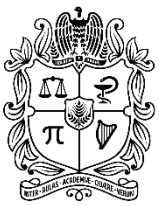


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Método para la detección de polaridad sobre una base de datos de reseñas de películas de cine, basado en una técnica combinada de selección y clasificación

Luis Miguel Ramirez Sandoval

Universidad Nacional de Colombia

Facultad de Minas

Medellín, Colombia

2022

Método para la detección de polaridad sobre una base de datos de reseñas de películas de cine, basado en una técnica combinada de selección y clasificación

Luis Miguel Ramirez Sandoval

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Analítica

Director (a):

Ph.D. Albeiro Espinosa Bedoya

Línea de Investigación:

Profundización

Universidad Nacional de Colombia

Facultad de Minas

Medellín, Colombia

2022

A mis padres, quienes siempre me han apoyado.

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Luis Miguel Ramirez Sandoval

Nombre

13/06/2022

Fecha

Resumen

Las críticas de cine tanto positivas como negativas tiene un impacto directo en la recaudación en taquilla de las producciones cinematográficas. Detectar si una crítica es positiva o negativa, representa un asunto de importancia para las empresas productoras, pues entre más críticas positivas reciban, mejores serán las probabilidades de que la película sea un éxito. El análisis de sentimientos se presenta como una herramienta útil para comprender la reacción del público y de la crítica hacia una película, especialmente cuando en la actualidad las redes sociales son una fuente de datos de las opiniones que hacen los usuarios sobre estas. Mediante el uso de técnicas de procesamiento de lenguaje natural, diferentes investigaciones han explorado la mejor forma para detectar la polaridad de una crítica; estas se han enfocado en dos aspectos principalmente, la selección de características y el método de clasificación, no obstante, ninguna de las investigaciones estudiadas plantea una técnica combinada de un método de selección y de clasificación, como lo son los métodos de ensamble. Para esta investigación se seleccionaron 3 arquitecturas descritas en la literatura y sus resultados fueron utilizados como las líneas base; luego, se propuso un modelo combinado con una arquitectura de dos caminos para realizar la detección de polaridad, y finalmente, se compararon y validaron los resultados de los modelos de referencia y del modelo propuesto. La arquitectura diseñada, que se construyó con base a los métodos de ensamble y de fusión, alcanzó un accuracy del 96% y un recall del 90%, además, al ser una arquitectura de dos caminos permitió analizar la importancia de diferentes variables, así como también de diferentes métodos en la predicción final.

Palabras clave: procesamiento de lenguaje natural, análisis de sentimientos, redes neuronales recurrente, redes neuronales convoluciones, métodos de ensamble.

Abstract

Method for polarity detection on a database of movie reviews, based on a combined selection and classification technique

Movie reviews both positive and negative have a direct impact at the box office of the cinematographic productions. Detect if a review is positive or not, represents a matter of importance by the production's companies, because more positive reviews a movie receives the better the probabilities to the movie be a success. Sentiment analysis is a useful tool to understand the audience and the professional critics reaction towards a movie, specially nowadays that people use social media networks to share their opinions. With the use of natural language processing techniques, different research had explored the best way to determine a movie's review polarity, most of them had focused on two main aspects, feature selection and the classification method, however, none of them proposed a combine technique of a feature selection method and a classification method, like the ensemble methods. For this investigation, 3 studies were selected, and their result were used as benchmarks; then, a new method was proposed with a two-way architecture, to detect the polarity of movie reviews, after this process was completed, the results of the benchmarks and the proposed model were compared and validated. The new proposed architecture, based on ensemble methods, achieved a 96% accuracy and 90% recall, also, this architecture allowed to analyze the importance of different variables in the model as well as the importance of different methods in the final prediction.

Keywords: natural language processing, sentiment analysis, recurrent neural networks, convolutional networks, ensemble methods.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

Las críticas de cine se han establecido como una fuente confiable de la calidad o el éxito de una película tanto para las empresas productoras como para el público en general, que esperan leer las reseñas de las películas de su interés. Encontrar reseñas positivas o negativas sobre una película puede ser determinante para influenciar al público para asistir a verla o, por el contrario, podría desanimar a los potenciales espectadores, lo que se traduciría en un fracaso para la empresa productora. Si bien, las reseñas escritas por críticos de cine profesionales cumplen un rol más de predictores que de influenciadores, es decir, sus reseñas pueden ser usadas para predecir si una película será un éxito (reseña positiva), poco hacen para influenciar al público para asistir a verlas. No obstante, el fenómeno de Word of Mouth (WoM), o voz a voz, y su efecto influenciador ha sido ampliamente estudiado; este fenómeno consiste en “correr la voz” de una opinión, buena o mala, y esto hará que las personas que reciben esa opinión sean influenciadas a favor o en contra (ir o no ir a cine), este fenómeno se ve potenciado en la actualidad por el masivo uso de las redes sociales, en donde las opiniones se replican y comparten rápidamente. Es por este fenómeno y su capacidad de influencia, que las empresas productoras se enfocan en obtener más y mejores reseñas, para con esto aumentar sus probabilidades de éxito.

Analizar las reseñas, opiniones y comentarios sobre una película y detectar su polaridad, se presenta entonces como un campo de estudio de gran importancia y que ha despertado gran interés por parte de la comunidad académica, que ha ido incrementando su producción de investigaciones sobre esta materia durante los últimos años. Determinar la polaridad de un texto (en este caso reseñas de películas de cine), basa su funcionamiento en técnicas de procesamiento del lenguaje natural y técnicas de análisis de sentimientos. Diferentes son las aproximaciones que se pueden encontrar en la literatura, en las cuales se plantean dos grandes enfoques para resolver esta tarea, el primero desde la selección de las características, en donde se busca encontrar las palabras que mejor puedan

predecir la polaridad de la reseña analizada, y el segundo enfoque que busca mejorar el método de clasificación que es el encargado de entregar la predicción final.

Si bien todas las aproximaciones disponibles son viables y sus resultados respaldan su implementación, no se encuentra una arquitectura combinada de selección y clasificación, bajo la cual se pueda de forma paralela optimizar los dos enfoques (selección y clasificación) y que además haga uso de técnicas más novedosas como lo son los métodos de ensamble y las técnicas de fusión, en donde diferentes arquitecturas y modelos son combinados para mejorar su desempeño.

Esta investigación tiene como principal objetivo proponer un método combinado de selección y clasificación para determinar la polaridad de reseñas de películas de cine, para esto, primero se seleccionaron tres arquitecturas descritas en diferentes artículos de investigación y se trató de reproducir sus resultados, entonces fueron utilizados como referencia contra los cuales el modelo propuesto por esta investigación fue comparado. Con estas referencias listas, se propuso un modelo combinado, basado en métodos de ensamble y técnicas de fusión; una vez obtenidos los resultados de este modelo, se comparó el desempeño del modelo propuesto contra aquel obtenido de los modelos de referencia.

Este documento se estructura de la siguiente forma: en el capítulo introductorio se describen los hallazgos de una revisión sistemática de la literatura sobre el estado del arte en el campo del análisis de sentimientos; en este capítulo también se describe de forma detallada del problema, su justificación y objetivos. En el segundo capítulo se presentan las arquitecturas seleccionadas como referencias y su desempeño. El tercer capítulo es dedicado a la construcción de un modelo combinado de selección y clasificación, allí se describe el proceso de desarrollo de esta propuesta de forma detallada, su arquitectura y sus métricas de desempeño. En el capítulo cuarto se realiza el análisis, comparación y validación de los resultados obtenidos por el modelo propuesto por esta investigación, con respecto a las métricas de desempeño de los modelos seleccionados como referencia. El último capítulo de este documento describe las conclusiones y recomendaciones sugeridas por esta investigación.

1. Capitulo introductorio

El cine es una de las formas de entretenimiento más populares alrededor del mundo; congrega semana tras semana a miles y miles de personas en diferentes salas de cine, para ver producciones de todos los idiomas, en todos los formatos, de todos los géneros y para todas las edades. Para conseguir estas grandes acogidas, las compañías productoras invierten grandes sumas de dinero en campañas publicitarias en donde se utilizan comerciales de televisión, pautas en redes sociales, pasacalles, entre otras, con el fin de aprovechar al máximo un fenómeno estudiado llamado Word of Mouth (WoM) [1]. Este fenómeno, que también se conoce como “voz a voz”, se produce cuando las personas comparten su opinión, o recomiendan una película, un restaurante, etc., y por esta razón los demás, deciden o seguir compartiendo esas opiniones a otras personas, o bien, deciden ir a ver la película o ir a cenar al restaurante que les fue recomendado; es entonces probable que si un conocido, familiar, amigo, recomiende un película los demás irán a verla, lo mismo pasaría si se una opinión se difunde masivamente, se estaría más propenso a ir a ver la película al haber escuchado de gran cantidad de personas de que es buena o por el contrario podría desanimar al público de ir a verla si lo que se escucha es que la película es mala.

Sang Ho Kim et al., en [1] concluyen sobre el importante efecto que tiene la constante publicación de críticas, opiniones, comentarios, etc., en blogs, post, en donde gracias a estos las ganancias de una película se han visto aumentadas. Se describe en este estudio, como las críticas hechas por críticos profesionales tiene un valor importante para el público en general, y como estas críticas afectan el comportamiento del consumo de la audiencia, una crítica buena o positiva lograra atraer a más personas que una crítica mala o negativa.

Aumentar entonces el flujo de espectadores en la taquilla se vuelve primordial, y en este sentido, S. Follows en [2] afirma que el 49% de las películas no logran generar una

ganancia; si bien esta cifra puede parecer incorrecta, lo cierto es, que este fenómeno no es extraño en la industria del cine. Para entender el porqué de esto, S. Follows en [3] describe la distribución de los gastos asociados a una película con un presupuesto de más de 100 millones de dólares; allí se puede apreciar que, para lograr una ganancia se debe, como mínimo, recaudar el doble del dinero invertido tanto en la producción del negativo (la película en sí), como en las campañas de marketing, costos contractuales, etc. Esta situación se hace especialmente compleja para producciones más limitadas o independientes, que no cuentan con el apoyo de un estudio de producción grande, con una distribución mucho más pequeña y una campaña de publicidad prácticamente nula.

Con el propósito de entender que hace a una película ser un éxito o un fracaso, el estudio presentado por Saurabh Kumar et al. en [4], realiza un análisis exploratorio, en donde se analizan diferentes componentes de una producción cinematográfica, como por ejemplo, el director, el director de fotografía, el elenco, si hay presencia de actores nominados o ganadores de un premio Oscar, también se incluyeron los puntajes dados por los críticos profesionales, el puntaje dado por el público y el puntaje ponderado dado por la plataforma de IMDb (Internet Movie Database, por sus siglas en inglés, esta página asigna un puntaje a cada película que resulta de la combinación de los puntajes dado por los críticos profesionales y el público), puntajes que pueden ser muy diferentes entre sí. El estudio mencionado estableció que la presencia de uno u otro director, e incluso si la película cuenta en su reparto con actores ganadores del premio de la academia, entre otros factores, no permite concluir que una película en particular sea un éxito gracias exclusivamente a uno de estos elementos, más aún, no encontró que estos aspectos (director, director de fotográfica, actores) sean un componente de peso para que una cinta sea un éxito en taquilla. No obstante, si se pudo observar, que las películas con un puntaje alto o favorable por parte de los críticos y de IMDb, tenían una mayor probabilidad de ser un éxito. Resultado similar se pudo encontrar con el puntaje dado por el público.

Es con estos antecedentes, que las investigaciones se han enfocado en utilizar diferentes métodos, estrategias, arquitecturas, etc., para encontrar las partes de la crítica (características del mensaje) que mejor describen una crítica positiva y una negativa. Las investigaciones descritas en [5] [6] [7], han implementado métodos de selección como la correlación de Pearson, ensamble features, bag of words (BoW) y ganancia de información

(information gain); estos en combinación con el método de clasificación de Naive-Bayes, dan como resultado valores de accuracy entre el 80 y el 90%, es válido anotar que estos estudios se realizaron sobre tweets en hindi, y para una cantidad variable de comentarios. Dentro de los hallazgos encontrados, se puede mencionar que, para hacer la clasificación correcta de una crítica, un texto más extenso resultará en una clasificación más acertada. También, pudieron concluir que cuando se realiza una selección de palabras mediante la estrategia de bag of words, los resultados de la clasificación mejoran. Sin embargo, técnicas como las redes neuronales recurrentes no fueron implementadas, dejando de lado sus beneficios.

Buscando mejorar el acierto en la clasificación desde la perspectiva de la selección de características se han implementado métodos como el N-grams, bag of words, Word2Vec, utilizados en [8] [9] [10], consiguiendo porcentajes de accuracy de poco menos del 80%. Estos estudios, además, utilizaron como clasificador técnicas como soporte de máquinas vectoriales, Naive-Bayes, e incluso redes neuronales recurrentes, en [10] también se implementó máquinas de soporte vectorial y regresión logística, sin embargo, sus resultados no fueron mejores que los vistos anteriormente. Al igual que en las investigaciones anteriores, estos estudios tampoco implementan técnicas que permitan entender el contexto de las características seleccionadas, como se menciona en [11], el manejo de la negación es un aspecto importante y que puede mejorar los resultados de la clasificación del mensaje.

Con la llegada de nuevas técnicas de aprendizaje de máquina, tal como se vió en [10], se da la implementación de diferentes tipos de redes neuronales tanto en la selección como en la clasificación, es así como en [12] [13], se utilizaron redes neuronales convolucionales para la etapa de clasificación, y para la selección de características se utilizaron también redes convolucionales y una combinación de redes convolucionales y recurrentes (en [12]). Estas combinaciones mejoran los resultados en la clasificación subiendo los porcentajes a entre el 91% y el 95%. Sin embargo, no se presentan estrategias para el manejo de las variaciones en la interpretación de los mensajes, como se describe en [11], en donde se planea una estrategia para el manejo de la negación de las reseñas.

Es de notar que en ningún caso se hace referencia a estrategias combinadas, es decir, selección de características y técnicas de aprendizaje de máquina, que permitan

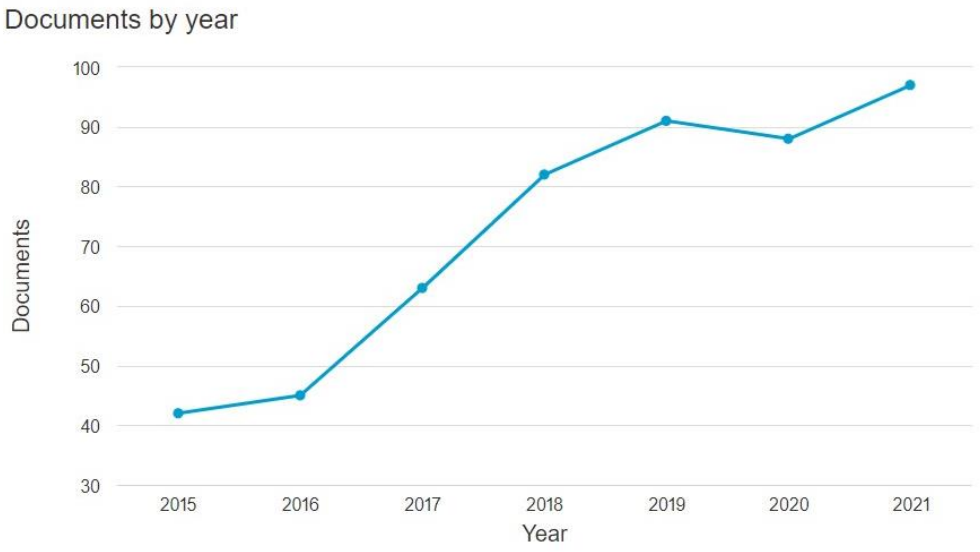
robustecer el proceso de la clasificación, así como tampoco se contemplan las variaciones de interpretación de los mensajes, configurándose de esta forma, un campo de investigación amplio, en donde hay cabida para la experimentación en la implementación de nuevas técnicas, mejorando el desempeño en la precisión de la clasificación.

De este modo, se ha encontrado en la literatura diferentes estrategias, diferentes aproximaciones de técnicas para la selección y clasificación de las reseñas de películas, por lo cual es la intención de este trabajo explorar técnicas combinadas, tanto de selección como de clasificación para aportar claridad sobre aquellas que arrojen los mejores resultados. El estudio de este tipo de datos supone para el área de investigación de analítica de sentimientos, un escenario ideal, en donde la fuente de datos es directamente recolectada de los críticos y de la audiencia, permitiendo la implementación de nuevas y novedosas técnicas de análisis, cuyos resultados contribuirán a un mejor entendimiento de las técnicas más apropiadas para este tipo de análisis. Además, se podrá obtener información sobre como la audiencia entiende las reseñas de una película, y como esta afecta su decisión de ver o no el filme.

1.1 Revisión sistemática

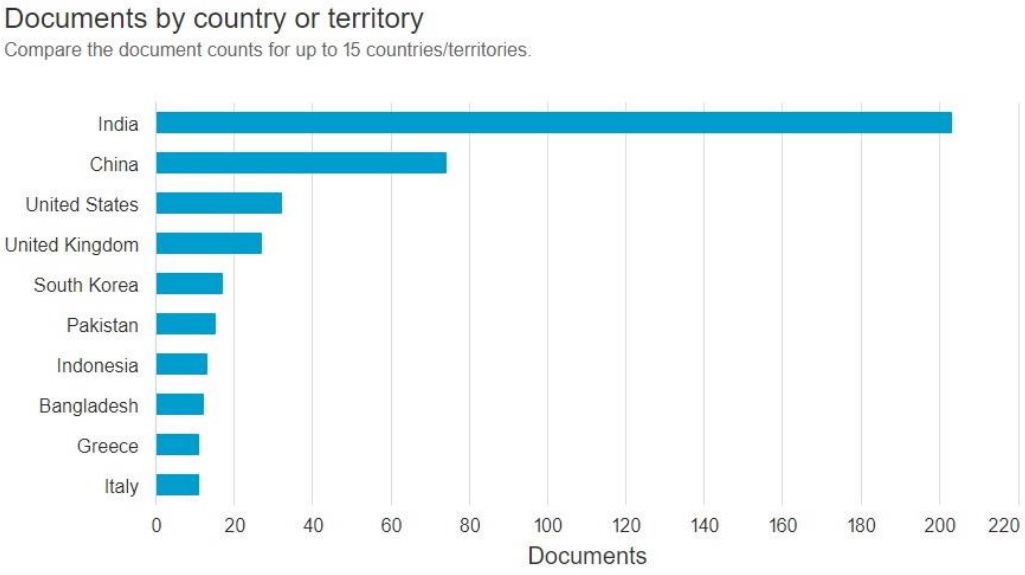
Como se expuso anteriormente, aumentar las ganancias en taquilla de las películas estrenadas alrededor del mundo, es de mucha importancia por lo que esto representa para las casas productoras; es así como se ha observado un incremento del interés en el área de la analítica de sentimientos y su impacto en el desempeño en la taquilla de producciones cinematográficas, que hoy en día y gracias a la expansión de las redes sociales, han generado un efecto importante en la audiencia potencial y que ha llamado la atención de las casas productoras; esto fenómeno, el cómo la opinión subyacente de una crítica impacta la recaudación en taquilla, llevo al aumento de estudios enfocados en dicha situación. La Figura **1-1**, muestra como han aumentado la cantidad de estudios con temáticas relacionadas a la tratada en esta investigación.

Figura 1-1: Variación en el tiempo de estudios realizados en el área de interés de la investigación.



Ahora bien, en la Figura **1-2**, muestra que uno de los países en donde más investigación de este tipo se ha realizado es en la India, dato que se evidencia en la cantidad de documentos que utilizan base de datos en idioma Indi.

Figura 1-2: Cantidad de estudios realizados por país en el área relacionada a la investigación.



Con el interés que se ve reflejado a través del aumento de las publicaciones con temas relacionados, se plantearon diferentes preguntas de investigación, que en esencia guiaron todo el proceso de esta revisión sistemática.

1.1.1 Preguntas de investigación

Las preguntas de investigación que se relacionan a continuación sirvieron como guía para la realización de la revisión sistemática de literatura en el área de interés de esta investigación.

- ¿Existe una relación entre la taquilla de películas de cine y las críticas que recibe?
- ¿Qué técnicas de procesamiento de texto son las más empleadas para el análisis de críticas de cine?
- ¿Cuáles son los principales aspectos tratados en los estudios realizados acerca del análisis de críticas de cine?
- ¿Se han analizado en los estudios previos el efecto de la selección de características frente a la precisión de la estimación?

1.1.2 Proceso de búsqueda

El proceso de búsqueda de las referencias bibliográfica se realizó de forma manual; estos a su vez fueron revisados en su totalidad con el fin de garantizar que sus métodos, resultados y conclusiones, estuvieran enmarcados en alguna de las preguntas formuladas en esta investigación. Se procuro, además, que el idioma de las referencias revisadas estuviera en idioma inglés.

1.1.3 String de búsqueda

El *string* de búsqueda empleado en el proceso de búsqueda se describe a continuación; en ambos motores de búsqueda se filtraron los años de 2015 a 2021:

- Para la plataforma Scopus:

  TITLE-ABS-KEY ("movies reviews" OR "film reviews") AND TITLE-ABS-KEY ("sentiment analysis" OR "polarity detection") AND PUBYEAR > 2014 AND PUBYEAR < 2022

- Para el motor de búsqueda de Google Scholar:
  Sentiment analysis in movie reviews

1.1.4 Palabras clave

Algunas de las palabras claves que se buscó estuvieran presentes dentro de los artículos evaluados fueron: *movies, sentiment analysis, movie reviews, machine learning*.

1.1.5 Criterios de calidad

Para evaluar la calidad de los documentos encontrados se establecieron los siguientes criterios:

- ¿Los métodos implementados están enmarcados en el campo de investigación planteado por las preguntas de investigación?
- ¿Se describe la metodología detallada de los métodos implementados?
- ¿Las conclusiones presentadas responden o contribuyen a la respuesta de alguna de las preguntas de investigación planteadas?
- ¿Los documentos plantean discusiones sobre los resultados obtenidos y métodos implementados?

Basado en los criterios anteriores, se plantearon las siguientes respuestas: **Si**, **No** y **Parcial**. Donde **Si** equivale a un puntaje de 1, **No** a un puntaje de 0 y **Parcial** a un puntaje de 0.5; siendo así, el puntaje máximo de calidad que puede ser obtenido por un documento es de 4.

1.1.6 Criterios de inclusión y exclusión

A continuación, se describen los criterios de inclusión y exclusión establecidos para la revisión sistemática de literatura:

- Artículos publicados en revistas indexadas o procedimientos expuestos en conferencias, con una antigüedad no mayor a 5 años.
- Artículos cuya área de estudio estuviera relacionada con el problema de investigación planteado por esta investigación, así como también, que dieran

respuesta a una o varias de las preguntas de investigación planteadas en el literal **1.1.1** de este documento.

- Artículos cuyos procedimientos y metodologías estuvieran claramente descritos; así como sus objetivos, resultados y conclusiones.

Es de aclarar que literatura que no cumpliese con estos tres criterios aun pudo ser incluida como parte de los referentes de esta investigación (como en el caso de los artículos de la tabla **1-3**), esto después de evaluar que su contenido sustentaba y contextualizaba esta investigación; cabe anotar, además, que el idioma de las bases de datos utilizadas por los respectivos estudios no fue considerado como criterio de exclusión.

Como parte de los criterios para incluir o excluir un documento, se consideró el puntaje de los criterios de calidad de los documentos seleccionados, siendo este el último filtro para incluir o excluir un documento, es así, que solo aquellos documentos con un puntaje superior a 3 fueron finalmente incluidos en la revisión.

Por último, se dio por excluido un documento cuando no cumpliera con los criterios de inclusión y su puntaje en los criterios de calidad fuera menor a 3.

1.1.7 Documentos seleccionados

La Tabla **1-1** muestra los documentos seleccionados que cumplieron con los criterios descritos en el literal **1.1.6** y que fueron arrojados por las plataformas de búsqueda.

Tabla 1-1: Documentos filtrados por los string de búsqueda en las plataformas de búsqueda.

Id	Titulo	Idioma - BD
1	How critical are critical reviews? The box office effect of films critics, star power and budgets	Inglés - IMDB
2	A data mining approach to analysis and prediction of movie rating	Inglés - IMDB
3	Movie Master: Hybrid Movie Recommendation	
4	A Data mining Technique for Analyzing and Predicting the success of Movies	Indi - IMDB

5	A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis	Inglés - IMDB, Amazon
6	A Technique to Handle Negation in Sentiment Analysis on Movie Reviews	Inglés - IMDB
7	A Word Vector based Review Vector method for Sentiment Analysis of Movie Reviews Exploring the applicability of the Movie Reviews	Inglés, Mandarín - IMDB (solo ingles)
8	Deep Learning Based Sentiment Analysis Using Convolution Neural Network	Indi
9	Do consumer and expert reviews affect the length of time a film is kept on screens in the USA	Ingles - Box Office Mojo, Rotten Tomatoes (2004 - 2015)
10	Feature Extraction and Classification of Move Reviews	Inglés - IMDB
11	Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet	Inglés - Rotten Tomatoes
12	Sentiment Analysis from Movie Reviews Using LSTMs	Inglés - IMDB
13	Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Featured Selection	Twitter
14	Twitter Sentiment Analysis of Movie Reviews Using Ensemble Features Based Naïve Bayes	Indonesian - Twitter
15	Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier	Twitter
16	An Improved Study of Multilevel Semantic Network Visualization for Analyzing Sentiment Word of Movie Review Data	Inglés - NAVER
17	Exploring contextual factors from consumer reviews affecting movie sales: an opinion mining approach	Inglés - IMDB
18	Movie Reviews Sentiment Analysis and Classification	Inglés - IMDB
19	Movie Success Prediction Using Data Mining	Inglés - IMDB (Kaggle)
20	An Ensemble method with sentiment features and clustering support	Pang & Lee movie reviews Stanford Twitter Sentiment Stanford Treebank

		IMDB
21	Sentiment Analysis on Movie Reviews	Ingles – IMDB
22	Effective Approach for Sentiment Analysis on Movie Reviews	Indi – Twitter API
23	Sentiment Analysis Using Gini Index Feature Selection, N-Gram and Ensemble Learners	No se especifica

1.1.8 Recolección de datos

Los datos recolectados de cada estudio revisado se presentan a continuación:

- Fuente (artículo, publicación, procedimiento de conferencia)
- Año de publicación
- Base de datos utilizada (idioma y fuente)
- Métodos de selección implementados
- Métodos de clasificación implementados
- Desempeño (métricas)

1.1.9 Evaluación de calidad

De acuerdo con la información recolectada de cada fuente y basado en los criterios de calidad descritos, cada documento revisado obtuvo el puntaje que se muestra en la Tabla **1-2**.

Tabla 1-2: Puntaje de calidad para los documentos revisados.

Fuente	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Puntaje
1	0	0	1	1	2
2	1	0	0	0	1
3	0	0	0	0	0
4	1	1	0	0	2
5	1	1	1	1	4

6	1	1	1	1	4
7	1	1	1	1	4
8	1	1	1	1	4
9	0	0	0	1	1
10	1	1	1	1	4
11	1	1	1	1	4
12	1	1	1	1	4
13	1	1	1	1	4
14	1	1	1	1	4
15	1	1	1	1	4
16	1	0	0	1	2
17	0	1	1	0	2
18	1	1	1	1	4
19	0	1	1	1	3
20	1	1	1	1	4
21	1	0	0	0	1
22	1	0	0	0	1
23	1	0	0	0	1

Una vez filtrados los documentos con base a su puntaje de calidad, se seleccionaron los relacionados en la Tabla **1-3.**

Tabla 1-3: Documentos seleccionados como referencias para la investigación.

Fuente	Titulo	Puntaje
1	How critical are critical reviews? The box office effect of films critics, star power and budgets	2

5	A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis	4
6	A Technique to Handle Negation in Sentiment Analysis on Movie Reviews	4
7	A Word Vector based Review Vector method for Sentiment Analysis of Movie Reviews Exploring the applicability of the Movie Reviews	4
8	Deep Learning Based Sentiment Analysis Using Convolution Neural Network	4
10	Feature Extraction and Classification of Move Reviews	4
11	Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet	4
12	Sentiment Analysis from Movie Reviews Using LSTMs	4
13	Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Featured Selection	4
14	Twitter Sentiment Analysis of Movie Reviews Using Ensemble Features Based Naïve Bayes	4
15	Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier	4
18	Movie Reviews Sentiment Analysis and Classification	4
19	Movie Success Prediction Using Data Mining	3
20	An Ensemble method with sentiment features and clustering support	4

1.2 Revisión y discusión

Como resultado del proceso de revisión sistemática se seleccionaron los documentos que se relacionan en la Tabla **1-3**. De la revisión de estos, se evidenció que la mayoría tenía como objetivo principal determinar la polaridad de la crítica analizada, objetivo que está alineado con el propósito de esta investigación; sin embargo, también se seleccionaron

otros artículos que soportan desde una perspectiva más descriptiva, la relación entre diferentes características de la crítica y el desempeño de la película en la taquilla.

La opinión de expertos es de suma importancia en muchos campos de la industria, estas llegan hoy en día a muchas más personas en el mundo, y es así como una opinión puede llevar, por ejemplo, a subir o bajar el precio del petróleo hasta el punto de llevar empresas a la quiebra; el mercado de valores también se ve afectado por comentarios y opiniones en las redes sociales; la industria cinematográfica no es la excepción. El estudio expuesto en [14], se enfoca en 3 objetivos principales: 1. Como se afecta el desempeño de una película dado el rol (y su opinión) de un crítico; 2. El efecto de un critica negativa versus una crítica positiva; y 3. Como la presencia de ciertos actores y el presupuesto, disminuye el impacto negativo en la taquilla. La base de datos utilizada por los investigadores almacenaba registros de 200 películas, algunas de las variables incluidas eran la crítica hecha sobre la película, la ganancia por semana, polaridad de la crítica, presupuesto, poder del reparto, entre otras. Para realizar este análisis los autores realizaron un estudio de correlación entre las variables mencionadas.

Para responder al primer interrogante los investigadores describen dos roles bajo los cuales la opinión de un crítico se puede analizar; uno es el rol de influenciador, esto es el poder de influenciar al público de ir a ver la película en algún momento durante su estadía en los cines; el segundo rol que un crítico puede ejercer es el del predictor, es decir, que a través de su opinión se pueda predecir el éxito en taquilla de la película. El resultado del estudio sobre este aspecto muestra que, si bien ambos roles tienen un impacto en la taquilla, especialmente en las primeras semanas de exhibición, los críticos presentan un mejor desempeño bajo el rol de predictores, es decir, con base en sus opiniones se puede predecir de manera confiable el éxito en taquilla de la película, sin embargo, su opinión puede no ser tan convincente como para influenciar al público de ir a ver la película.

Con respecto al aspecto número 2, el estudio concluye que ambos tipos de críticas (positivas y negativas) tienen un efecto en la taquilla de la película. No obstante, el efecto de ambas es diferente, las críticas negativas tienen un impacto mayor en la taquilla que el impacto de críticas de positivas, esto se podría traducir a que las criticas positivas no contribuyen significativamente en el éxito de una película.

Sobre el aspecto número 3, se concluye que los actores presentes en la película y el presupuesto tienen un efecto pequeño en películas con criticas mayormente positivas, sin embargo, mitigan fuertemente el impacto de críticas negativas.

Siguiendo con el análisis exploratorio y descriptivo para identificar posibles relaciones entre diferentes aspectos de una película, el estudio realizado en [4] contó con un aproximado de 651 películas, datos recolectados de IMDb, en donde se incluyen variables como: puntaje dado por IMDb, puntaje dado por los críticos, el puntaje de la audiencia, director, actores, titulo, estudio, genero, entre otros. Para realizar este análisis se realizó un análisis de dispersión entre las diferentes variables.

En este estudio se evidenció que existe una correlación fuerte (valor de 0.7) entre el puntaje dado por un crítico y el puntaje dado por los usuarios, es decir, que la audiencia y los críticos tienden a calificar de la misma forma las películas, esto fortalece el rol de predictor del crítico, basado en el hecho que para este estudio en particular el éxito de una película se basa en su popularidad, una mayor popularidad se traduce en una mejor taquilla, es decir que un crítico puede correctamente predecir el éxito de la película.

Sobre si la presencia de actores, directores, entre otros, con un premio de la academia de cine (Oscar) o con alguna nominación, no necesariamente por la película en particular, puede o no predecir el éxito de una película, la investigación encontró una débil correlación entre estos. Por lo que, si bien un Oscar es símbolo de prestigio dentro de la industria, un actor, director, etc., necesitaría de otro incentivo para congregar una mayor cantidad de público. Continuando en este apartado, y soportado en el prestigio de un premio Oscar, el estudio continúo analizando el caso en el cual la película hubiera conseguido una nominación o hubiese ganado un premio de la academia; se concluyó que no existe un efecto contundente sobre el éxito de la película. También se puede concluir, y no sería de extrañar, que las películas con algún tipo de nominación o premio ostentan mejores criticas tanto por parte de la crítica especializada como de la audiencia.

Otros análisis como la relación entre el puntaje de la crítica con la duración de la película, o con el formato de lanzamiento (teatro, dvd, etc.), o con el género de la película; no presentaron correlaciones fuertes, por el contrario, el comportamiento del puntaje con relación a esas variables no presenta mayor dispersión, y se presentan un sesgo hacia la

derecha, es decir, generalmente se entregan calificaciones altas para todas las películas, géneros, actores, etc.

Una metodología que involucra en el proceso de selección de características la correlación de Pearson es la propuesta en [5]; el análisis propuesto en esa investigación se realizó sobre unos 500 tweets sobre críticas de películas que se caracterizaron mediante dos técnicas: características de ensamble y Bag of Words. Con estas técnicas se descompone el tweet en diferentes partes como verbos, pronombres, artículos, etc., estas son llamadas partes del discurso, y con ellas se crean bolsas de palabras que comparten significados similares o que su utilización en un texto puede tener el mismo significado; también se puede incluir en esos grupos de palabras, características específicas del medio, como hashtags, emoticones, estilos de escritura (negritas, subrayados, mayúsculas, etc.), y las ya mencionadas partes del discurso. Es a partir de estos grupos de palabras de los que la correlación de Pearson selecciona aquellos que tendrían un mejor resultado al momento de clasificar la polaridad del tweet. Como métodos de clasificación se implementaron diferentes modelos de clasificadores bayesianos como: Bernoullie Naïve Bayes para datos binarios, Gaussian Naïve Bayes para datos continuos y Naïve Bayes multinomial para datos numéricos. Dentro de los resultados presentados por el estudio, con una selección de características limitada al 20% del total de características disponible, se obtuvo un accuracy del 82%, una precisión del 86% y un recall del 79.62%. Como se puede ver, los valores en esa etapa del experimento podrían estar algo por debajo de otros estudios, sin embargo, no son despreciables en ningún sentido. Ahora bien, los investigadores realizaron una estandarización manual de los tweets que incluyó corregir errores de ortografía, eliminar palabras con el mismo significado (ej: malo, terrible, pésimo) con esto limitando la redundancia, entre otras tareas. Una vez realizado este proceso, los investigadores observaron un aumento de un 8% en los resultados obtenidos. Si bien este proceso manual mejoró el desempeño del modelo propuesto, es una desventaja para cualquier modelo tener dentro de una de sus etapas un proceso manual crítico para el resultado, esto dado que una persona podría identificar más o menos errores de ortográfica o de redundancia, por lo cual la clasificación final podría terminar sufriendo de la subjetivad de la persona que realizó este proceso. Esta desventaja es también anotada por los investigadores, quienes recalcaron las dificultades que representan los mensajes escritos (ej: ortografía, sinónimos), por lo cual presentan como el mejor resultado aquel obtenido antes de realizar el proceso de estandarización manual. Es entonces pertinente que para

trabajos futuros se tengan presente esas observaciones y con ello desarrollar modelos que implementen esas tareas de estandarización de forma automática. Es también un campo por explorar en esa investigación nuevas técnicas de selección y clasificación que se han probado como muy efectivas en estas tareas como las redes neuronales, o las máquinas de soporte vectorial.

Con una metodología similar a la de la metodología anterior, la investigación expuesta en [7] realiza una combinación de características, conocidas como ensamble de características, en esta técnica fueron combinados diferentes tipos de características: características específicas del tweet (ej: hashtags, emoticones, urls), características del texto (ej: longitud del texto, longitud promedio de una palabra, cuantas mayúsculas, cuantas minúsculas, signos de exclamación, signos de interrogación, entre otros), características propias del lenguaje (conocidas como partes del discurso, ej: verbos, pronombres, adjetivos, etc.), características basadas en lexicones (cantidad de palabras conocidas como positivas o negativas presentes en los tweets), y bag of words (bolsa de palabras). Esta combinación de características es entregada un clasificador bayesiano encargado de etiquetar los tweets entre positivos o negativos. La base de datos utilizada por este estudio estuvo compuesta por 500 tweets, con 250 positivos y 250 negativos. Todos los tweets estaban en indonesio. Esta investigación realizó varios experimentos en donde se cambió el conjunto de características manteniendo siempre el mismo clasificador. Es así, que el mejor resultado fue obtenido por las características de bag of words con una precisión del 96%, le siguió la combinación de todas las características con un 91%. Se puede observar en el artículo que cuando se utilizó como entrada de análisis tipos de características como las específicas de twitter o las partes del discurso, se obtuvieron precisiones del 40% y el 56%, estos dos tipos de características están relacionadas por la misma naturaleza del tweet, en donde muchas personas prefieren utilizar otras formas de expresarse diferentes al texto escrito tradicional, por lo cual una mala caracterización de estas nuevas formas pudo derivar en el pobre desempeño de estos dos tipos de características, y de igual forma, estos pudieron tener un impacto en la combinación de todos los tipos de características reduciendo su efectividad para clasificar. Si bien la comparación entre las diferentes combinaciones de características brinda una visión ampliada de las características de un mensaje que pueden resultar más valiosas en futuros métodos, contar solamente con un método de clasificación limita la investigación pues no

es posible ver el efecto de las características en técnicas de clasificación mucho más robustas y con un buen desempeño al momento de clasificar.

La investigación expuesta en [6], continua con técnicas basadas en el clasificador bayesiano, este estudio se realizó sobre una base de datos de tweets compuesta por 8231 mensajes sobre 12 películas elegidas manualmente bajo el criterio de su popularidad en el 2016. Una vez recolectado esta información, se realizaron diferentes tareas de preprocesamiento como la tokenizacion (proceso en el cual cada frase es separada en partes individuales del mensaje), la eliminación de stop words (proceso de eliminar palabras con poco valor como artículos, conectores, preposiciones, etc.), conversión de verbos a su raíz (proceso conocido como Stemming); una vez realizado este proceso los mensajes son enviados a un selector de características basado en la técnica de ganancia de información, esta técnica se fundamenta en el cálculo de la entropía, esta mide el "caos" dentro la información, se busca obtener la menor entropía posible y con esto la máxima cantidad de información. El estudio establece una ganancia de información del 0.006. Continuando con la metodología propuesta, se pasó a clasificar el tweet con un clasificador Naïve Bayes. Con esta configuración, el estudio encontró un accuracy del 82.16%. Como se puede ver este estudio no compara con otras técnicas ni de selección ni de clasificación lo cual limita la generalización de los resultados. Es también una debilidad del estudio haber seleccionado ciertas películas en particular, esto puede derivar en un sesgo si todas las películas seleccionadas obtuvieron buena o mala recepción por parte del público, ciertamente la popularidad de una película no es reflejo fiel de la opinión que el público o la crítica tengan sobre ella. Este estudio también concluye que esta metodología logró mejores resultados al identificar tweets neutrales, efecto que pudo ser causado de una selección poco cuidadosa de las películas a analizar.

Con la misma idea de que un proceso de selección de características produce un mejor desempeño al momento de la clasificación, el estudio presentado en [8] propone un método basado en la técnicas de Word2Vec, en esta técnica una frase es descompuesta en cada una de sus palabras y con estas se forma un vector; como se cree este vector dependerá de una técnica especifica, así, el estudio implementó la técnica de Skip-diagram, esta técnica permite predecir palabras alrededor de otras palabras o párrafos, esto lo que significa de forma general es, que a partir de una palabra o una frase o párrafo, Skip-diagram es capaz de encontrar las palabras que complementan a esa palabra, frase o párrafo. Este vector se obtiene del corpus de todas las palabras dentro de la base de datos,

este es llamado el word vector. Ahora bien, el estudio propone un review vector, que proviene de las reseñas que se van a analizar, es decir, del texto completo de las reseñas (no separadas). Luego, se mide la distancia entre cada review vector y cada grupo de palabras del word vector. Una vez que estos vectores son definidos, son enviados al clasificador para la detección de su polaridad. Para este estudio se implementaron diferentes clasificadores, entre ellos: soporte de máquinas vectoriales, regresión logística y redes neuronales. Con estas combinaciones (word2vec y los diferentes clasificadores), los investigadores realizaron diferentes experimentos:

- El primero fue identificar la influencia que tiene la cantidad de palabras dentro de un word vector para lograr una clasificación más adecuada. Para esto se seleccionaron vectores compuestos por 100, 200, 300 y 400 palabras. Si bien se podría pensar que una mayor cantidad de palabras podría significar un desempeño mayor, este experimento evidencia que esto no es así. De hecho, para todas las combinaciones de métodos, no hubo una que destacara por encima de las demás, sumado a esto, el porcentaje de accuracy estuvo un poco por encima del 77% pero no mayor al 87%. Este experimento solo se realizó con clasificador de soporte de máquinas vectoriales.
- El segundo experimento consistió en determinar la influencia de la longitud promedio de una reseña en la clasificación. Este experimento dio como resultado desempeños similares al experimento anterior, sin embargo, el menor accuracy se dio alrededor del 80%, por lo cual la longitud de la reseña claramente mejora el desempeño de la clasificación. Para la realización de este experimento se utilizaron las mismas longitudes del experimento 1 y con el mismo clasificador.
- El tercer experimento busco identificar la influencia del idioma en el método propuesto, en este experimento se compararon todos los clasificadores mencionados con anterioridad. Aquí concluyeron que para ambos idiomas utilizados en el estudio (inglés y chino-mandarín), el método presenta buenos resultados. No obstante, el accuracy no supero el 86.1%, la combinación que mejor resultados presentó fue la compuesta por un clasificador de soporte de máquinas vectoriales y con un word vector de 300 palabras.
- Este último experimento consistió en utilizar el tamaño de las reseñas completo y entender con esto el efecto que esto podría tener en la clasificación. Esto dio como resultado un accuracy del 87.6%. Identificando nuevamente a la combinación con

el clasificador de soporte de máquinas vectoriales como el que mejor desempeño presento.

Este estudio presenta puntos de vista de análisis muy interesantes, al tener en cuenta la longitud de las reseñas, la cantidad de palabras en las reseñas, el idioma; estos aspectos han sido notoriamente marginados en los estudios descritos con anterioridad, y claramente pueden ser aspectos que mejoren o empeoren el rendimiento del método propuesto si se toma a la ligera. Si bien el desempeño descrito por el método propuesto por este estudio no obtuvo porcentajes de accuracy superiores o considerablemente mejores que otros estudios, es claro que tener en cuenta aspectos como los ya mencionados representan un punto de exploración importante para el desarrollo de una investigación de este tipo. El estudio finalmente concluye que el método es los suficientemente efectivo bajo las condiciones establecidas y con las técnicas propuestas, proponen en futuros trabajos, realizar una exploración al área de aprendizaje profundo para mejor el desempeño obtenido.

Continuando en el campo de la selección de características, la investigación expuesta en [9] basa su metodología en comparar diferentes métodos de selección de características acompañándolos con un método supervisado de aprendizaje de máquina. Los métodos de selección implementados fueron Bi-grams, TF-IDF, y wordcount vector. La técnica de selección Bi-gram es una técnica en donde se crean grupos de palabras, similar a como lo hace bag of words, sin embargo, en Bi-grams los grupos de palabras están compuestos solo por dos palabras. TF-IDF (Term Frequency - Inverse Document Frequency), es una técnica que refleja la importancia de una palabra dentro del documento. Una palabra que es frecuentemente utilizada dentro de un texto podría no ser tan importante dado la alta frecuencia con la cual aparece (ejemplo de estos casos pueden ser los artículos: “el”, “la”, entre otras). Por último, la técnica de wordcount vector, consiste en contar las veces que una palabra parece en un documento, con esto sería posible comparar dos documentos de acuerdo con la cantidad y frecuencia de palabras que aparecen en ambos. Para clasificar estas características se utilizaron 3 clasificadores: Naïve Bayes multinomial, soporte de máquinas vectoriales y random forest. Con estas técnicas, y analizando una base de datos compuesta por 25000 reseñas de películas recolectada por la universidad de Stanford, los resultados muestran valores de accuracy entre el 84% y el 88%; siendo las mejores combinaciones las compuestas por TF-IDF y el clasificador de soporte de

máquinas vectoriales, y Bi-grams en conjunto con el clasificador Naïve Bayes multinomial. Este estudio concluye resaltando la importancia de la elección de los métodos para selección y para clasificación, propone como futuros trabajos continuar la exploración de más técnicas para realizar las dos tareas en búsqueda mejorar los resultados obtenidos. Como se menciona en las conclusiones de este artículo, faltó hacer una recolección más amplia de métodos tanto de selección como de clasificación, pues se dejan por fuera técnicas como las redes neuronales que han mostrado, como se ha visto con anterioridad, excelentes resultados al momento de clasificar.

Una técnica conocida como SentiWordnet, es la base del modelo propuesto en [15], esta técnica que se asemeja a un árbol, una palabra es colocada en la raíz del árbol, esta palabra es especial pues debe de ser una palabra que tenga un significado amplio, de forma tal que otras palabras puedan ir definiendo su significado más específico. Estas palabras que van cercando el significado de la palabra raíz, van creando ramas dentro del árbol. El árbol se va definiendo de forma tal que a cada palabra se le va asignando un puntaje que al final es el que determina si el mensaje que se quiere clasificar es positivo o negativo. Este modelo también implementa un preprocesamiento estándar (tokenizacion, lematización, remoción de stop words, etc.), y una vez surtidos estos dos procesos (primero el preprocesamiento y luego la selección de características con SentiWordnet), el vector resultante es entregado al algoritmo de regresión logística para su posterior clasificación. Con esta configuración, se obtuvieron resultados de precisión, recall y F-score iguales a 97%. Aclarando que la base de datos utilizada en esta investigación fue obtenida del agregador de reseñas de cine, Rotten Tomatoes. El estudio concluye que el buen desempeño de este modelo se debe en parte a la inclusión de otras características como el puntaje dando por expertos (en Rotten Tomatoes, los expertos además de publicar sus criticas también pueden asignar un puntaje numérico a la película) e incluso el puntaje entregado por SentiWordnet. Esto podría haber derivado en una redundancia de características pues el puntaje de estos puede estar estrechamente correlacionados con la polaridad de la reseña, lo que deja en segundo plano el análisis del texto y que solo con este se pueda determinar su polaridad. El estudio compara sus resultados con SentiWordnet operando en solitario al igual que con el puntaje de expertos, estos métodos obtuvieron 51% y 91% respectivamente, para las métricas mencionas. Se puede ver entonces que la inclusión de estas características al modelo pudo influir en el excelente

desempeño del modelo propuesto por la fuerte correlación entre los puntajes y la polaridad de la reseña, no obstante, esta observación no es identificada por los autores.

Un modelo propuesto para la evolución del proceso de clasificación con técnicas tradicionales es expuesto en [16], este estudio subraya la importancia de este proceso dentro de cualquier modelo, por lo cual, la elección del método para realizar esta función no puede ser dejado a ligera. El modelo propuesto en esta investigación compara varios clasificadores, entre los que se menciona: Naïve Bayes, arboles de decisión, soporte de máquinas vectoriales, redes bayesianas, vecinos más cercanos (KNN), random forest, Ripper Rule Learning y gradiente estocástico descendiente. El proceso de preprocesamiento aplicado estuvo compuesto por tokenizacion, lematización, stemming, remoción de stop words. El proceso de selección de características se basó en Gain Ratio. Al evaluar todas las combinaciones se llegó a que el mejor resultado de clasificación fue obtenido por el clasificador random forest con un 96% de accuracy, seguido de cerca por vecinos más cercanos y arboles de decisión, con 92.9% y 91.28% respectivamente. Los resultados de este estudio son muy representativos pues son comparables con los resultados entregados por otras técnicas más avanzadas, por lo cual se puede cuestionar que estrategia sería la más adecuada y que proceso de preprocesamiento y selección de características son los más adecuados en técnicas más avanzadas (como las redes neuronales que se expondrán a continuación).

Ahora bien, desde la perspectiva de mejorar la clasificación del mensaje e implementando técnicas más novedosas de clasificación, el estudio descrito en [10] propuso utilizar una variante de las redes neuronales recurrentes, conocida como redes neuronales de memoria de corto plazo. La idea de esta técnica es que en muchos casos las cadenas de texto pueden ser muy largas, por lo cual el diseño de una red neuronal normal conlleva a que se pierda información en el proceso, pues no se retiene ni se conecta información previamente procesada. Este estudio realizo un proceso de selección de características basado en word embeddings, más específicamente en word2vec y skip-gram, con estas características seleccionadas; una vez obtenido el vector de características se construyó un modelo de red neuronal de memoria de corto plazo (LSTM). Este estudio implementó diferentes arquitecturas para identificar la polarización de la reseña, sin embargo, se mantuvieron las capas de la red, variando los hiperparametros, se crearon 3 capas: una capa de entrada (que recibe el vector de características encontrado previamente), una capa LSTM y una capa dense (capa de tipo softmax en donde al final se obtiene la

probabilidad de que la reseña sea positiva o negativa). Los hiperparametros variaron de la siguiente forma:

- Épocas: Se realizaron los experimentos con 3, 10 y 50 épocas.
- Unidades en la capa LSTM: 50, 100 y 200 unidades
- Longitud del vector de entrada: 500 y 1000

Dentro de los resultados obtenidos se evidencia que, la arquitectura compuesta por un entrenamiento de 50 épocas, una cantidad de unidades en la capa LSTM de 100 y una longitud del vector de entrada de 500, se obtuvo un accuracy del 88.46%, siendo este resultado el mejor desempeño obtenido por todas las arquitecturas. El estudio señala que cuando el número de unidades en la capa LSTM se aumenta a 200 o cuando la longitud del vector de entrada aumenta a 1000, se presenta una caída en el desempeño, que los investigadores atribuyen a un problema de sobreajuste. El estudio compara sus resultados con el desempeño de otros clasificadores como los soportes de máquinas vectoriales, regresión logística, un perceptrón multicapa y una red neuronal convolucional, sin embargo, ninguno superó el desempeño presentado por la arquitectura propuesta por el estudio. El artículo menciona que dentro de futuros trabajos se podría explorar diferentes métodos para la selección de características.

Otro ejemplo de implementación de redes neuronales se puede hallar en [12], allí se propone un modelo hibrido de word2vec, una red neuronal convolucional y una LSTM. Al igual que el estudio anterior, la selección de características se basó en la técnica de word2vec, el vector resultante de esta selección se utilizó como vector de entrada para una red neuronal convolucional compuesta por 7 capas: 3 capas convolucionales, 3 capas de pooling y por último una capa fully connected de 20 unidades a la salida. Esta salida se conectó a una capa de max global pooling y luego a una capa de dropout; en este punto se conecta la red LSTM, que a su vez termina en una capa softmax en donde se estima la probabilidad de pertenecer a una clase o a otra. Con esta arquitectura, se obtuvo un accuracy del 91%; al compararse con técnicas de clasificación como soporte de máquinas vectoriales, Naïve Bayes, redes convolucionales solas y redes LSTM solas; la arquitectura propuesta sobrepaso el desempeño de los otros métodos, aunque no por un amplio margen ya que Naïve Bayes y soporte de máquinas vectoriales también presentaron un desempeño comparable y muy cercano. Se puede concluir de esta investigación, que el

modelo propuesto entrega muy buenos resultados de en la clasificación, sin embargo, como ya se ha visto en otros estudios, técnicas más tradicionales también presentan desempeños comparables, así pues, que este estudio deja de lado el campo de la selección de características, que se ha evidencia es parte fundamental para un desempeño optimo en el a clasificación.

Siguiendo en la misma perspectiva de encontrar una mejor técnica para clasificar, el estudio en [17] expuso un método en donde se implementan tanto redes neuronales convolucionales (CNN), como redes LSTM y redes neuronales para la clasificación del mensaje; si bien esta combinación pudiera parecer igual a algunas de las ya descritas, el enfoque tomado por los autores fue la de adicionar el ensemble features y el clustering. Con ensemble features se buscó caracterizar grupos de palabras por su sentido semántico y no por el sentimiento que este grupo representa (es decir, en la frase “me gusto la película”, se buscó caracterizar ese conjunto de palabra como un “gusto” por la película, mas no, la polaridad o el sentimiento subyacente de la oración, que para este ejemplo podría pensarse que es positivo). Con clustering, se buscó principalmente agrupar los grupos de palabras con semánticas similares, pero no por el sentimiento de los grupos. Finalmente, el estudio propone una arquitectura de red neuronal de 3 capas para realizar la clasificación del mensaje. Este estudio utilizo 5 bases de datos diferentes, algunas con críticas de cine, el mejor resultado de accuracy fue del 94.6% para la base de críticas de cine de IMDB. Para las otras bases de datos se obtuvieron resultados de alrededor del 85% de accuracy.

Continuando con técnicas avanzadas de aprendizaje de máquina como las redes neuronales, el estudio descrito en [13], propone una metodología basada en redes neuronales convolucionales para alcanzar el mejor desempeño en la clasificación. Dentro de su metodología, el estudio propone un método de selección de características basado en word2vec; este vector es la entrada a la red neuronal convolucional, que para el experimento se utilizaron diferentes arquitecturas, en donde algunos de los hiperparametros que se variaron fueron el número de capas convolucionales (entre 2 y 3), el número de capas ocultas (6 y 7), la cantidad de filtros y el tamaño del filtro. La arquitectura también incluyó una capa de dropout, así como también, una red fully connected, y fue entrenada con 5 épocas y el tamaño del batch de 64. Con estas combinaciones, el modelo propuesto obtuvo un accuracy del 95%, sin embargo, la base de datos utilizada estaba en idioma hindi, por lo cual sus resultados están ajustados a este

idioma, no obstante, la arquitectura de su modelo puede ser funcional para otros idiomas y para otras bases de datos, dado la cantidad de combinaciones que se realizaron y que de forma general obtuvieron buenos resultados. Tal y como lo concluyen los autores, este modelo presenta un desempeño excepcional comparándolo con otros artículos previamente vistos, pero como se mencionó con anterioridad, el idioma puede representar una facilidad para este tipo de técnicas, sin embargo, el estudio no analiza el efecto de este dentro del proceso de clasificación.

De las referencias citadas previamente, es de notar la falta de una estrategia única para realizar la detección de polaridad, y aunque todas las estrategias vistas pueden ser comparadas, dado lo cercano de sus resultados, ninguna es superior a la otra y, por el contrario, dejan abierto el campo de análisis para nuevas combinaciones y la implementación de mejores técnicas.

1.3 Definición del problema

Como se describe en [1], el efecto de Word-of-Mouth (WoM) es innegable y tiene un gran impacto en el éxito o en el fracaso en taquilla de una película. La incidencia de la opinión de los críticos y del público también ha sido registrada en [18], allí se concluye que, las opiniones positivas sobre una película, tienen un impacto positivo en la taquilla, por lo tanto contribuyendo al éxito de la misma, se relaciona, que si bien las opiniones del público así como de la de críticos profesionales, generan este fenómeno, la opinión de un crítico experto produce un impacto mayor en la taquilla, una opinión positiva generará mayor recaudación, mientras que una opinión negativa persuadirá al público a no ver el filme.

El análisis de la polaridad de las reseñas representa un objeto de estudio muy interesante e importante, ya que técnicamente significa seleccionar que palabras, formas de hablar, expresiones, etc., son las que identifican un critica como positiva o negativa, y como a partir de ellas se logra determinar si la crítica es positiva o no. En el campo de la analítica de sentimientos es importante la selección de características, es decir, las partes del mensaje analizado que definen o caracterizan la presencia de un sentimiento en el mensaje, así pues, si un mensaje contiene palabras como “feliz”, “alegría”, etc., denotarían un mensaje con un sentimiento de felicidad. Sin embargo, hay mensajes en donde el sentimiento no se expresa explícitamente, es así, como en las investigaciones descritas

anteriormente, no se presentan estrategias para el manejo de esta clase de mensajes, es así como en [11] se propone una estrategia para el manejo de la negación en las reseñas de películas; allí se tipifican diferentes clases de negaciones, y se propone una estrategia para el manejo de estos al momento de clasificar si la crítica es positiva o negativa.

Tanto para la selección de características como para la clasificación, es posible encontrar diferentes aproximaciones, con resultados similares, esto se evidencia en las aproximaciones propuestas en la literatura revisada y descrita previamente. No obstante, ninguno se muestra concluyente, y, por el contrario, dejan abiertos espacios para nuevas propuestas, nuevas combinaciones e implementación de nuevas técnicas.

Al no evidenciar una estrategia clara, única, el problema de investigación se configura en encontrar métodos de selección, acompañados de un algoritmo de clasificación, soportado en este trabajo con redes neuronales, que permitan determinar si una reseña de una película es positiva o negativa. ¿Es posible encontrar nuevos métodos de selección que permitan mejorar los resultados de la clasificación de las reseñas? ¿Cuáles son las técnicas que mejorar resultado presenta al momento de clasificar? ¿Qué información adicional se puede extraer del proceso de selección de características de las reseñas que permitan analizar desde otra perspectiva el efecto de la polaridad en las críticas de cine?

Con los interrogantes anteriores, se puede sugerir que la implementación de las mejores o más relevantes técnicas para la selección de características, así como de técnicas para la clasificación, acompañados por una efectiva estrategia para el manejo de las variaciones en la interpretación del lenguaje (ambigüedad, negación, sarcasmo, etc.), permitirán encontrar una arquitectura para el análisis de las reseñas, lo suficientemente robusta para interpretar y clasificar correctamente las reseñas, aportando una nueva mirada sobre cómo se deben entender estas expresiones, que por su naturaleza de un lenguaje informal, sarcástico, exagerado en ocasiones, representan un desafío para las técnicas propias de la análisis de sentimientos y un reto para el campo del procesamiento del lenguaje natural.

1.4 Justificación

En los últimos años, la industria cinematográfica ha visto como año tras año sus ingresos han aumentado, rompiendo récords en taquilla y marcando hitos en la historia de la

industria. El fenómeno del WoM (Word of Mouth) [1] ha tenido un efecto directo en la recaudación de las películas en cartelera, impulsado por el auge de redes sociales como Facebook o Twitter. Sin embargo, gran cantidad de películas no generan una ganancia, como se describe en [2], tan solo el 51% de las producciones, generan una ganancia real (Figura **1-3**); esto debido a los costos asociados, en su mayoría desconocidos para el público, que incluyen entre otras cosas, las campañas de marketing, distribución, copias, entre otros tantos como los descritos en [3], de donde se puede concluir que para lograr que una película genere ganancia, esta debe recaudar al menos el doble de su presupuesto inicial estimado (Figura **1-4**).

Figura 1-3: Porcentaje de ganancia/perdida con relación al presupuesto de una película producida en Hollywood.

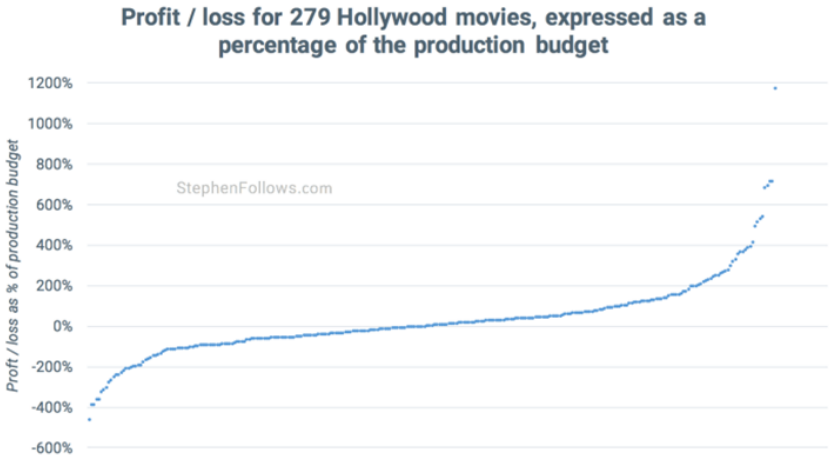
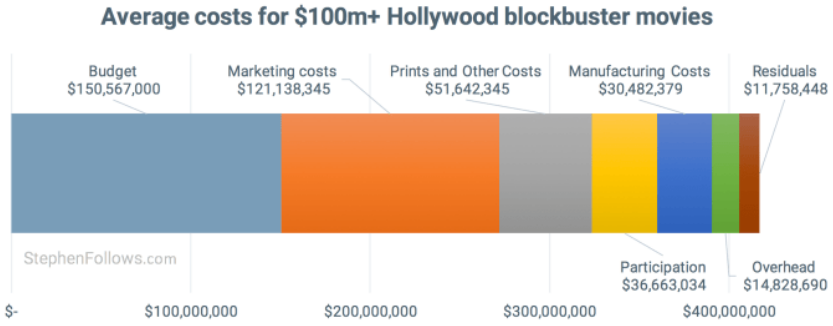


Figura 1-4: Distribución del costo de una película con presupuesto mayor a 100 millones de dólares.



El acceso directo que se tiene a las reseñas de críticos y del público, representan una fuente ideal para la implementación de técnicas de análisis de sentimientos, como las aplicadas en [5] [6], en donde por medio de técnicas como la correlación de Pearson o el método de ganancia de información, en donde se tiene como objetivo clasificar una reseña como positiva o negativa. También, como lo muestran las investigaciones [10] [12], técnicas de selección de características y de clasificación más novedosas, como las redes neuronales recurrentes, demuestran el excelente desempeño que presentan este tipo de arquitecturas al momento de seleccionar partes relevantes de un mensaje y con estas clasificar correctamente la polaridad de un mensaje.

Es por esto que en el campo de análisis de sentimientos sobre esta base de datos de reseñas, se presentan líneas de investigación diferentes, unas en donde se explora mejorar la clasificación desde la selección de características; otras en las cuales el enfoque del estudio se encuentra en mejorar los algoritmos que clasifican el mensaje, sin embargo, todas ellas dejan por fuera de su alcance importantes estrategias como la propuesta en [11], en donde se sugiere una metodología para el manejo de los diferentes tipos de negación en las reseñas; así como tampoco presentan técnicas de análisis combinadas, como se describe en [8] o [13].

Encontrar entonces aquellas técnicas que mejor desempeño presenten tanto al momento de seleccionar características como clasificar, y que además, contribuyan a una mejor compresión del comportamiento del público ante una reseña, hace de este estudio un proceso de alto interés, no solo por los costos de pérdidas de dinero asociados que se pueden reducir, sino como también, permiten la innovación al momento de implementar nuevos métodos de análisis de sentimientos y procesamiento del lenguaje natural.

1.5 Objetivos

1.5.1 Objetivo general

Proponer un método para la detección de polaridad sobre una base de datos de reseñas de películas de cine, basado en una técnica combinada de selección y clasificación.

1.5.2 Objetivos específicos

1. Analizar el comportamiento de los principales métodos de detección de polaridad reportados en la literatura, empleando una base de datos de reseñas de películas de cine.
2. Proponer un método para la detección de polaridad sobre una base de datos de reseñas de películas de cine empleando una técnica combinada de selección y clasificación.
3. Validar el método propuesto empleando un conjunto de métricas de desempeño en comparación con los principales métodos identificados en el primer objetivo.

2. Métodos para la detección de polaridad reportados en la literatura

Los métodos implementados en la actualidad con el objetivo de detectar la polaridad de un mensaje escrito han sido variados y han involucrado diferentes técnicas; de forma general, la labor de detectar la polaridad se basa en la identificación de las características del mensaje que permitan determinar la polaridad del mensaje, es así como, esta tarea normalmente se ha dividido en dos partes, la primera parte de selección y la segunda encargada de la clasificación.

En este capítulo se describen las técnicas más utilizadas para la detección de polaridad, que resultan de cumplir el objetivo específico 1 planteado en el numeral **1.5.2**., esto se logró al desarrollar una revisión sistemática de la literatura, en donde se seleccionaron aquellas investigaciones y estudios que abordaran la problemática de investigación y cuyos resultados fueran relevantes para este estudio.

Con los estudios seleccionados, resultado de la revisión sistemática desarrollada, se escogieron 3 arquitecturas, cuyos resultados fueron tomados como base para esta investigación y contra los cuales se realizó la comparación y se evaluó el desempeño del modelo propuesto por este estudio.

2.1 Marco teórico

2.1.1 Estado del arte en métodos de selección de características para la detección de polaridad

La selección de características dentro del campo del procesamiento del lenguaje natural consiste en seleccionar aquellas palabras, conjunto de palabras, expresiones, emoticones,

hashtags, entre otras, que hacen parte de un mensaje y con las cuales se puede determinar la polaridad del mensaje.

2.1.1.1 Bag of Words

También llamada bolsa de palabras o BoW, es una técnica que consiste en representar un texto por medio de la cantidad de ocurrencias de una o varias palabras dentro del mismo. Es así, que una bolsa de palabras está compuesta por: un listado de palabras conocidas y la frecuencia con la que ocurren en el texto. Este modelo de la selección de características es bastante popular por ser un concepto sencillo y un método lo suficientemente versátil para aplicar a diferentes problemas. Sin embargo, cuando la bolsa de palabras es muy grande, su manipulación se puede volver complicada, también dentro de sus falencias podemos nombrar que, al separar un texto por palabras, se desprecia el contexto del texto completo, por lo cual su clasificación se puede ver afectada [19] [20].

2.1.1.2 N-grams

Este método es un modelo en el cual se crean conjuntos de N palabras. Así pues, un bigram son dos palabras (Ej. “carro malo”), un trigram son 3 palabras (ej. “carro esta malo”), etc. Este modelo brinda un mejor contexto sobre el texto que se está analizando, pues crear conjuntos de varias palabras reduce la posibilidad de malinterpretación sobre palabras que cambian de significado dependiendo al contexto (ej. “El carro es blanco”, “Jugar tiro al blanco”) [19].

2.1.1.3 Term Frecuency Inverse Document Frecuency

También conocido como TF-IDF (por sus siglas en ingles). En este modelo de selección, un texto es representado por palabras (o también bag of words), y con la idea de que un término (palabra o bolsa), que se repite muy frecuentemente tiene muy poco peso o relevancia al momento de poder clasificar dicho texto. Es así, que, en un banco de reseñas de películas, la palabra “película” se repetiría muchas veces, y, sin embargo, esa palabra serviría de poco para poder identificar el sentimiento subyacente de la reseña; este término, bajo el modelo de TF-IDF, seria fuertemente penalizado, por lo que su importancia como característica para la clasificación sería muy poca [20].

2.1.1.4 Word2Vec

Este modelo parte del concepto conocido como Word Embeddings, que consiste en que se puede agrupar un conjunto de palabras, que pueden ser utilizadas en el mismo contexto en un texto, por lo que en esencia podrían significar lo mismo, siendo así representadas por un vector en donde se encuentra reflejado el contexto de la palabra. Ahora bien, para el caso específico de Word2Vec, existen dos métodos: CBOW (Continuos Bag of Words), en este método se busca encontrar la palabra (en un espacio de palabras definido) a partir del contexto en donde es utilizada; el otro método es Skip-gram, que consiste en encontrar el contexto de la palabra analizada [19].

2.1.1.5 Information Gain

Ganancia de información es un método en el cual, el texto es subdivido en conjuntos de palabras más pequeños, con ellos se mide la entropía (medida que describe la cantidad de información que aporta el texto analizado), así, el objetivo de la ganancia de información es encontrar aquellos grupos de palabras que más información aportan, basados en la probabilidad de ocurrencia de cada palabra dentro del texto, similar a TF-IDF, si una palabra tiene una probabilidad de ocurrencia muy alta, es también probable que no aporte mucha información [21].

2.1.2 Estado del arte en métodos de clasificación para la detección de polaridad

La clasificación de un mensaje entre positivo o negativo (polaridad) es una tarea para la cual han sido utilizados tanto métodos clásicos (Naïve Bayes por ejemplo) como métodos más avanzados (redes neuronales).

2.1.2.1 Naïve Bayes

Técnica de clasificación en la cual se parte de la probabilidad que tiene un elemento de pertenecer a una clase en particular. Se pueden presentar diferentes enfoques en esta técnica, uno en el cual la probabilidad de ocurrencia de una clase es conocida (a priori) y otro en el cual la probabilidad de una clase es desconocida y solo se determina después de clasificar un elemento (a posteriori) [22].

2.1.2.2 Random Forest

Es una técnica de clasificación que basa su funcionamiento en combinar el resultado obtenido de diferentes arboles de decisión. Un árbol de decisión es como su nombre lo indica, un árbol a partir del cual se van separando una a una las variables de decisión de un problema, la bifurcación de una decisión es llamada rama. El nodo final de una rama es llamado hoja y esta representa el resultado de la clasificación del problema [22].

2.1.2.3 Red Neuronal

Una red neuronal, como su nombre lo indica, hace referencia a una neurona humana, y comparten un comportamiento similar; ante una entrada, una neurona, entregara una salida. La función aplicada dentro de cada neurona para entregar esa respuesta puede variar, sin embargo, es comúnmente utilizada una función lineal. Las redes neuronales son métodos muy poderosos para realizar tareas de clasificación, sin embargo, son computacionalmente costosos, por lo que su utilización es recomendada solo en casos en el que ese esfuerzo de procesamiento sea absolutamente necesario. Con el tiempo se han desarrollado diferentes arquitecturas, dentro de ellos se puede mencionas: CNN – Red Neuronal Convolucional, esta arquitectura es especialmente utilizada en el procesamiento de imágenes, en donde a través de la aplicación de diferentes filtros sobre la imagen, se pueden encontrar características dentro de la imagen; aunque su principal aplicación es en el procesamiento de imágenes, estas también se han utilizado para el procesamiento de texto. Otra arquitectura muy potente, pero en el campo del análisis de texto son LSTM – Red Neuronal de Memoria de Corto Plazo, estas redes tienen la capacidad de recordar valores de la entrada, de forma tal que puedan ser retroalimentados nuevamente a la red con el fin de lograr un mejor aprendizaje; esta arquitectura es especialmente aplicada a textos, pues en estos, el contexto de una palabra puede cambiar por completo como la red interpreta esa palabra, por lo cual, si se considera una palabra como entrada de una red, es importante para esta recordar uno o varias palabras ya vistas, de forma tal que con estas pueda determinar el contexto de la entrada actual y así arrojar un resultado acertado [19] [22].

2.1.3 Estado del arte en técnicas de preprocesamiento sobre datos de texto

Una tarea fundamental dentro de cualquier proceso de clasificación es el preprocesamiento de los datos a utilizar. Durante esta etapa se identifican las características a utilizar, el formato en el que se encuentran (numéricas, categóricas, texto, etc.), la cantidad de datos nulos, datos incorrectos, etc.; en el procesamiento del lenguaje natural, se han identificado algunas tareas de importantes de limpieza y transformación, que pueden afectar directamente el desempeño de los modelos de clasificación.

2.1.3.1 Segmentación

En el procesamiento de lenguaje natural, se puede buscar dividir un texto en partes más pequeñas, así, el texto de un libro podría dividirse en páginas, las páginas en párrafos, estos a su vez se podrían dividir en frases.

2.1.3.2 Tokenizacion

La tokenizacion es el proceso en el cual se separa una palabra, frase o texto, estos se llaman *tokens*. Estos se pueden considerar como los bloques más pequeños a partir de los cuales se pueden construir frases, párrafos, etc. A diferencia de la segmentación, los tokens son la base del procesamiento del lenguaje natural, es a estos a los que se realizan los diferentes pasos de preprocesamiento y con estos se construyen los modelos y métodos.

2.1.3.3 Lematización

La lematización consiste en encontrar a partir de la forma flexionada de una palabra, es decir, una palabra en plural, conjugada, en femenino, etc., su lema o raíz de la cual proviene la palabra en cuestión. Por ejemplo, la palabra *jugando*, esta conjugada, por lo que se lema será *jugar*, lo mismo sucederá con palabras como *trabajando -> trabajar*, *malísima -> malo*.

2.1.3.4 Stemming

El stemming es el proceso de remover la parte de final de una palabra, normalmente las ultimas letras, con el fin de obtener la raíz de la palabra; aunque este procedimiento es similar a la lematización, el proceso de stemming puede no retornar una palabra valida, por ejemplo, para la palabra *viendo* la lematización encontraría la palabra *ver*, mientras que el proceso de stemming posiblemente entregaría como resultado la palabra *vie* o *vien*.

2.1.3.5 Remoción de stop words

El termino *stop words* hace referencia a palabras que se usan frecuentemente en cualquier texto, pero que por su naturaleza no aportar mucha información y por lo tanto su procesamiento solo significaría ocupar espacio de almacenamiento, y por su frecuencia de ocurrencia, alargarían el tiempo de ejecución.

2.2 Resultados de estudios previos

Basados en los artículos seleccionadas a partir de la revisión sistemática realizada para esta investigación, se escogieron 3 implementaciones que abarcaran las diferentes estrategias para cumplir con el objetivo de la clasificación, se replicó su arquitectura y se obtuvieron sus respectivas métricas de desempeño.

La primera arquitectura implementada fue la descrita en [13]; este estudio realizó diferentes experimentos con diferentes arquitecturas variando en ellas algunos hiperparametros de las mismas, buscando con esto encontrar aquellos que mejor rendimiento entregaban. De acuerdo con los resultados expuestos, la arquitectura más sobresaliente se compone de dos capas convolucionales de 50 filtros cada una, con un tamaño de filtro de 3 y 4. Esta red se conecta a una red fully-connected y esta a su vez, se conecta con una capa tipo softmax que entrega la predicción de polaridad de la reseña analizada. La Tabla **2-1** muestra los valores de hiperparametros utilizados y detalles específicos de la arquitectura implementada. La Figura **2-1** muestra de forma gráfica la arquitectura descrita.

Tabla 2-1: Hiperparametros utilizados en la implementación de la arquitectura descrita en [13].

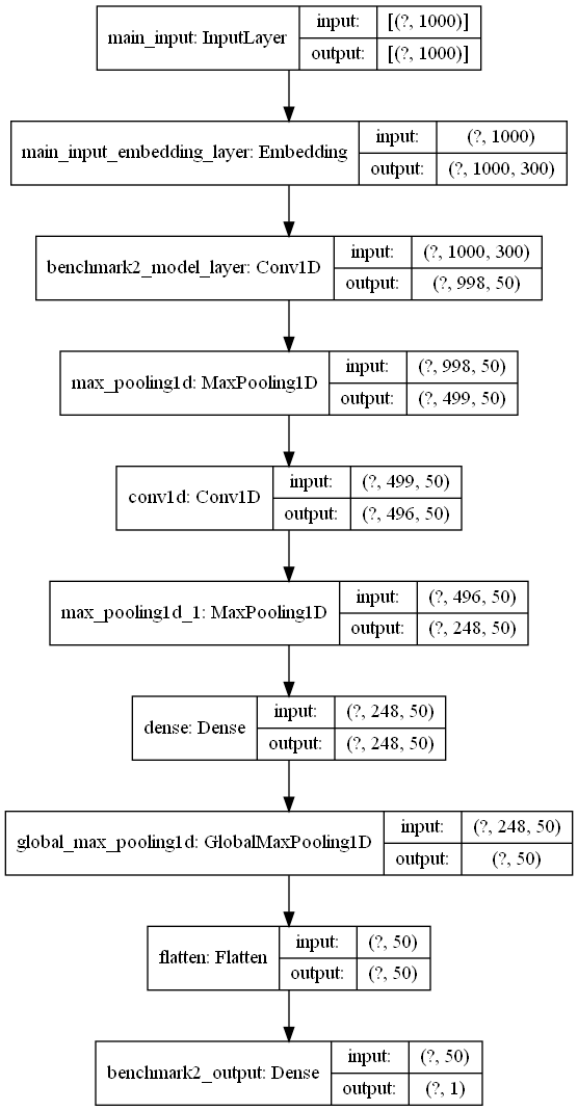
Hiperparametro	**Valor**
Capas de convolución	2
Capas ocultas	7
Funcion de activación	ReLu
Numero de filtros	50
Tamaño del filtro	3,4
Regularizador	L2
Dropout	0.5
Épocas	5
Tamaño del batch	64

Los resultados obtenidos de esta arquitectura se muestran en la Tabla **2-2.**

Tabla 2-2: Resultados obtenidos de la arquitectura descrita en la Tabla **2-1**.

Accuracy	**Recall**	**Precision**	**F1-Score**
87%	86%	88%	87%

Figura 2-1: Modelo gráfico de la arquitectura descrita en [13].



El estudio [10] describe la segunda arquitectura implementada, en ella se hace uso de las redes neuronales recurrentes (LSTM), luego, esta se conecta a una capa de una red neuronal sencilla, en cuya salida se realiza una función de activación de tipo *softmax*, esto con el fin de que esta última capa y mediante esa función se realice la predicción

(clasificación) de la reseña analizada. La Tabla **2-3** muestra los diferentes hiperparametros utilizados en esta arquitectura y la Figura **2-2** muestra de forma gráfica el modelo descrito.

Tabla 2-3: Hiperparametros utilizados en la implementación de la arquitectura descrita en [10].

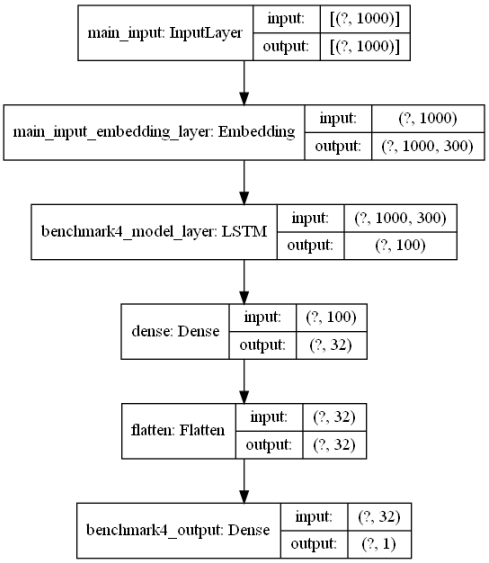
Hiperparametro	**Valor**
Optimizador	Adam
Learning rate	0.001
Learning rate decay	0.97
Regularizador	L2
Parametro de regularización	0.001
Dropout	0.5
Épocas	50
Tamaño del batch	128

Los resultados obtenidos de implementar la arquitectura descrita anteriormente se pueden observar en la Tabla **2-4**.

Tabla 2-4: Resultados obtenidos de la arquitectura descrita en la Tabla **2-4**.

Accuracy	**Recall**	**Precision**	**F1-Score**
90%	89%	89%	89%

Figura 2-2: Modelo gráfico de la arquitectura descrita en [10].



Por último, el estudio en [9] describe la tercera arquitectura implementada como referencia; en ella, se plantea el uso de la teoría de *Term Frequency-Inverse Document Frequency* (TF-IDF), con este método se realiza la selección de características, es decir, aquellas palabras que más información podrían aportar al momento de determinar la polaridad de la reseña. El método de clasificación utilizado en esta arquitectura fue el de máquinas de soporte vectorial (SVM por sus siglas en ingles). Al igual que en el estudio, diferentes configuraciones fueron probadas, variando tanto el método de selección como de clasificación, se pueden mencionar entre ellos Naive-Bayes multinomial, Random Forest, Counter vector y Bi-grams, no obstante, la arquitectura combinada de TF-IDF y SVM, presento un mejor desempeño. La Tabla **2-5** muestra los resultados obtenidos al implementar esta arquitectura.

Tabla 2-5: Resultados obtenidos de la arquitectura descrita en [9].

Accuracy	**Recall**	**Precision**	**F1-Score**
85%	88%	84%	86%

2.3 Análisis y discusión

En la Tabla **2-6** se resume los diferentes métodos de detección de polaridad que fueron descritos en las secciones anteriores.

Tabla 2-6: Resumen de métodos detección de polaridad utilizados como referencia.

Modelo	Características	Fortalezas	Debilidades	Resultados			
				Accuracy	Recall	Precision	F1-Score
[13]	Uso de redes convolucionales como técnica para la selección de características y una red fully-connected para la clasificación. Uso de dropout como técnica para prevenir el sobreajuste. Uso de activación de tipo softmax como técnica para obtener la probabilidad de que la polaridad de la reseña sea positiva.	El uso de redes convolucionales fortalece la selección de características, por lo cual el resultado final se basa en las palabras que mejor predicen la polaridad.	El modelo deja de lado la relación entre las palabras y su contexto, no retiene información previamente vista.	87%	86%	88%	87%
[10]	Uso de red LSTM. Uso de red fully-connected para la clasificación de la reseña evaluada. Uso de activación de tipo softmax para obtener la probabilidad de que la polaridad de la reseña sea positiva.	El uso de redes LSTM en este modelo garantiza que se toma en consideración el contexto de las palabras analizadas. Su habilidad para preservar información hace que se cree un contexto general de la reseña y no solo basado en algunas palabras consideradas importantes.	El modelo deja de lado la selección de características, en su lugar se enfoca que generar una predicción sobre todo el conjunto de palabras analizados, esto aun cuando presenta un buen desempeño, no considera técnicas de selección de características que permitieran mejorar el resultado en la predicción e incluso mejor el tiempo de entrenamiento.	90%	89%	89%	89%
[9]	Uso de TF-IDF como técnica principal de selección. Uso de máquinas de soporte vectorial para	Es un método de estructura sencilla, menos profunda.	Este modelo deja de lado las ventajas y beneficios de técnicas más nuevas como	85%	88%	84%	86%

	realizar la clasificación final.	Implementas métodos tradicionales, que presentan buenos resultados	las rede neuronales, por lo cual su desempeño se ve limitado en comparación a modelos donde las rede neuronales son la base de la arquitectura.				

Las diferentes técnicas utilizadas de los modelos de referencia seleccionados muestran la gran variedad de técnicas y arquitecturas que son utilizadas para la detección de polaridad; técnicas tradicionales como TF-IDF, máquinas de soporte vectorial, hasta técnicas algo más nuevas como las redes neuronales convolucionales, las redes LSTM, son solo algunas de las más empleadas para esta tarea. Ahora bien, los modelos que emplean técnicas relativamente nuevas como las redes neuronales, en todas sus formas, son los que mejores resultados entregan; las redes neuronales convoluciones son muy efectivas al momento de realizar selección de característica, pues por su diseño, son capaces de encontrar características o detalles dentro del texto analizado, esto se logra a través la aplicación de diferentes filtros sobre el texto, como se describió para el modelo de referencia propuesto en [13], se implementaron 50 filtros con una ventana de tamaño de entre 3 y 4; la función de los filtros es ir recorriendo el texto, con esto se va seleccionando aquellas palabras más importante dentro de la porción de texto que es seleccionada por la ventana. Es claro que esta forma de selección entrega buenos resultados, no obstante, debe existir un balance en el tamaño de la ventana a utilizar (tamaño del filtro), un filtro muy grande resultara en una ventana muy grande, que se procesara rápidamente las palabras analizadas pero por consiguiente iría en contravía en su objetivo de seleccionar las palabras más representativas que permitan obtener una predicción de polaridad suficientemente acertada; un filtro pequeño, como el usado en el modelo referencia, deriva en que solo se analiza un conjunto pequeño de palabras, despreciando con esto un contexto más amplio, como se ha mencionado previamente, la polaridad general de la reseña analizada puede no estar enmarcada en un pequeño número de palabras, sino, que su polaridad final puede estar ligada al análisis completo del texto; el uso de la ironía o del sarcasmo dentro del texto, puede afectar la predicción final al hacer un filtrado con una ventada pequeña, pues bien podría estarse analizando únicamente las palabras

usadas en las recursos del lenguaje mencionados, por lo cual la red convolucional podría perder efectividad.

El modelo en [10] se propone, a diferencia del modelo previo, el uso de redes neuronales de tipo LSTM, estas redes son muy utilizadas en el procesamiento del lenguaje natural y el análisis de sentimientos, su gran fortaleza radica en que este tipo de redes realiza un análisis de contexto mucho más amplio del texto. Por su diseño, las redes LSTM están en la capacidad de recordar o preservar información (texto) previamente analizado y decidir si este es de importancia para el texto actual analizado o si por el contrario se desecha. Esta forma de funcionamiento permite realizar un análisis más completo y general de la reseña estudiada, por lo que la predicción final será una ponderación de todas las frases y palabras que componente el texto. Es de notar el beneficio de utilizar esta técnica, esto permite paliar el efecto de los recursos del lenguaje como la ironía y el sarcasmo, no obstante y a pesar de su buen desempeño, no ofrece una técnica fuerte de selección de características, que bien podría reducir la cantidad de texto analizado y a su vez entregar un resultado mejor, pues basaría su predicción en la relación y el contexto entre las palabras que mejor pueden predecir la polaridad de la reseña, sin perder con esto, el análisis general de todo el texto, que es al final, la gran cualidad de este tipo de redes.

El modelo descrito en [9] hace uso de técnicas tradicionales, reconocidas por su efectividad en el procesamiento de lenguaje natural como lo son TF-IDF y las máquinas de soporte vectorial. Como técnica principal de selección es utilizada la técnica TF-IDF, en donde se evalúa la importancia de las palabras en la reseña con base a la frecuencia con la que aparecen en el texto; si bien esta técnica tiene un buen desempeño, presenta la misma de debilidad de las redes neuronales convolucionales, en donde no se considera un contexto amplio de las palabras, ni tampoco, la polaridad general de la reseña, que se basa en el análisis completo y conectado de todas las palabras del texto. El uso de las máquinas de soporte vectorial (SVM) como técnica de clasificación es la encargada de entregar la predicción de la polaridad final, las SVM son ampliamente utilizadas en tareas de clasificación por su gran desempeño, y como es evidente en los resultados de este modelo, la clasificación de las reseñas es superior al 80%, con un valor comparable a los otros modelos vistos. Es claro que las técnicas utilizadas por este modelo entregan un muy buen

desempeño, no obstante, deja de lado técnicas más especializadas como las redes neuronales, que como se pudo ver también tienen un gran desempeño.

Las diferentes técnicas implementadas en los modelos estudiados, como se describió, presentan un muy buen desempeño, razón por la cual son ampliamente utilizadas para la detección de polaridad, no obstante, nuevas formas de seleccionar, clasificar, combinar, se han desarrollado con el tiempo, los métodos de ensamble o las técnicas de fusión en redes neuronales, por mencionar algunos; se puede concluir entonces, que a pesar de los buenos resultados descritos por los diferentes modelos de referencia estudiados, no se explora nuevas aproximaciones para solucionar la tarea de detección de polaridad, que en la actualidad han demostrado su eficacia, además de que abren la puerta a modelos más flexibles que pueden integrar diferentes fuentes de datos para con esto ser aún más acertados en sus predicciones.

Es con este análisis que se cumple con el objetivo específico 1 planteado en el numeral **1.5.2**., concluyendo con este que aún hay un vacío por explorar con nuevas técnicas de procesamiento del lenguaje natural y análisis de sentimientos, impulsados por las cualidades de las nuevas técnicas, como lo son los métodos de ensamble, que permiten robustecer no solo la etapa de selección de características sino también la etapa de clasificación, y combinando estos resultados otras características externas (por ejemplo una fuente de datos diferente, o aplicando un preprocesamiento distinto a los datos), que permitirán capturar de forma más completa la polaridad general del texto analizado.

3. Método propuesto para la detección de polaridad en una base de datos con reseñas de películas de cine

Los métodos de ensamble, entendidos dentro del área del aprendizaje de máquina, son métodos que basan su funcionamiento en combinar el resultado de diferentes modelos, generalmente son modelos con desempeños bajos, pero que al ser combinados, el desempeño del modelo en general aumenta. Los métodos de ensamble no solo son utilizados para cuando se quiere potencializar modelos débiles, sino también cuando se cuenta con múltiples fuentes de datos, cuyos datos pueden estar en diferentes formatos.

En este capítulo se realiza la propuesta de un modelo para la detección de polaridad en una base de datos que contiene reseñas de películas de cine; para el desarrollo de este modelo se tomaron como base los métodos de ensamble y las técnicas de fusión, que se describen en este capítulo. Este modelo cumple con objetivo específico 2 del literal **1.5.2**. en donde se plantea realizar una propuesta con un modelo nuevo para la detección de polaridad. Se realizaron varios experimentos cambiando la arquitectura del modelo desarrollado, seleccionando aquella que entrego los mejores resultados con respecto a las demás arquitecturas ensayadas.

3.1 Marco teórico

3.1.1 Métodos de ensamble y métodos de fusión

Dentro de los métodos de ensamble se pueden mencionar tres clases: *Bagging, Stacking* y *Boosting*.

El método de bagging consiste en separar los datos de entrada en x cantidad de muestras, luego cada una de las muestras se pasa por igual cantidad de métodos (generalmente es

el mismo método); la predicción final se obtiene mediante algún mecanismo simple como votación o promedio. Random forest es un ejemplo de un método de ensamble del tipo bagging, en el, los métodos utilizados son arboles de decisión y las predicciones por ellos entregadas son combinadas por medio de votación (mayoría).

Otro tipo de método de ensamble es el conocido como boosting, en él se busca cambiar los datos de entrada de forma que se preste especial atención a aquellos datos que no fueron correctamente clasificados. Es así, que todos los métodos reciben los mismos datos para ser entrenados, y también, recibirán una entrada en donde se indica al siguiente método sobre qué datos se debe enfocar. La idea principal de este método es ir corrigiendo los errores cometidos por el método anterior, y con esto, mejorar el desempeño del modelo en su predicción final. AdaBoost y XGboost son métodos de este tipo.

Finalmente, el método de ensamble de tipo stacking, consiste en entrenar de forma paralela diferentes modelos, todos ellos reciben como entrada el mismo conjunto de datos, no obstante, cada modelo puede necesitar un preprocesamiento distinto, esto puede hacerse como parte del modelo o en una etapa previa. Los resultados de los modelos entrenados son enviados a un nuevo modelo que los combina, y que se entrena haciendo uso de estos, es así como la predicción final es el resultado de entrenar un nuevo modelo con las predicciones de los modelos anteriores.

Tener diferentes fuentes de datos representa también una labor particular en cuanto al preprocesamiento, bien sea por su naturaleza o por los requerimientos de los diferentes modelos a implementar, los datos pueden necesitar diferentes tratamientos para poder ser utilizados en el método de ensamble deseado; esto puede beneficiar aún más el resultado del método utilizado, pues se pueden obtener más y mejores características a partir de los distintos tratamientos realizados.

Los métodos de ensamble también traen consigo otra ventaja y es el hecho de que es posible combinar los diferentes modelos en distintos puntos de cada arquitectura, es decir, no es estrictamente necesario hacer la combinación con el resultado de la predicción, sino, que se puede tomar el resultado de algo punto especifico del modelo y realizar la combinación allí. Esta ventaja es especialmente útil al utilizar redes neuronales, dado que es posible tomar la salida de una capa oculta en particular y unir estos a las salidas de otra

capa de otra red (otro modelo). Estas formas de combinar los modelos se conocen como fusión.

Dentro de los métodos de fusión se pueden mencionar tres clases: temprana, media y de score. En la fusión temprana, la arquitectura de la red neuronal concatena dos tipos de entradas diferentes, como, por ejemplo, una imagen y un stream de video, estas dos fuentes de datos pueden o no recibir un preprocesamiento diferente antes de la fusión, luego las dos entradas son concatenadas o unidas, creando así un vector de características con una dimensión mayor y es este el que se utiliza para entrenar la red neuronal final. La fusión media consiste en realizar un proceso de selección de características a cada una de las entradas, y luego cada nuevo vector de características es concatenado o unido para ser utilizado por la arquitectura subsecuente.

La fusión de score consiste en concatenar o unir los resultados de clasificación entregados por cada uno de los caminos implementados, y con estos realizar una nueva clasificación.

3.2 Método propuesto de múltiples caminos para la detección de polaridad

Basados en los métodos de ensamble y las técnicas de fusión descritas previamente, en esta sección se propone una arquitectura de múltiples caminos, en donde cada una de las entradas es procesada por modelos diferentes, en este caso redes neuronales, redes convolucionales y redes LSTM.

3.2.1 Método de selección

El método de selección de características implementado se compone principalmente de una red convolucional. Esta red se encarga de encontrar aquellas palabras dentro de la reseña analizada que mejor pueden predecir la polaridad final del texto analizado. Esta labor es una de las principales ventajas de este tipo de redes; para esto es importante el tamaño de la ventana o el filtro con el cual la red convolucional va recorriendo la reseña, es así como se van analizando las palabras que se encuentran dentro de la ventana, y con la utilización de capas de tipo MaxPooling, se va seleccionando aquellas palabras que

mejor permiten predecir la polaridad del texto analizado y así se va reduciendo el tamaño de la reseña.

3.2.2 Método de clasificación

El proceso de clasificación del modelo propuesto se basa en la combinación de diferentes entradas, que en este modelo son tres: la salida de la red convolucional que se encarga de la selección, la salida de una red LSTM y el parámetro de negatividad de la reseña. La combinación de estas tres entradas, mediante una arquitectura de múltiples caminos, deriva en una predicción más acertada, que se soporte en un análisis más completo del texto de la reseña.

La red LSTM en este modelo, busca establecer relaciones entre las palabras dentro de la reseña analizada, no solo entre el conjunto de palabras cercanas, sino, también en reseñas largas, en donde la polaridad del texto no se puede determinar solo por algunas frases o palabras dentro del texto, sino, que se debe determinar a partir del texto completo; el uso de este tipo de red ayudara a una predicción más acertada.

Ahora bien, las arquitecturas implementadas basan su funcionamiento en el análisis de las palabras presentes en el texto, no obstante, las reseñas pueden contener gran número de palabras que comúnmente se utilizan de forma negativa sin que esto signifique, necesariamente, que la reseña tiene una polaridad negativa; esto puede deberse al uso de la ironía y el sarcasmo, estas figuras retoricas del lenguaje consisten en dar entender lo opuesto a lo que se está diciendo, la diferencia entre las dos radica en el hecho de que el sarcasmo, se utiliza como forma de burla, y en ocasiones, es una forma más agresiva e insultante que la ironía. Es entonces razonable pensar que, los modelos pueden verse “engañados” por el uso de estos recursos del lenguaje. Basado en esta premisa y por medio de un proceso previo, realizado con ayuda de la librería TextBlob, se estableció el parámetro de negatividad de cada reseña, que consiste en un valor numérico que indica que tanta negatividad hay en la reseña basada en la cantidad de palabras que frecuentemente se utilizan de forma negativa.

Con las tres entradas descritas anteriormente, se logra entonces, una sola gran entrada, que contiene toda la información que se desea tener en cuenta para realizar una predicción acertada. Estas entradas son tomadas por una red neuronal, que cumple una función de

agregación, y que, en su última capa, aplica una función de activación de tipo Softmax, que entrega la probabilidad de que una reseña sea positiva; se estableció que, si esta probabilidad era superior al 50%, la reseña se entendía por positiva, en caso contrario, se daría como negativa.

3.2.3 Arquitectura e hiperparametros del modelo propuesto

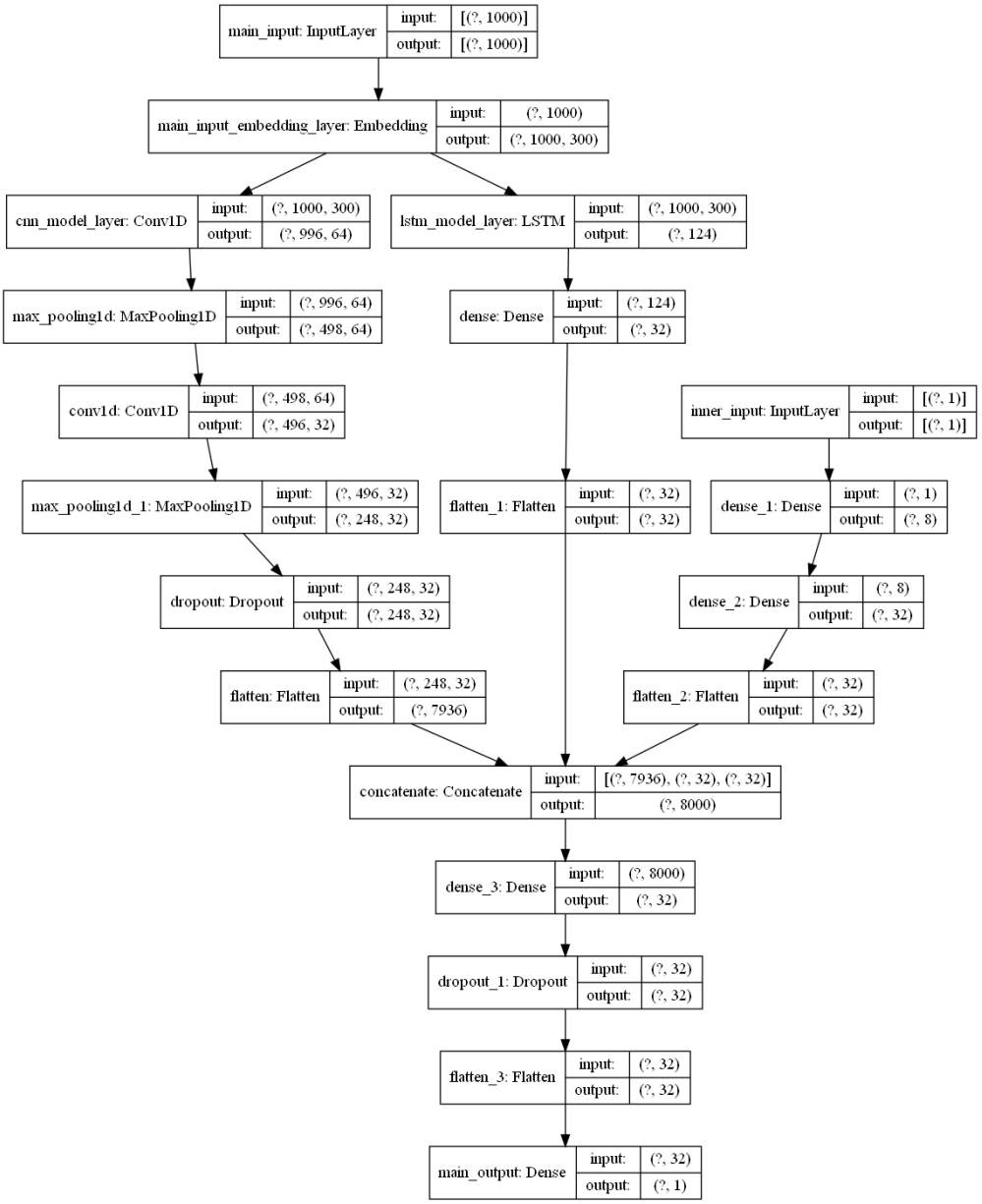
La Figura **3-1** muestra la arquitectura completa del modelo propuesto, en este se utilizó la fusión tardía y se incluyó la negatividad como una entrada intermedia al modelo, otras técnicas de fusión fueron evaluadas y sus resultados se detallan en el capítulo **4**. El modelo en la Figura **3-1**, muestra que la red convolucional utilizada se compone de 2 capas convolucionales, con 5 y 3 filtros respectivamente y con función de activación ReLu; cada capa convolucional es seguida por una capa de MaxPooling y al final como método para prevenir el sobreajuste, se aplicó una capa de Dropout de 0.2.

En la misma Figura **3-1** se ve la composición de la red LSTM, que para este modelo contiene 1 capa LSTM con 124 unidades y un Dropout 0.5, esta capa se conecta a una capa oculta con 32 unidades con una función de activación ReLu.

Ahora bien, la entrada de negatividad es conectada a una red neuronal compuesta por 2 capas ocultas con 8 y 32 unidades cada una y ambas con función de activación tipo ReLu.

Estas 3 salidas son concatenadas y luego es conectada a una red neuronal sencilla de 1 capa oculta con 32 unidades y una capa de Dropout con un valor de 0.2. La salida final de esta última arquitectura es la que entrega la predicción final y para esto se conecta a una capa de tipo Softmax con 1 unidad de salida. Esta unidad entrega la probabilidad de que la reseña analizada tenga una polaridad positiva, si esta probabilidad es menor al 50%, la reseña es clasificada como negativa.

Figura 3-1: Arquitectura de múltiples caminos para la detección de polaridad.



La Tabla **3-1** muestra los hiperparametros utilizados para el entrenamiento del modelo propuesto.

Tabla 3-1: Hiperparametros utilizados en la implementación de la arquitectura del modelo propuesto.

Hiperparametro	**Valor**
Optimizador	Adam
Learning rate	0.001
Epsilon	1e-8
Épocas	15
Tamaño del batch	128

3.3 Experimentación

3.3.1 Base de datos

Los datos utilizados por esta investigación fueron tomados de las bases de datos del sitio web especializado en cine, IMDb (*Internet Movie Database*). Este sitio web es reconocido por ser un gran agregador de críticas de cine y televisión, en donde, tanto la crítica especializada como la audiencia en general, puede escribir una reseña sobre alguna película o programa de televisión. Algunas investigaciones como [9] [10] han hecho uso de esta misma base de datos. Está compuesta por 50000 reseñas de cine con su respectiva etiqueta (polaridad – negativa o positiva) en idioma inglés. Esta base de datos se encontraba en formato CSV (*comma separated values*).

La base de datos se dividió en dos conjuntos de datos, el 70% como datos de entrenamiento y el restante 30% como datos de prueba.

3.3.2 Preprocesamiento

La etapa de preprocesamiento consistió en aplicar los siguientes procesos a los dos conjuntos de datos (entrenamiento y test).

- Remoción de etiquetas HTML.
- Remoción de signos de puntuación.
- Remoción de stop-words (palabras comunes que no aportan información importante).

- Conversión de las contracciones detectadas en la base de datos del idioma inglés a su forma completa, por ejemplo, *I'm* se convirtió en *I am*, o *It's bad* se convirtió en *It is bad*.
- Remoción de caracteres no alfanuméricos.
- Normalización (todas las palabras se convirtieron a minúsculas).
- Stemming.
- Lematización.

La aplicación de los procesos anteriores se realizó de forma secuencial, esto porque cada paso era requerido para el siguiente. El proceso de lematización requería, por ejemplo, que las palabras no tuvieran contracciones. Esta etapa de preprocesamiento solo se ejecutó una vez, y se almaceno el resultado en el mismo formato CSV de la base de datos original, esta nueva base de datos preprocesada fue la que se usó para realizar el entramiento del modelo propuesto.

Una vez realizado lo anterior se procedió a construir una matriz de embeddings; esta matriz es utilizada como los datos de entrada, tanto para el modelo propuesto como para los modelos de referencia, que hacen uso de esta técnica. Los métodos utilizados dentro del modelo para detectar la polaridad de las reseñas, no funcionan con texto propiamente, es decir, no hacen uso de palabras, la teoría bajo la cual operan estos métodos es esencialmente ecuaciones matemáticas, por lo que cada palabra o reseña analizada debe ser "convertida" a una representación numérica para que sea útil dentro del modelo propuesto; el proceso para realizar esta conversión y con ello, construir la matriz de embeddings, consiste en primero igualar la longitud de las reseñas a analizar, matemáticamente la matriz debe estar completa, las reseñas largas pueden ser recortadas y las reseñas cortas son completadas con palabras con ningún valor, lo que se podría interpretar como agregar un cero. Una vez surtido este primer paso, se procede a encontrar para cada palabra dentro de la reseña una serie de sinónimos que bajo el contexto en el cual la palabra está siendo utilizada puede entrar a sustituirla, esta serie o conjunto de palabras se conoce como vector, cada elemento del vector (que representa un sinónimo), es un valor numérico, que indica la similitud de ese sinónimo con la palabra analiza en el contexto especifico de uso, el termino similitud se asocia a que tan distantes se encuentran dos palabras en una espacio vectorial, es así que dos palabras que son sinónimos estarán

más cerca y por lo tanto serán más similares que palabras que sean antónimos, por ejemplo, la palabra “país” será cercana y similar a la palabra “Colombia”, dado que ambas se refieren a un país, pero será lejana a la palabra “silla”, porque a primera vista “país” y “silla” no tienen nada en común, no obstante, esta relación podría cambiar de acuerdo al contexto en la que alguna de las palabras sea utilizada. Con la matriz de embeddings construida, se procedió al entrenamiento del modelo propuesto.

3.4 Resultados

Los resultados obtenidos de la arquitectura implementada del modelo propuesto para la detección de polaridad se presentan en la Tabla **3-2.**

Tabla 3-2: Resultados obtenidos de la arquitectura en la Figura 3-1.

Accuracy	**Recall**	**Precision**	**F1-Score**
96%	92%	84%	88%

Durante la etapa de resultados también se recolectaron las métricas de desempeño de las entradas correspondientes a la red convolucional y a la red LSTM; estos se pueden ver en la Tabla **3-3.**

Tabla 3-3: Resultados obtenidos de la arquitectura en la figura 3-1, de las entradas independientes de la red convolucional y de la red LSTM.

Entrada	**Accuracy**	**Recall**	**Precision**	**F1-Score**
Red Convolucional	79%	83%	77%	80%
Red LSTM	88%	87%	88%	88%

3.5 Análisis

Como se ha descrito en los estudios [10] [13], el uso de redes convolucionales y redes LSTM, presentan un muy buen desempeño al momento de determinar la polaridad de una reseña, aunque estos estudios utilizan cada una de ellas por separado, la base de su funcionamiento es claramente útil para esta tarea. Es basado en estos buenos resultados

y con el uso de la teoría de los métodos de ensamble, que se construyó el modelo propuesto.

Los métodos de ensamble, en los cuales se basa el modelo propuesto, busca mejorar los resultados por cada arquitectura de forma individual, mediante la combinación de sus predicciones, es en este punto en donde las diferentes técnicas de fusión juegan también un papel fundamental, pues el punto en donde las arquitecturas individuales se combinan puede afectar la predicción final.

Se puede observar los buenos resultados que entrega el modelo propuesto y que se reportan en la Tabla **3-2**, y es también importante resaltar los resultados reportados por cada una de las arquitecturas utilizadas como entradas para el modelo general (red convolucional y red LSTM), cuyas métricas se reportan en la Tabla **3-3**. Es claro que cada una de las redes de forma individual, presentan un buen resultado, no obstante, cuando sus salidas son combinadas su potencial se maximiza y el resultado aumenta en promedio un 5%. Estos resultados comprueban, que la premisa de los métodos de ensamble se cumple y que la combinación de las diferentes arquitecturas potencializa el desempeño del modelo propuesto.

Es de notar, que el modelo propuesto, se beneficia también del parámetro de negatividad, que ingresa al modelo como una entrada nueva, esta flexibilidad de poder combinar características específicas, de las cuales se desea conocer su impacto en la predicción final, crean en este modelo una ventaja adicional, pues permitiría combinar nuevas entradas con formatos distintos, con preprocesamientos distintos, que pueden complementar el análisis realizado por el modelo propuesto.

Con los resultados expuestos, se concluye que el modelo propuesto cumple con el objetivo específico número 2 del literal **1.5.2.**; este modelo sigue el comportamiento esperado de los métodos de ensamble, y también evidencia que se diseñó una buena arquitectura para la selección de características, así como, la arquitectura complementaria que constituye el proceso de clasificación y que en un unión con el parámetro de negatividad, se presenta como una opción válida para la detección de polaridad de reseñas de películas de cine.

4. Validación del método propuesto

En este capítulo se valida el desempeño del método propuesto, mediante la comparación de sus resultados con las métricas entregadas por cada uno de los modelos de referencia implementados. Las métricas registradas fueron: *accuracy, precision, recall* y *f-score*.

Si bien todas las métricas seleccionadas sirven para describir el desempeño de cada arquitectura, el *recall* fue la métrica sobre la cual se determinó que arquitectura fue la mejor. Esta métrica hace referencia a detectar dentro de las reseñas analizadas y que han sido etiquetadas con polaridad positiva, cuantas fueron correctamente clasificadas por cada uno de los modelos implementados. Realizar esta validación y comparación de los desempeños registrados durante esta investigación, sirven para evidenciar los pros y los contras de cada arquitectura, además de abrir el camino para nuevas propuestas con técnicas más novedosas, como la implementada en el modelo propuesto que fue basada en métodos de ensamble.

Durante el proceso de desarrollo de una propuesta de modelo de múltiples caminos para la detección de polaridad se realizaron diferentes experimentos, en donde, como se describió en el capítulo **3**, se implementaron diferentes puntos de fusión, además de evaluar el desempeño de incluir o no la negatividad como una entrada nueva al modelo.

La Tabla **4-1** muestra las métricas de desempeño de implementar diferentes técnicas de fusión (tardía y de score). En esta tabla se puede ver como la técnica de fusión tardía es la que entrega mejor desempeño, con un recall de un 6% mayor a la técnica de fusión por score.

Tabla 4-1: Métricas de desempeño para las diferentes técnicas de fusión en el modelo propuesto.

Fusion	Accuracy	Recall	Precision	F1-Score
Tardía	91%	92%	84%	88%
Score	88%	88%	86%	87%

Utilizando como base el modelo que implementa la fusión tardía, se evaluó también el efecto en el desempeño de incluir como una entrada intermedia la negatividad. La Tabla **4-2** muestra las métricas de desempeño de incluir o no esta característica al modelo.

Tabla 4-2: Métricas de desempeño del modelo propuesto al incluir y excluir la negatividad como una entrada intermedia.

	Accuracy	Recall	Precision	F1-Score
Con negatividad	90%	92%	84%	88%
Sin negatividad	87%	86%	83%	84%

Se puede observar de la Tabla **4-2** que incluir la característica de negatividad al modelo como una entrada intermedia, incrementa un 6% el recall del modelo. Incluir este parámetro ayuda al modelo a determinar la polaridad de la reseña no solo analizando las palabras y las conexiones dentro del texto, sino también, entendiendo como esas mismas palabras son comúnmente utilizadas y qué tan frecuentemente se usan de forma negativa.

En la Tabla **4-3** se resumen las diferentes métricas de los modelos de referencia y del modelo propuesto. En ella se puede observar que el modelo propuesto presenta un recall aproximadamente un 4% mayor que los otros modelos implementados.

Tabla 4-3: Métricas de desempeño de los modelos de referencia y del modelo propuesto.

Modelo	Accuracy	Recall	Precision	F1-Score
Modelo de referencia 1	87%	86%	88%	87%

Modelo de referencia 2	90%	89%	89%	89%
Modelo de referencia 3	85%	88%	84%	86%
Modelo propuesto	91%	92%	84%	88%

De los resultados descritos se concluye que el método propuesto, que basa su diseño en los métodos de ensamble y las técnicas de fusión, presenta un desempeño no solo comparable sino superior a los modelos tomados de la revisión de literatura. Una arquitectura de múltiples caminos como la propuesta, permite combinar lo mejor de cada modelo implementado y al mismo tiempo, compensar sus debilidades.

Si bien la arquitectura del modelo propuesto es más robusta y profunda, esto debido a las diferentes redes que se interconectan antes de entregar la predicción final, es una arquitectura que permite evidenciar claramente el efecto que tiene incluir cierto tipo de entradas o como puede influir el tratamiento realizado a una entrada en particular. Para este modelo propuesto, se pudo observar como la característica de negatividad podía influenciar el resultado de la predicción; si bien su inclusión significo un aumento de solo el 6% de recall, en comparación al modelo donde no se incluida esta característica, es preciso anotar que tener la posibilidad de medir estos efectos y de incluir diferentes fuentes de datos que pueden venir de procesos independientes (como en este caso), es muy importante, especialmente hoy en día que se genera gran cantidad de información con diferentes formatos que requieren distintos tratamientos.

El análisis realizado en este capítulo cumple con el objetivo específico 3 del literal **1.5.2.** en donde se estableció como método de validación de desempeño del modelo propuesto, realizar la comparación de los resultados de las métricas de desempeño del modelo desarrollo y de los modelos de referencia escogidos.

5. Conclusiones y recomendaciones

5.1 Conclusiones

Diferentes métodos han sido utilizados para la detección de polaridad en mensajes de texto, como las reseñas de cine; enfoques distintos que abarcan desde modelos para la selección de características hasta arquitecturas que buscan mejorar la clasificación final.

La revisión sistemática de literatura evidencio la implementación de métodos tradicionales como los basados en Naives-Bayes, máquinas de soporte vectorial, e incluso, el uso de técnicas más novedosas y robustas como las redes neuronales convolucionales, redes LSTM, entre otros. Todos los modelos y arquitecturas revisados presentan métricas de desempeño superiores al 80%; se puede concluir que de la revisión literatura hecha, que las estrategias y modelos planteados, son eficientes al momento de detectar la polaridad de la reseña.

De acuerdo con lo expuesto en el capítulo **2**, y con base a las diferentes técnicas utilizadas para al procesamiento del lenguaje natural y el análisis de sentimientos, se concluye que los métodos seleccionados como modelos referencia, [9] [10] [13], son una buena representación de las distintas aproximaciones que se han implementado para resolver este problema; no solo por reportar un desempeño superior al 80%, sino también por las diferentes técnicas de las que hacen uso como las redes neuronales, máquinas de soporte vectorial, TF-IDF, entre otras, que permiten comparar de una forma razonable métodos tradicionales con métodos más novedosos, incluyendo el método propuesto por esta investigación.

Las arquitecturas de varios caminos, como el método propuesto en el capítulo **3**, son una evidencia que confirma el hecho de que varios métodos débiles, hacen uno fuerte; el método propuesto, que se compone de redes neuronales de dos tipos, convolucionales y LSTM, con una entrada extra (negatividad), para luego ser combinadas, entrega un mejor

desempeño que cuando cada una de esas arquitecturas es evaluada por separado. El poder realizar estas combinaciones, abren la puerta para realizar análisis detallados sobre arquitecturas particulares (ej: ¿Cómo mejorará el desempeño la adición de una arquitectura nueva con más capas en el modelo existente?), también para sirve para evaluar la importancia de características específicas, como en este caso se evidencio con la entrada del valor de la negatividad.

De los resultados obtenidos, se puede concluir que el modelo propuesto es comparable con otros modelos implementados para la misma tarea, e incluso supera el desempeño reportado por los modelos de referencia; en el capítulo 4 se describió el resultado entregado por cada uno de los modelos aquí implementados, allí se pudo observar que el resultado de recall del método propuesto fue en promedio de un 6% superior a los demás modelos. La razón fundamental para establecer esta métrica como la mejor medida de desempeño de esta investigación, es el hecho de que esta es un indicador de la correcta detección de las críticas positivas, y como ya se ha visto, es de gran importancia para las productoras de cine aumentar significativamente las reseñas positivas con el fin de influenciar al público a ir al cine. Si bien el recall es la métrica principal para este estudio, el desempeño del método propuesto en las otras métricas como el accuracy y presicion, son comparables con los otros métodos, e incluso llegando a ser mejor.

5.2 Recomendaciones

Dado que el lenguaje está en constante evolución se sugiere para trabajos o investigaciones futuras incluir nuevas formas de expresión y con esto evaluar los efectos de su inclusión y como afectan el desempeño del modelo propuesto.

También se sugiere continuar con la exploración de nuevos métodos de detección que permitan la combinación de distintas fuentes de datos, técnicas basadas en los métodos de ensamble parecen ser técnicas con un excelente desempeño al realizar las tareas como la analizada en este estudio.

Bibliografía

[1] Sang Ho Kim, Namkee Park y Seung Hyun Park, «Exploring the Effects of Online Word of Mouth and Expert Reviews on Theatrical Movies'Box Office Success,» *Journal of Media Economics,* vol. 26, nº 2, pp. 98-114, 2013.

[2] S. Follows, «Do Hollywood movies make a profit?,» 25 July 2016. [En línea]. Available: https://stephenfollows.com/hollywood-movies-make-a-profit/. [Último acceso: September 2019].

[3] S. Follows, «How movies make money: $100m+ Hollywood blockbusters,» 10 July 2016. [En línea]. Available: https://stephenfollows.com/how-movies-make-money-hollywood-blockbusters/. [Último acceso: September 2019].

[4] Saurabh Kumar, Avinay Mehta y Joy Pal, «Movie Success Prediction using Data Mining,» Vellore Institute of Technology, 2019.

[5] Fachrul Rozy Saputra Rangkuti, M. Ali Fauzi, Eka Dewi Lukmana Sari y Yuita Arum Sari, «Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Featured Selection,» *International Conference on Sustainable Information Engineering and Technology,* pp. 88-91, 2018.

[6] Sari Widya Sihwi, Insan Prasetya Jati y Rini Anggrainingsih, «Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naive Bayes Classifier,» *International Seminar on Application for Technology of Information and Communication,* 2018.

[7] Rosy Indah Permatasari, M. Ali Fauzi, Eka Dewi Lukmana Sari y Putra Pandu Adikara, «Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Navie Bayes,» *International Conference on Sustainble Information Engineering and Technology,* pp. 92-95, 2018.

[8] Yin Fulian, Pan Xingyi, Wang Yanyan y Su Pei, «A Word Vector based Review Vector method for Sentiment Analysis of Movie Reviews Exploring the applicability

of the Movie Reviews,» de *3rd International Conference on Computational Intelligence and Applications*, 2018.

[9] Nhamo Mtetwa, Awukam Ojang Awukam y Mehdi Yousefi, «Feature Extraction and Classification of Movie Reviews,» *5th International Conference on Soft Computing and Machine Intelligence,* 2018.

[10] Jyostna Devi Bodapati, N. Veeranjaneyulu y Shareef Shaik, «Sentiment Analysis from Movie Reviews Using LSTMs,» *Ingenierie des Systemes d'Information,* vol. 24, nº 1, pp. 125-129, 2019.

[11] Swastika Pandey, Santwana Sagnika y Bhabani Shankar Prasad Mishra, «A Technique to Handle Negation in Sentiment Analysis on Movie Reviews,» de *International Conference on Communication and Signal Processing*, India, 2018.

[12] Anwar Ur Rehman, Ahmad Kamran Malik, Basit Raza y Waqar Ali, «A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis,» *Multimedia Tools and Applications,* vol. 78, nº 18, pp. 26597-26613, 2019.

[13] Sujata Rani y Parteek Kumar, «Deep Learning Based Sentiment Analysis Using Convolution Neural Network,» *Arabian Journal for Science and Engineering,* vol. 44, nº 4, pp. 3305-3314, 2019.

[14] S. Basuroy, S. Chatterjee y S. A. Ravid, «How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budges,» *Journal of Marketing,* vol. 67, pp. 103-117, 2003.

[15] Suhariyanto, Ari Firmanto y Riyanarto Sarno, «Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet,» *International Seminar on Application for Technology of Information and Communication,* 2018.

[16] Mais Yasen y Sara Tedmori, «Movie Reviews Sentiment Analysis and Classification,» *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology,* 2019.

[17] H. T. N. y M. L. N. , «An ensemble method with sentiment features and clustering support,» *Neurocomputing,* vol. 370, pp. 155-165, 2019.

[18] Thaís L. D. Souza, Marislei Nishijima y Ana C. P. Fava, «Do consumer and expert reviews affect the length of time a film is kept on screens in the USA?,» *Journal of Cultural Economics,* vol. 43, nº 1, pp. 145-171, 2019.

[19] M. Bramer, Principles of Data Mining, Springer, 2013.

[20] Y. Goldberg, Neural Network Methods for Natural Language Processing, Morgan & Claypool Publishers, 2017.

[21] C. D. Manning, P. R. y H. Schütze, An Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2009.

[22] T. Hastie, R. Tibshirani y J. Friedman, The Elements of Statistical Learning, Springer, 2009.