópicos usando textos cortos en español. *Research in Computing Science*, *134*(1), 151–162. https://doi.org/10.13053/rcs-134-1-12

Rodríguez-Rodríguez, J., & Reguant-Álvarez, M. (2020). Calcular la fiabilitat d'un qüestionari o escala mitjançant l'SPSS: el coeficient alfa de Cronbach. *REIRE Revista d'Innovació i Recerca En Educació*, *13*(2), 1–13–1–13. https://doi.org/10.1344/REIRE2020.13.230048

Salazar Loor, S. A., & Ponce Intriago, K. E. (2018). Análisis del uso de data mining de las redes sociales y su influencia en la competitividad de las PYMES. *Revista Científica Ciencia y Tecnología*,

18(Vol. 18 Núm. 18 (2018)), 177–191. https://doi.org/10.47189/rcct.v18i18.181

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLoS ONE*, *16*(8 August), e0254937. https://doi.org/10.1371/journal.pone.0254937

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591. https://doi.org/10.2307/2333709

Sieminski, A., Kozierkiewicz, A., Nunez, M., & Ha, Q. T. (Eds.). (2018). *Modern Approaches for Intelligent Information and Database Systems*. *769*. https://doi.org/10.1007/978-3-319-76081-0

SimilarWeb. (2021). *Tráfico del sitio web: compruebe y analice cualquier sitio web | Similarweb*. https://www.similarweb.com/es/

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AAAI Workshop - Technical Report*, *WS-06-06*, 24–29. https://doi.org/10.1007/11941439_114

Southerton, D. (2014). Likert Scales. In *Encyclopedia of Consumer Culture*. https://doi.org/10.4135/9781412994248.n322

Stefanikova, Ľ., Rypakova, M., & Moravcikova, K. (2015). The Impact of Competitive Intelligence on Sustainable Growth of the Enterprises. *Procedia Economics and Finance*, *26*, 209–214. https://doi.org/10.1016/s2212-5671(15)00816-3

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Sun, J., Long, C., Zhu, X., & Huang, M. (2009). Mining Reviews for Product Comparison and Recommendation. *Polibits*, *39*, 33–40. https://doi.org/10.17562/pb-39-5

Taboada, M. (2017). *SFU Review Corpus | Maite Taboada*. https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html

Taheri, S. M., & Hesamian, G. (2013). A generalization of the Wilcoxon signed-rank test and its applications. *Statistical Papers*, *54*(2), 457–470. https://doi.org/10.1007/S00362-012-0443-4

Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). A Survey on Text Mining Techniques. *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*, 1022–1026. https://doi.org/10.1109/ICACCS.2019.8728547

TASS Team. (2012). *TASS: Workshop on Semantic Analysis at SEPLN*. TASS - SEPLN. http://tass.sepln.org/tass_data/download.php?auth=etKassVsC4AeqvyeFrj

Taulé, M., M.A. Martí, M. R. (2008). Ancora: Multilingual and multilevel annotated corpora. *Proceedings of 6th International Conference on Language Resources and Evaluation*, 96–101. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.5816&rep=rep1&type=pdf

Tjong Kim Sang, E. F., & de Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task:

Language-Independent Named Entity Recognition. *Proceedings of the 7th Conference on Natural Language Learning, CoNLL 2003 at HLT-NAACL 2003*, 142–147. https://arxiv.org/abs/cs/0306050

Tripathi, G. (2015). FEATURE SELECTION AND CLASSIFICATION APPROACH FOR SENTIMENT ANALYSIS. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(2). https://doi.org/10.5121/mlaij.2015.2201

Tripathi, G., & S, N. (2015). Feature Selection and Classification Approach for Sentiment Analysis. *Machine Learning and Applications: An International Journal*, *2*(2), 01–16. https://doi.org/10.5121/mlaij.2015.2201

Tsirakis, N., Poulopoulos, V., Tsantilas, P., & Varlamis, I. (2017). Large scale opinion mining for social, news and blog data. *Journal of Systems and Software*, *127*, 237–248. https://doi.org/10.1016/j.jss.2016.06.012

Turney, P. D. (2002). *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*. https://doi.org/10.48550/arxiv.cs/0212032

Twitter, I. (2019). *Información sobre las API de Twitter*. https://help.twitter.com/es/rules-and-policies/twitter-api

ur Rehman, I. (2019). Facebook-Cambridge Analytica data harvesting: What you need to know. *Library Philosophy and Practice*, *2019*. https://core.ac.uk/download/pdf/215162147.pdf

Varathan, K. D., Giachanou, A., & Crestani, F. (2017). Comparative opinion mining: A review. *Journal of the Association for Information Science and Technology*, *68*(4), 811–829. https://doi.org/10.1002/asi.23716

Vera Kristanti Dewi, M., & Sri Darma, G. (2019). The Role of Marketing & Competitive Intelligence In Industrial Revolution 4.0. *Jurnal Manajemen Bisnis*, *16*(1), 1. https://doi.org/10.38043/jmb.v16i1.2014

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2013). A supervised approach to opinion mining on Spanish tweets based on linguistic knowledge. *A Supervised Approach to Opinion Mining on Spanish Tweets Based on Linguistic Knowledge*, *51*(51), 127–134. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4880

Wagstaff, K. L. (2012). Machine learning that matters. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, *1*, 529–534. https://doi.org/10.48550/arxiv.1206.4656

Wang, W., Zhao, T. J., Xin, G. D., & Xu, Y. D. (2015). Exploiting Machine Learning for Comparative Sentences Extraction. *International Journal of Hybrid Information Technology*, *8*(3), 347–354. https://doi.org/10.14257/ijhit.2015.8.3.31

Webb, G. I. (2016). Naïve Bayes. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1–2). https://doi.org/10.1007/978-1-4899-7502-7_581-1

Weischedel, R., Pradhan, S., Ramshaw, L., Kaufman, J., Franchini, M., El-Bachouti, M., Xue, N.,

Palmer, M., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., & Houston, A. (2010). Ontonotes release 4.0. *Catalog.Ldc.Upenn.Edu*. https://catalog.ldc.upenn.edu/docs/LDC2011T03/OntoNotes-Release-4.0.pdf

Weng, C. G., & Poon, J. (2008). A new evaluation measure for imbalanced datasets. *Conferences in Research and Practice in Information Technology Series*, *87*, 27–32.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, *18*(1), 45–55. https://doi.org/10.1177/016555159201800106

Wu, S., Fang, Z., & Tang, J. (2012). Accurate product name recognition from user generated content. *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, 874–877. https://doi.org/10.1109/ICDMW.2012.129

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49–73. https://doi.org/10.1007/s10462-017-9588-9

Xu, K., Liao, S. S., Lau, R. Y., Tang, H., & Wang, S. (2009). Building comparative product relation maps by mining consumer opinions on the Web. *15th Americas Conference on Information Systems 2009, AMCIS 2009*, *3*, 1653–1661. http://aisel.aisnet.org/amcis2009/179

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems*, *50*(4), 743–754. https://doi.org/10.1016/j.dss.2010.08.021

Xue, J. H., & Titterington, D. M. (2008). Do unbalanced data have a negative effect on LDA? *Pattern Recognition*, *41*(5), 1558–1571. https://doi.org/10.1016/J.PATCOG.2007.11.008

Xue, Y., Zhou, Y., & Dasgupta, S. (2018, June 26). Mining competitive intelligence from social media: A case study of IBM. *Proceedings of the 22nd Pacific Asia Conference on Information Systems - Opportunities and Challenges for the Digitized Society: Are We Ready?, PACIS 2018*. https://aisel.aisnet.org/pacis2018/313

Yadav, S., & Shah, D. (2019). Opinion Mining from Customer Reviews for Predicting Competitors. *International Research Journal of Engineering and Technology*, 1180. www.irjet.net

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 2145–2158. http://arxiv.org/abs/1910.11470

Younis, U., Asghar, M. Z., Khan, A., Khan, A., Iqbal, J., & Jillani, N. (2020). Applying Machine Learning Techniques for Performing Comparative Opinion Mining. *Open Computer Science*, *10*(1). https://doi.org/10.1515/COMP-2020-0148/HTML

Zanasi, A. (1998). Competitive intelligence through data mining public sources. *Competitive Intelligence Review*, *9*(1), 44–54. https://doi.org/10.1002/(sici)1520-6386(199801/03)9:1<44::aid-cir8>3.0.co;2-a

Zhang, X. Da. (2020). A matrix algebra approach to artificial intelligence. In *A Matrix Algebra Approach to Artificial Intelligence*. Springer Singapore. https://doi.org/10.1007/978-981-15-

2770-8

ANEXO 1: CUESTIONARIO REALIZADO EN LA EVALUACIÓN

Evaluación sobre detección y análisis de competidores en el contexto del Sector Textil a través de las Redes Sociales.

Esta encuesta tiene como objetivo evaluar la utilidad de la aplicación para la detección de competidores en base a tres de las fuerzas competitivas que moldean la estrategia según Michael Porter.

Link Dashboard: https://app.powerbi.com/view?r=eyJrIjoiNjc2MzNiOTMtNjkyNC00MjA1LThmZjEtODhlMzkzOGNlODRkIiwidCI6IjhmNDY4MTZhLTcyMjAtNDg1MS04ZWYzLTY4MWI2MGM3ZmYwZiIsImMiOjR9&pageName=ReportSection32869605aed1a10ec905

Nota: Los datos que van a ser evaluados son recolectados de las diferentes plataformas desde el 1 de enero del 2022 al 5 de abril del 2022

patricio.fajardo96@ucuenca.edu.ec Cambiar de cuenta

*Obligatorio

Correo *

apatricio.fajardoc@gmail.com

Introducción

En la actualidad con el avance de la tecnología y también debido a la pandemia el uso de redes sociales por los consumidores ha incrementado enormemente, en estos medios los consumidores pueden expresar abiertamente comentarios tanto positivos como negativos sobre empresas y/o productos. Por lo tanto, realizar un análisis de datos de estas plataformas es importante para que las empresas puedan ganar ventaja competitiva. En esta investigación se realizan procesos para generar corpus en el idioma español que permita crear modelos de machine learning para detectar posibles competidores. Luego para evaluar la utilidad del corpus generado se crean algunos modelos y se presentan los resultados en un Dashboard.

Para la evaluación de este corpus se tiene en consideración tres fuerzas competitivas [1] de Michel Porter [2] sobre el análisis competitivo. Michael Porter es profesor de la Universidad de Harvard en la escuela de negocios y un reconocido investigador sobre entornos competitivos entre empresas.

[1] PORTER, Michael E. The five competitive forces that shape strategy. Harvard business review, 2008, vol. 86, no 1, p. 25-40.
[2] Michael Porter. https://scholar.google.com/citations?user=g9WIbh0AAAAJ&hl=es&oi=ao

Proceso Para la Evaluación

Este proceso consta de dos pasos:

- Presentación del Dashboard (10min)
- Llenar una encuesta sobre lo observado en el Dashboard (5min)

Siguiente

Borrar formulario

Evaluación sobre detección y análisis de competidores en el contexto del Sector Textil a través de las Redes Sociales.

patricio.fajardo96@ucuenca.edu.ec Cambiar de cuenta

Encuesta

En base a lo observado en la aplicación, y los resultados que fueron presentados.
Por favor responda a las siguientes preguntas.
Nota. Cada pregunta de la encuesta tendrá 5 opciones:

- Totalmente de acuerdo (5)
- De acuerdo (4)
- Indeciso (3)
- En desacuerdo (2)
- Totalmente en desacuerdo (1)

Página 2 de 5

Atrás Siguiente Borrar formulario

Nunca envíes contraseñas a través de Formularios de Google.

Evaluación sobre detección y análisis de competidores en el contexto del Sector Textil a través de las Redes Sociales.

patricio.fajardo96@ucuenca.edu.ec Cambiar de cuenta

*Obligatorio

FUERZA 1. Rivalidad entre competidores actuales

La rivalidad entre los competidores existentes adopta muchas formas familiares, incluidos descuentos de precios, lanzamiento de nuevos productos, campañas publicitarias y mejoras en el servicio. La alta rivalidad limita la rentabilidad de una industria. El grado en que la rivalidad reduce el potencial de ganancias de una industria depende, en primer lugar, de la intensidad con la que compiten las empresas y, en segundo lugar, de la base sobre la que compiten.

F1.1 ¿Puede determinar si existen muchos o pocos competidores en las redes sociales? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo

F1.2 ¿Puede determinar si existe una competencia de precios en la industria? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo

F1.3 ¿Puede determinar si los productos seguirán mostrando interés a los clientes en el futuro? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo

F1.4 ¿Puede determinar que productos tienen una alta o baja demanda durante cierto período de tiempo? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo

Página 3 de 5

Atrás Siguiente Borrar formulario

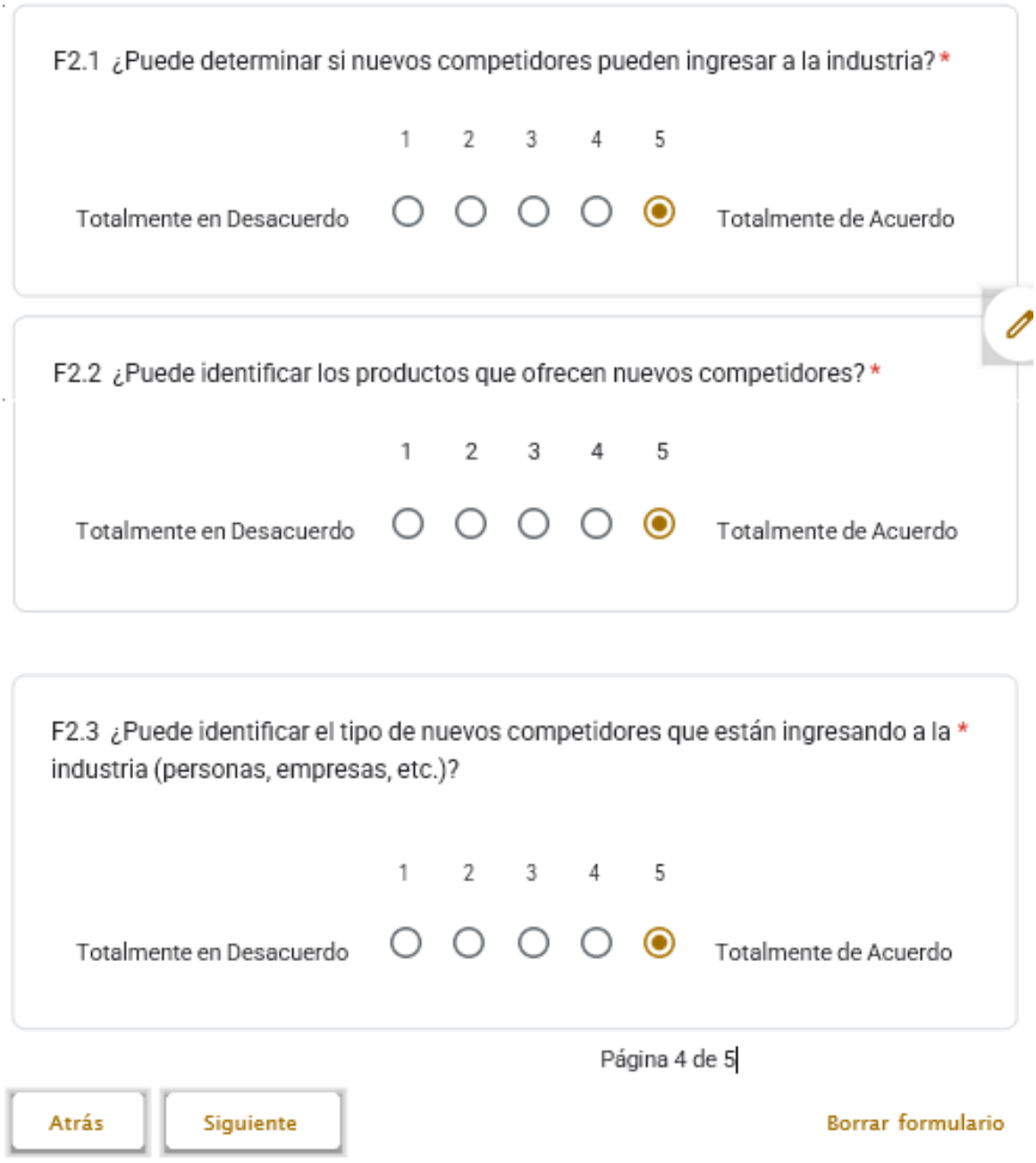
Evaluación sobre detección y análisis de competidores en el contexto del Sector Textil a través de las Redes Sociales.

patricio.fajardo96@ucuenca.edu.ec Cambiar de cuenta

*Obligatorio

FUERZA 2. Amenaza de competidores potenciales

Los nuevos participantes en una industria aportan nueva capacidad y un deseo de ganar participación de mercado que ejerce presión sobre los precios, los costos y la tasa de inversión necesaria para competir. Particularmente cuando los nuevos participantes se están diversificando de otros mercados, pueden aprovechar las capacidades existentes y los flujos de efectivo para sacudir la competencia, como lo hizo Pepsi cuando ingresó a la industria del agua embotellada, Microsoft cuando comenzó a ofrecer navegadores de Internet y Apple cuando ingresó el negocio de distribución de música.



F2.1 ¿Puede determinar si nuevos competidores pueden ingresar a la industria? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo

F2.2 ¿Puede identificar los productos que ofrecen nuevos competidores? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo

F2.3 ¿Puede identificar el tipo de nuevos competidores que están ingresando a la industria (personas, empresas, etc.)? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo

Página 4 de 5

Atrás Siguiente Borrar formulario

Nunca envíes contraseñas a través de Formularios de Google.

Evaluación sobre detección y análisis de competidores en el contexto del Sector Textil a través de las Redes Sociales.

patricio.fajardo96@ucuenca.edu.ec Cambiar de cuenta

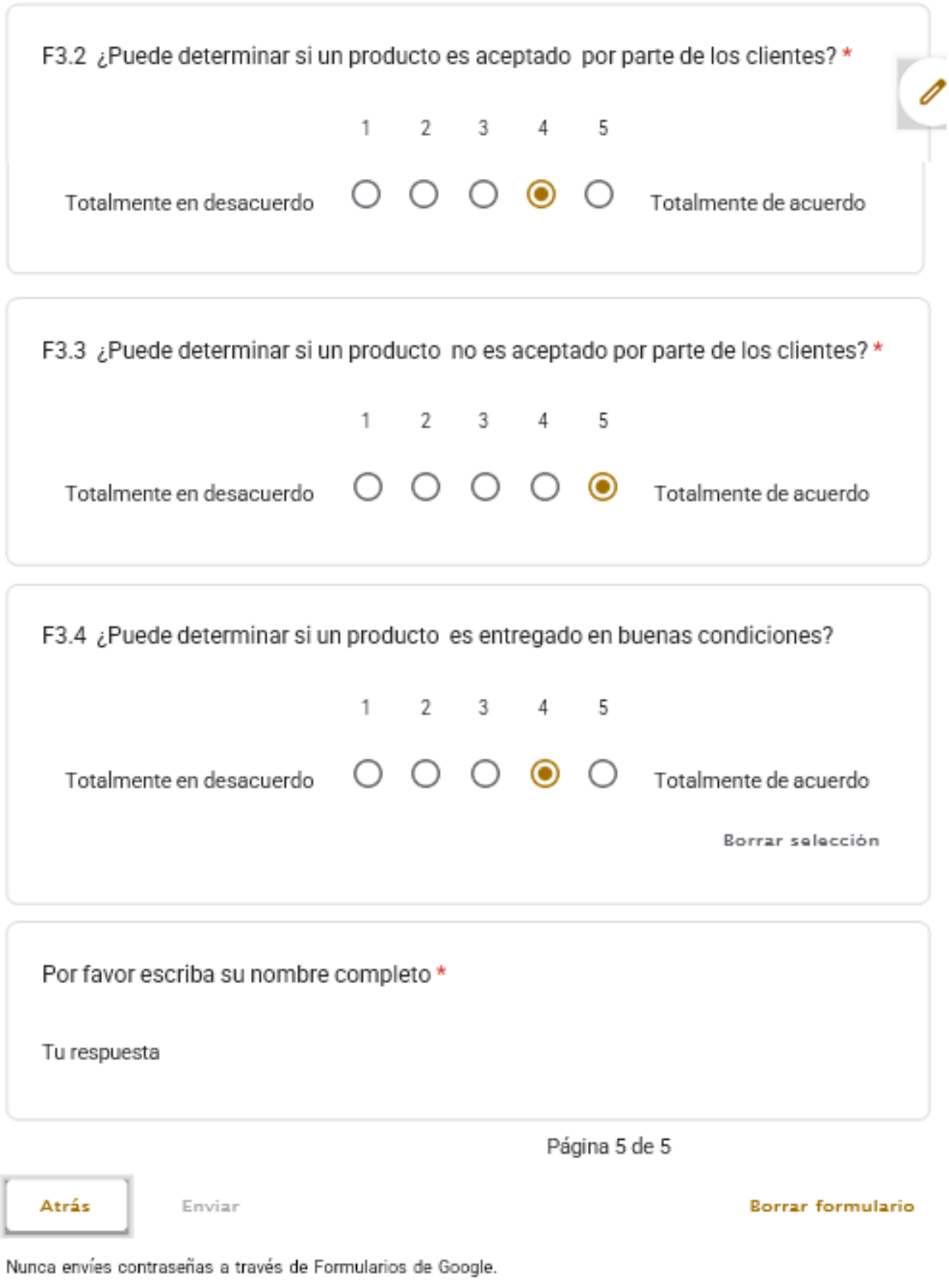
*Obligatorio

FUERZA 3. Poder de negociación de clientes

Los clientes poderosos pueden capturar más valor forzando a bajar los precios, exigiendo una mejor calidad o más servicio (lo que eleva los costos) y, en general, enfrentando a los participantes de la industria entre sí, todo a expensas de la rentabilidad de la industria.

F3.1 ¿Puede determinar si un competidor es aceptado o no por parte de los clientes? *

	1	2	3	4	5	
Totalmente en Desacuerdo	○	○	○	○	◉	Totalmente de Acuerdo



F3.2 ¿Puede determinar si un producto es aceptado por parte de los clientes? *

	1	2	3	4	5	
Totalmente en desacuerdo	○	○	○	◉	○	Totalmente de acuerdo

F3.3 ¿Puede determinar si un producto no es aceptado por parte de los clientes? *

	1	2	3	4	5	
Totalmente en desacuerdo	○	○	○	○	◉	Totalmente de acuerdo

F3.4 ¿Puede determinar si un producto es entregado en buenas condiciones?

	1	2	3	4	5	
Totalmente en desacuerdo	○	○	○	◉	○	Totalmente de acuerdo

Borrar selección

Por favor escriba su nombre completo *

Tu respuesta

Página 5 de 5

Atrás Enviar Borrar formulario

Nunca envíes contraseñas a través de Formularios de Google.

ANEXO 2: ARTICULO DE TIC.EC 2021

https://doi.org/10.37815/rte.v33n2.839
Artículos originales

Plataforma para Análisis de Mercado a través de Datos de Redes Sociales

Platform for Market Analysis through Social Network Data

Ángel Patricio Fajardo Cárdenas[1] https://orcid.org/0000-0002-7292-6983, Néstor Ariel Bravo Chuqui[2] https://orcid.org/0000-0002-7217-9455, Andrés Vinicio Auquilla Sangolqui[3] https://orcid.org/0000-0002-3754-041X, Paúl Fernando Vanegas Peña[4] https://orcid.org/0000-0002-3805-4130

[1] *Facultad de Ingeniería, Universidad de Cuenca*, Cuenca, Ecuador
patricio.fajardo96@ucuenca.edu.ec

[2] *Facultad de Ingeniería, Universidad de Cuenca*, Cuenca, Ecuador
ariel.bravo@ucuenca.edu.ec

[3] *Facultad de Ingeniería, Departamento de Ciencias de la Computación, Universidad de Cuenca*, Cuenca, Ecuador
andres.auquilla@ucuenca.edu.ec

[4] *Facultad de Ciencias Químicas, Departamento de Espacio y Población, Universidad de Cuenca*, Cuenca, Ecuador
paul.vanegas@ucuenca.edu.ec

Enviado: 2021/07/11
Aceptado: 2021/09/28
Publicado: 2021/11/30

Para obtener el articulo completo:
http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/839/535

ANEXO 3: REPOSITORIO CON LOS CÓDIGOS DESARROLLADOS

https://drive.google.com/drive/folders/1jvhFr4tBUFDtJY0rbsL4A1M4AiMipnYy?usp=sharing