INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE INGENIERÍA MECÁNICA Y ELÉCTRICA

UNIDAD PROFESIONAL "ADOLFO LÓPEZ MATEOS", ZACATENCO

"COMPILACIÓN DE CORPUS PARA LA DETECCIÓN DE NOTICIAS FALSAS EN ESPAÑOL"

T E S I S

PARA OBTENER EL TÍTULO DE

INGENIERO EN COMUNICACIONES Y ELECTRÓNICA

PRESENTAN:

RAMIREZ CRUZ JUAN MANUEL
PALACIOS ALVARADO SILVIA URSULA
FRANCA TAPIA KARIME ELENA

ASESORES:

M. EN C. ELIBETH MIRASOL MELÉNDEZ
DR. GRIGORI SIDOROV
DR. JUAN PABLO FRANCISCO POSADAS DURÁN

CIUDAD DE MÉXICO, JUNIO 2019

INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE INGENIERÍA MECÁNICA Y ELÉCTRICA
UNIDAD PROFESIONAL "ADOLFO LÓPEZ MATEOS"

T E M A D E T E S I S

QUE PARA OBTENER EL TÍTULO DE	**INGENIERO EN COMUNICACIONES Y ELECTRÓNICA**
POR LA OPCIÓN DE TITULACIÓN	**TESIS COLECTIVA Y EXAMEN ORAL INDIVIDUAL**
DEBERA (N) DESARROLLAR	**C. KARIME ELENA FRANCA TAPIA** **C. JUAN MANUEL RAMIREZ CRUZ** **C. SILVIA URSULA PALACIOS ALVARADO**

"COMPILACIÓN DE CORPUS PARA LA DETECCIÓN DE NOTICIAS FALSAS EN ESPAÑOL"

COMPILAR UN CORPUS DE NOTICIAS FALSAS EN ESPAÑOL E IMPLEMENTAR UN SISTEMA PARA LA DETECCIÓN DE NOTICIAS FALSAS A TRAVÉS DEL ANÁLISIS DE TEXTO UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL Y ESTRATEGIAS DE APRENDIZAJE AUTOMÁTICO.

- **INTRODUCCIÓN**
- **ESTADO DEL ARTE**
- **MARCO TEÓRICO**
- **MÉTODO PROPUESTO**
- **EXPERIMENTACIÓN Y RESULTADOS**
- **CONCLUSIONES**

CIUDAD DE MÉXICO, A 05 DE JUNIO DEL 2019.

A S E S O R E S

M. EN C. ELIBETH MIRASOL MELÉNDEZ

DR. GRIGORI SIDOROV

DR. JUAN PABLO FRANCISCO POSADAS DURÁN

ING. GABRIEL VEGA REYES
JEFE DEL DEPARTAMENTO DE
INGENIERÍA EN COMUNICACIONES Y ELECTRÓNICA

AUTORIZACIÓN DE USO DE OBRA

Instituto Politécnico Nacional
P r e s e n t e

Bajo protesta de decir verdad, los que suscriben: **JUAN MANUEL RAMIREZ CRUZ**, **SILVIA URSULA PALACIOS ALVARADO** y **KARIME ELENA FRANCA TAPIA**, manifestamos ser autores y titulares de los derechos morales y patrimoniales de la obra titulada **"COMPILACIÓN DE CORPUS PARA LA DETECCIÓN DE NOTICIAS FALSAS EN ESPAÑOL"**, en adelante **"La Tesis"** y de la cual se adjunta una copia impresa y un cd, por lo que por medio del presente y con fundamento en el artículo 27 fracción II, inciso b) de la Ley Federal del Derecho de Autor, otorgamos al **Instituto Politécnico Nacional**, en adelante **El IPN**, autorización no exclusiva para comunicar y exhibir públicamente total o parcialmente en medios digitales o en cualquier otro medio; para consulta y/o apoyo a futuros trabajos de investigación relacionados con el tema de **"La Tesis"** por un periodo de **5 años** contando a partir de la fecha de la presente autorización, dicho periodo se renovará automáticamente en caso de no dar aviso expreso a **El IPN** de su terminación. En virtud de lo anterior, **El IPN** deberá reconocer en todo momento nuestra calidad de autores de **"La Tesis"**.

Adicionalmente, y en nuestra calidad de autores y titulares de los derechos morales y patrimoniales de **"La Tesis"**, manifestamos que la misma es original y que la presente autorización no contraviene ninguna otorgada por los suscritos respecto de **"La Tesis"**, por lo que deslindamos de toda responsabilidad a **El IPN** en caso de que el contenido de **"La Tesis"** o la autorización concedida afecte o viole derechos autorales, industriales, secretos industriales, convenios o contratos de confidencialidad o en general cualquier derecho de propiedad intelectual de terceros y asumimos las consecuencias legales y económicas de cualquier demanda o reclamación que puedan derivarse del caso.

Ciudad de México a 21 de Octubre del 2019

Atentamente

Juan Manuel Ramírez Cruz

Silvia Ursula Palacios Alvarado

Karime Elena Franca Tapia

Agradecimientos

A mi madre quien con su amor, paciencia y esfuerzo me ha permitido llegar a cumplir hoy un sueño más, gracias por inculcar en mí el ejemplo de esfuerzo, dedicación y valentía, de no temer las adversidades.A mi padre, quien me enseñó que el mejor conocimiento que se puede tener es el que se aprende por sí mismo.A mi abuelo, quien me enseñó que incluso la tarea más grande se puede lograr si se hace un paso a la vez. A mi hermano, por su cariño, apoyo incondicional y por la paciencia que me tuvo durante todo este proceso.Agradezco a toda mi familia porque con sus oraciones, consejos y palabras de aliento hicieron de mí una mejor persona y de una u otra forma me acompañan en todos mis logros, sueños y metas. De igual manera doy mi profundo agradecimiento a todas las autoridades y personal que hacen la Unidad ESIME Zacatenco, por confiar en mi, abrirme las puertas y permitirme realizar todo el proceso de investigación dentro de su establecimiento educativo.

Finalmente quiero dedicar esta tesis a mi novio y todos mis amigos por apoyarme cuando más los necesito, por extender su mano en momentos difíciles y por el amor brindado cada día, de verdad muchas gracias, siempre los llevo en mi corazón.

Karime Elena Franca Tapia

A mis padres, Maximina C. M. y Lorenzo R. G., las dos personas mas importantes en mi vida, porque reconozco todos los sacrificios, que con todo el amor que los caracteriza han hecho para que yo pueda llegar hasta este punto de mi vida, por todos sus consejos, su esfuerzo, dedicación y sobre todo el amor transmitido.

A mis amigos, Lesli, Alejandro, Victor, Pacheco, Osvaldo, Karime y César, por que a lo largo largo de este trayecto se han convertido en mi segunda familia.

Juan Manuel Ramírez Cruz

A mi amada madre Silvia y a mi amado padre Rene,
a mi abuelita Alicia, a mi abuelito Rafael
y a mi querido Chopper.

Silvia Ursula Palacios Alvarado

Finalmente, en conjunto agradecemos al Dr. Juan Pablo Francisco Posadas Durán, al Dr. Grigori Sidorov, a la Dra. Helena Gómez Adorno, y a la Mtra. Elibeth Mirasol Meléndez; no solamente por su gran apoyo, sino también por el conocimiento, tiempo, enseñanzas y consejos brindados a cada uno de nosotros.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este capítulo se describen las generalidades de un proyecto enfocado a la creación de un corpus para la detección de noticias falsas en español y su detección automática. Se describe un panorama general sobre el problema que se quiere resolver, se presenta el objetivo del proyecto y se describen las motivaciones del proyecto.

1.1. Planteamiento del problema

La difusión de la información en la Web se puede definir como un proceso en el cual las noticias, eventos y opiniones, se publican, reciben y reenvían a través de los usuarios. La información difundida en las redes sociales sigue una ruta de un usuario a otro y de un sitio a otro, permitiendo que la información sea rastreable, sin embargo, debido a que la información se propaga de manera inmediata la información no es verificada antes de ser entregada.

Las noticias falsas proporcionan información que busca manipular a los medios y a las personas. A pesar de que en ocasiones las noticias falsas se originan por errores en la redacción, generalmente son creadas de manera intencionada. Sin importar la razón de su origen, las noticias falsas ocasionan confusión, preocupación, paranoia, tensión y desorden entre la audiencia que las consulta [1].

En los medios digitales es posible encontrar distintos tipos de contenido: sátira, opiniones, memes, imágenes, noticias, entre otros. No solamente las noticias son susceptibles de ser alteradas y utilizadas de manera malintencionada, también los videos e imágenes, por ejemplo, pueden ser alterados. La información en las redes sociales se extiende en segundos entre miles de personas, por ello resulta necesario desarrollar herramientas que ayuden a reducir la cantidad de información falsa existente en la Web.

Un sistema para la detección de noticias falsas tiene como objetivo ayudar a los usuarios a detectar y filtrar noticias potencialmente engañosas. Es posible utilizar técnicas de aprendizaje automático para detectar noticias falsas, sin embargo, estos sistemas requieren de corpus o conjuntos de datos para su entrenamiento, es decir, de noticias falsas y verdaderas previamente etiquetadas. El número de corpus sobre noticias falsas es escaso particularmente para el idioma español, surgiendo la necesidad de compilar un corpus para esta tarea.

En México existen algunos servicios que se dedican a la detección manual mediante el análisis realizado por expertos y en ocasiones por colaboraciones de algunos usuarios del servicio, como sucede por ejemplo en el sitio *VerificadoMX*[1] enfocado a publicaciones del ámbito político. En los trabajos reportados previamente sobre este tema, la mayoría de los sistemas que detectan de manera automática noticias falsas se encuentran enfocados hacia el idioma inglés. Una limitante para el desarrollo de este tipo de sistemas automáticos para el idioma español es la falta de los datos de prueba necesarios para poder realizar experimentos. En el presente proyecto de titulación se propone la compilación de un corpus de noticias falsas en el idioma español y se describe un prototipo de un sistema para la detección automática de noticias falsas, aplicando técnicas del procesamiento de lenguaje natural y aprendizaje automático.

1.2. Objetivo general

Compilar un corpus de noticias falsas en español e implementar un prototipo de sistema para la detección de noticias falsas a través del análisis de texto, utilizando técnicas de procesamiento de lenguaje natural y de aprendizaje automático.

[1] `https://verificado.mx/`

1.3. Objetivos particulares

- ✓ Compilar un corpus en español para el entrenamiento del sistema, que incluya noticias sobre distintos eventos.
- ✓ Diseñar un procedimiento para la recolección de noticias falsas de la Web.
- ✓ Etiquetar las noticias del corpus de acuerdo a su tema principal (ciencia y tecnología, deportes, economía, educación, espectáculos, política, salud, seguridad, sociedad).
- ✓ Diseñar un procedimiento para el etiquetado de las noticias con base en su veracidad (verdadera o falsa).
- ✓ Procesar los elementos del corpus, eliminando elementos innecesarios para el procesamiento de la información (imágenes, videos, hipervínculos, etc.) usando herramientas de procesamiento de lenguaje natural (lematización, segmentación a nivel de palabras, frecuencia de términos, normalización).
- ✓ Implementar módulos para la extracción de características lingüísticas de las noticias.
- ✓ Procesar los textos de las noticias para obtener una representación vectorial a partir de la frecuencia de los términos (palabras).
- ✓ Evaluar diferentes métodos de aprendizaje automático en el corpus de noticias falsas.

1.4. Justificación

Los medios digitales se han convertido en los medios de divulgación de mayor uso, desplazando a los medios de comunicación tradicionales como el periódico impreso, la radio o la televisión. Una de las razones del auge de los medios digitales es el hecho de que la información se propaga prácticamente en tiempo real y de forma masiva a través de ellas, además de que permiten recibir retroalimentación por parte de los usuarios. Debido a esta capacidad de retroalimentación que usuarios de medios digitales tienen es importante cuestionar el grado de confiabilidad de la información que se difunde a través de este medio, así como de las respectivas fuentes.

La publicación y difusión de noticias falsas no es un tema nuevo, sin embargo, en la actualidad su divulgación ha aumentado debido al crecimiento tecnológico haciendo que su identificación resulte ser un proceso complejo.

Algunos ejemplos del impacto que tiene la publicación de noticias falsas en medios digitales se mencionan a continuación. En el año 2017, el contenido falso difundido en la Web en contra de la candidata del Partido Demócrata Hillary Clinton influyó en el proceso electoral por la presidencia de Estados Unidos; a través de las redes sociales se difundían noticias que tenían encabezados como "*Hillary Clinton le dio 400 millones de dólares al Estado Islámico*" y "*Hillary Clinton gastó 200 millones de dólares en una mansión*", dichas noticias fueron difundidas por sitios pro Trump en Facebook, ya sea tergiversando o generando la información falsa [2]. Por otro lado, en México durante septiembre del mismo año tras el sismo que afectó a la CDMX y diferentes estados del país, empezaron a surgir rumores y publicaciones que aseguraban que la ONU había pronosticado un mega terremoto que ocurriría en México en días posteriores, ante lo cual la misma ONU se vio en la necesidad de desmentir tales publicaciones[2].

En el presente proyecto se plantea un prototipo de sistema que utilice técnicas de aprendizaje automático para detectar noticias falsas difundidas a través de la Web. Para dicho entrenamiento se propone la elaboración de un corpus que contenga ejemplos de noticias y que incluya casos en los que las noticias son verdaderas y casos en los que son falsas.

El trabajo se organiza de la siguiente forma: en el capítulo 2 se hablará acerca del estado del arte, en el capítulo 3 se hablará sobre las herramientas y conceptos clave necesarios para el proyecto, en el capitulo 4 se describe el método propuesto para el proyecto, en el capitulo 5 se muestran los resultados de los experimentos realizados para evaluar el sistema y finalmente en el capítulo 6 se muestran las conclusiones y el trabajo futuro.

[2] `https://twitter.com/CINUmexico/status/910302988375347200`

Capítulo 2

Estado del arte

En este capítulo se expondrán los avances, trabajos y desarrollos tecnológicos que se han logrado en el tema de detección de noticias falsas, describiendo generalidades de estos proyectos, dando un panorama general sobre el problema a resolver de cada uno de ellos, las soluciones que proponen, los parámetros y los métodos que utilizaron.

2.1. Análisis de noticias falsas basado en su contenido

Determinar si una noticia es verdadera resulta una tarea compleja, debido a que involucra diversos temas, por ejemplo, analizar los intereses detras de la decisión de publicar una determinada noticia, que pueden ser de tipo económicos, políticos, sociales o personales [3]. Para esta elección se tiene una agenda, que incluye los temas que la editorial considera relevantes y descarta otros, por lo que existe una cierta manipulación de la información [4].

El periodismo es una actividad que elabora y difunde información con el objetivo de revelar la verdad sobre un suceso. Sin embargo, algunas veces se tergiversa la realidad, por ejemplo, en las redes sociales donde existen diversas fuentes de información, es importante reconocer la subjetividad de las noticias publicadas para no tomarla completamente como verdadera [5]. Otro problema es la publicación de noticias falsas con el fin de hacer sátira de un hecho o

acontecimiento real e inclusive la publicación de una noticia falsa para afectar con alevosía a un tercero [5].

Se han elaborado algunas propuestas para resolver el problema de la detección de noticias falsas. Trabajos como *Fake News Detection on Social Media: A Data Mining Perspective* [6], reúne investigaciones sobre la detección de noticias falsas en las redes sociales, enfocándose en psicología, teorías sociales y algoritmos. Se revisan dos aspectos del problema: la caracterización de las noticias falsas y técnicas para su detección [6].

La etapa de extracción de características tiene como objetivo representar el contenido de las noticias y la información relacionada en una estructura matemática formal, para posteriormente construir un modelo de aprendizaje automático para diferenciar las noticias falsas de las noticias verdaderas. Se propone realizar la detección a partir de características del contenido de las noticias: título, texto del cuerpo, imagen y video [6].

En el proyecto *Fake News Challenge (FNC-1)*[1] consideran dividir el proceso de detección de noticias falsas en etapas, una etapa primaria es comprender lo que diversas organizaciones de noticias están escribiendo sobre el tema, proponiendo la automatización por medio de *Stance Detection* [7], haciendo cuatro clasificaciones: *agrees* (está de acuerdo), *disagrees* (desacuerdo), *discusses* (discute), *unrelated* (no relacionado) [8]. En este proyecto se busca estimar la postura a partir del cuerpo del texto en relación con el título, el texto del cuerpo puede estar de acuerdo, en desacuerdo, discutir o no estar relacionado con el título. Por lo que el objetivo de esta etapa es desarrollar herramientas que permitan organizar las noticias para que posteriormente personas puedan analizar e identificar de manera eficaz y rápida la información falsa de las noticias [7].

Otra forma en la que se ha abordado la clasificación de noticias falsas es la propuesta de *"Liar, Liar, Pants on Fire": A New Benchmark Dataset for Fake News Detection* [9], la cual considera seis etiquetas para las clasificaciones de veracidad: *pants-fire* (pantalones en llamas), *false* (falso), *barely true* (apenas verdad), *half-true* (mitad verdad), *mostly-true* (principalmente verdad) y *true* (verdadero) [9]. Estas clasificaciones se toman del conjunto

[1] `http://www.fakenewschallenge.org/`

de datos propuestos por periodistas de Politifact[2] que clasifican manualmente el conjunto de datos [9].

La obtención de las noticias necesarias para el entrenamiento que usa *Intelligent Disaster Response via Social Media Analysus – A Survey* [10], se realizó haciendo uso de corpus ya existentes como: *TweetTracker*[3], *AIDR* [4] y *Ushahidi*[5], esto con el fin de facilitar y acelerar el proceso de extracción y filtrado. El proceso de filtrado se basa en la extracción de tuits a partir de palabras clave y posteriormente a través de los metadatos correspondientes a la ubicación de las publicaciones [10].

2.2. Análisis de noticias falsas basado su origen y propagación

Existe un enfoque utilizado para la detección de las noticias falsas que consiste en analizar la fuente de la noticia. El trabajo *Intelligent Disaster Response via Social Media Analysis* [10], expone un método para la detección de noticias falsas relacionadas con desastres naturales por medio de las red social de Twitter, analizando las características del lenguaje de los tuits, así como sus variaciones a lo largo del periodo de tiempo en el que tiene lugar el desastre. Se plantea la problemática de la recolección de noticias en tiempo real, debido a que al inicio de las situaciones de desastre la cantidad de información es tal, que se dificulta seguirles la pista a todas las publicaciones para determinar su veracidad.

Otro punto que se considera en el trabajo mencionado anteriormente es la difusión de contenido no deseado como el spam, los rumores, las opiniones genéricas y el uso de *bots*. Los *bots* se han convertido en un asunto de relevancia debido a que son capaces de difundir grandes cantidades de información en un corto periodo de tiempo [10].

El sistema *Hoaxy* [11] es un sistema que permite almacenar y rastrear la propagación de las noticias, principalmente falsas, con el objetivo de tener un registro sobre cómo es que

[2] https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

[3] http://tweettracker.fulton.asu.edu/

[4] https://github.com/qcri-social/AIDR/wiki/AIDR-Overview

[5] https://www.ushahidi.com/

se dispersan y almacenan las noticias para posteriormente funcionar como un buscador que devuelve una lista con las noticias falsas que se han publicado sobre un tema, el número de veces que se ha compartido en Facebook o Twitter y en algunos casos artículos que desmientan o afirmen la veracidad de la información [11].

Otro proyecto que basa su funcionamiento comparando la fuente con otras es *BS Detector*[6], esta es una extensión de navegadores, el cual busca los enlaces de una página Web comparándola con referencias de una lista creada manualmente que contiene dominios de fuentes no confiables. Las clasificaciones propuestas por este proyecto son: noticias falsas, sátira, sesgo extremo, rumor, noticias estatales, ciencia basura, grupo de odio y ciberanzuelo (*clickbait*) [6].

El trabajo titulado *CSI: A Hybrid Deep Model for Fake News Detection* [12], propone tres atributos de las noticias falsas: el texto, las respuestas de los usuarios que las reciben y los usuarios fuente que las dispersan. A partir de estas características proponen un modelo llamado CSI integrado por tres módulos: *Capture* (captura), *Score* (puntuación) e *Integrate* (integrar). *Capture* se basa en la respuesta y el texto, *Score* en aprender la fuente por medio del comportamiento de los usuarios e *Integrate* es un clasificador de falso o verdadero [12].

2.3. Corpus de noticias falsas en inglés

El trabajo *Fake News Detection on Social Media: A Data Mining Perspective* [6], propone conjuntos de datos como *BuzzFeedNews*[7] que contiene 1,627 artículos, 826 *mainstream*, 356 de izquierda política y 545 artículos de derecha política. Otro conjunto es el de *BS Detector*[8] el cual a partir de la comparación de sus fuentes comentadas anterior ha logrado almacenar ya varios datos. *Credbank*[9]es un corpus recopilado entre el año 2014 y el 2015, su contenido son tuits clasificados como eventos o no eventos, eventos anotados con calificaciones de credibilidad; estos últimos están catalogados como ciertamente inexacto, probablemente inexacto, incierto, probablemente exacto, ciertamente preciso [6].

[6] http://bsdetector.tech/

[7] https://github.com/buzzfeednews/2016-10-facebook-fact-check/tree/master/data

[8] https://www.kaggle.com/MRISDAL/FAKE-NEWS

[9] http://compsocial.github.io/CREDBANK-data/

Por otro lado, en la primera etapa de *Fake News Challenge (FNC-1)* propone un corpus aprobado y hecho por varios periodistas[10]. El equipo *SOLAT en el SWEN* ganador del *Fake News Challenge (FNC-1)* propuso combinar varios modelos en un conjunto de un promedio de 50/50 entre Árboles de decisión impulsados por gradiente y una Red neuronal convolucional unidimensional (CNN), en el título y el texto del cuerpo, representados a nivel de palabra haciendo uso de los vectores establecidos de Google News, la salida de la CNN se envía a un Perceptor multicapa (MLP) con salida a las cuatro etiquetas que se habían propuesto. Respecto a los Árboles de decisión impulsados por gradiente (GBDT), se utilizaron características basadas en el título y el cuerpo, prediciendo la relación entre ellos, posteriormente las similitudes de medidas entre las palabras y el análisis de la Frecuencia de término – Frecuencia inversa de documento (TF-IDF) y la Descomposición del Valor Singular (SVD)[11].

En el trabajo de *"Liar, Liar, Pants on Fire": A New Benchmark Dataset for Fake News Detection* [9] los métodos usados para el aprendizaje automatizado fueron: Regresión logística regularizada (LR), un clasificador de Máquinas de vectores de soporte (SVM) [13] [14] y un modelo de Redes de memoria a corto plazo bidireccional (Bi-LSTMs) [15], y un modelo de Red neuronal convolucional (CNN) [16], implementado una Red neuronal para integrar texto y metadatos, este modelo híbrido requiere 5 etapas de entrenamiento. Para LR y SVM, se utilizó el LibShortText toolkit6 [17], para Bi-LSTM y CNN, también se hizo uso de TensorFlow[12] y word2vec[13] para iniciar las incrustaciones de texto [18]. Los resultados de los conjuntos de datos dieron una precisión de 0.204 y 0.208 en los conjuntos de validación y prueba. Los modelos con mejores resultados fueron el clasificador de texto SVM y modelos LR, sin embargo, las CNN superaron a todos los modelos, lo que resultó en una precisión de 0.270 en el conjunto de prueba. Uno de los aportes más significativos es que se mostró que al combinar metadatos con texto, mejoras son significativas se puede lograr para la detección de noticias falsas [9].

[10] https://github.com/FakeNewsChallenge/fakenewschallenge.github.io

[11] https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html

[12] https://www.tensorflow.org/

[13] https://pypi.org/project/word2vec/

Capítulo 3

Marco teórico

En este capítulo comenzaremos con la definición de los conceptos clave en el desarrollo de la propuesta como son procesamiento de lenguaje natural, bolsa de palabras, etiquetas gramaticales, lematización, palabras auxiliares y n-gramas. Posteriormente se describirá el modelo de espacio vectorial, el concepto de aprendizaje automático y los diferentes modelos de aprendizaje que se pueden utilizar para el entrenamiento de un clasificador. Por último la descripción de herramientas que se utilizaran para realizar el presente proyecto.

3.1. Procesamiento de lenguaje natural

El lenguaje se puede definir como la capacidad de un individuo para expresarse o comunicarse con el fin de transmitir una idea o mensaje. El lenguaje natural es la forma de comunicación entre humanos, en la cual, una parte del lenguaje puede ser expresado manera oral o escrito. El lenguaje escrito se define por medio de la gramática, sin embargo, es flexible ya que continúa evolucionando, dificultando su formalización [19].

Debido a la complejidad del lenguaje natural, es necesario descomponerlo en elementos más sencillos, dichos elementos se definen conforme a los siguientes niveles [19]:

- **Nivel fonológico**: Describe la relación entre las palabras y los sonidos que las representan.
- **Nivel morfológico**: Se refiere a la estructura interna de las palabras, su construcción y la extracción de raíces.
- **Nivel sintáctico**: Describe la unión de las palabras para formar oraciones fijando la estructura que ocupará cada una de ellas en la oración.
- **Nivel semántico**: Describe la extracción del significado de las palabras y como a su vez se unen para dar significado a la oración.
- **Nivel pragmático o discursivo**: Describe cómo se utilizan las oraciones en diferentes contextos y cómo ese uso afecta el significado de las oraciones.

Existen diferentes aplicaciones del procesamiento de lenguaje natural, por ejemplo, el análisis de sentimientos, traducción automática, recuperación de información, elaboración de resúmenes, entre otros. Cada una de las aplicaciones mencionadas anteriormente requiere de un grado de procesamiento diferente [19].

El procesamiento de lenguaje natural en una computadora requiere de analizar el texto de manera morfológica y sintáctica para posteriormente hacerlo de manera semántica y pragmática. Los algoritmos relacionados con el procesamiento de lenguaje natural no tienen la capacidad de trabajar con la definición estricta de las palabras (una secuencia de letras), se requiere de una representación simplificada de los textos, es decir, una representación en la que se omitan los detalles que carecen de importancia o que no dan información relevante [20].

Uno de los modelos utilizados con frecuencia para realizar procesamiento de lenguaje natural es el modelo de bolsa de palabras (*bag-of-words*), que hace referencia a una representación simplificada del documento en la que se asume que la ocurrencia de un elemento lingüístico (palabras, prefijos, sufijos, lemas, entre otros) es independiente de los elementos lingüísticos que le preceden. El modelo de bolsa de palabras permite analizar de forma independiente el comportamiento de las características que conforman el texto.

El modelo de bolsa de palabras permite obtener una representación vectorial de un texto en la que cada dimensión se indica el comportamiento de una característica. Existen diferentes estrategias para modelar un texto de acuerdo con el nivel de análisis requerido por el problema

a resolver (por ejemplo tf, idf y tf-idf) [21]. Una representación basada en el modelo de bolsa de palabras se puede obtener por medio de la frecuencia de ocurrencia de las palabras sin tomar en cuenta las características gramaticales o el orden de las palabras. En la tabla 3.1 se muestra un ejemplo de una representación en términos de la frecuencia de ocurrencia de las palabras, utilizando el modelo de bolsa de palabras considerando la siguiente oración como entrada: *A las seis de la tarde se levantó de la cama y se puso la corbata.*

Tabla 3.1: **E**jemplo de representación usando el modelo de bolsa de palabras

Palabra	**Frecuencia**
A	1
las	1
seis	1
de	2
la	3
tarde	1
se	2
levantó	1
cama	1
y	1
puso	1
corbata	1

Las palabras que se muestran en la tabla 3.1 se pueden clasificar de la siguiente manera:

- Palabras auxiliares (*Stopwords*): Son las palabras cuyo nivel de frecuencia en el texto es elevado. Por ejemplo: artículos, preposiciones, conjunciones, pronombres, entre otros [22].
- Palabra clave (*Keyword*): Son aquellas palabras cuya semántica es de ayuda para identificar el contexto o el tema cardinal del texto [22].

Eliminando las palabras auxiliares al ejemplo anterior, la representación del ejemplo mencionado en la tabla 3.1 queda como se muestra en la tabla 3.2.

Tabla 3.2: Bolsa de palabras, sin palabras auxiliares

Palabra	Frecuencia
seis	1
tarde	1
se	2
levantó	1
cama	1
puso	1
corbata	1

La lematización hace referencia a encontrar el lema o raíz de las palabras sin que esto signifique que haya una equivalencia entre el lexema y la palabra. Este proceso es elaborado a partir de un vocabulario y analizándolo de manera morfológica [23]. La importancia de la lematización se encuentra en que las palabras, independientemente de su idioma de origen, se encuentran frecuentemente flexionadas, es decir, una palabra está flexionada cuando tiene alguna o varias de las siguientes características: está escrita en plural, en femenino, conjugada (en el caso de los verbos), en diminutivo o superlativo. Por ejemplo: amigos, amiga, amigote, amiguita.

En algunas ocasiones se puede confundir la lematización con el *stemming* sin embargo este último corta las palabras sin una rigidez en el vocabulario como lo hace la lematización, intentado eliminar las derivaciones a partir de cortar los extremos de las palabras [24].

Las etiquetas de clases gramaticales o también llamado POS (*Part of speech*) es el proceso de asignarle a cada palabra una etiqueta a partir de su función en la oración, por ejemplo si es un artículo, sustantivo, verbo, adjetivo, preposición u alguna otra categoría. Hacer esta asignación dependerá del contexto por lo que las categorías gramaticales se apoyan en palabras previas y posteriores, ajustando la categoría [25]. En la tabla 3.3 se muestran ejemplos de palabras con sus respectivas lematizaciones y etiquetas de clase gramatical.

Los n-gramas se definen como una secuencia de elementos (palabras) que mantienen su estructura tal como se encuentran en el texto, estas pueden proporcionar información relevante de los documentos que se analizarán, la letra *n* nos indica el número de elementos a consi-

Tabla 3.3: Lematización y etiquetas de clase gramatical

Palabra	**Lematización**	**Etiqueta POS**
levantó	levantar	verbo
puso	poner	verbo
mil	mil	número
456	456	número

derar (la longitud de la secuencia), por ejemplo bigramas (2-gramas), trigramas (3-gramas), 4-gramas, etc. [20]. En la tabla 3.4 se muestran los n-gramas posibles de una oración de ejemplo.

Tabla 3.4: Ejemplo de n-gramas

Origial:	Ana comió una zanahoria cruda
2-gramas:	(Ana comió), (comió una), (una zanahoria), (zanahoria cruda)
3-gramas:	(Ana comió una), (comió una zanahoria), (una zanahoria cruda)
4-gramas:	(Ana comió una zanahoria), (comió una zanahoria cruda)

3.2. Representación de espacio vectorial

La representación de espacio vectorial se basa en un modelo algebraico que busca representar de manera formal a los objetos describiéndolos mediante características y valores de estas características. La construcción de una representación vectorial es a partir del análisis del problema en particular, ya que para cada problema las características a considerar son diferentes. Una vez teniendo los términos de las características y sus valores, se construye un espacio de N dimensiones, donde cada dimensión corresponde a una de las características. Un vector n-dimensional con valores en cada una de las dimensiones corresponde a una instancia de las entidades de interés. Un espacio vectorial se puede expresar de forma matricial, donde las columnas corresponden a los objetos y las filas a las características [20].

El espacio vectorial nos ayuda a comparar los objetos de manera formal, haciéndolo necesario para aplicar los métodos de aprendizaje supervisado [20] que se ocupan en el presente trabajo.

En la tabla 3.5 se muestra un ejemplo de representación vectorial de tres oraciones utilizando el modelo de bolsa de palabras. Las palabras corresponden a las dimensiones del espacio vectorial y los valores en cada dimensión corresponden a la frecuencia de ocurrencia de cada una de ellas.

Tabla 3.5: Matriz término-documento

El	Club	**América**	se	encuentra				**en**	crisis			
El		**América**			remonta			**en**		dos	minutos	
El		**América**				no	participa	**en**				torneo
1	1	1	1	1	0	0	0	1	1	0	0	0
1	0	1	0	0	1	0	0	1	0	1	1	0
1	0	1	0	0	0	1	1	1	0	0	0	1

3.3. Aprendizaje automático

El aprendizaje automático (*Machine Learning, ML*) es una rama de la Inteligencia Artificial (IA), enfocada a la construcción de modelos de la inteligencia humana, donde a partir de la capacidad de las computadoras para procesar grandes volúmenes de información se han creado modelos que permiten obtener características o patrones importantes de los datos. Los algoritmos operan mediante la construcción de modelos basados en conjuntos de datos de entrenamiento, estos algoritmos son heurísticos, ya que no existe un modelo determinado para resolver un problema [26].

El aprendizaje automático se utiliza cuando el objetivo es la predicción o clasificación automatizada y existen datos históricos disponibles para entrenar. Cuando se implementa el

aprendizaje automático se busca automatizar los procesos, además puede ser utilizado en cualquier organización que necesite herramientas para entender los datos que se tienen y ser capaces de reconocer los datos que se tendrán más adelante. La solución de un problema dependerá del tipo de sistema a realizar, estos se clasifican en aprendizaje supervisado y el aprendizaje no supervisado [26].

El aprendizaje supervisado consiste en un algoritmo que requiere de datos etiquetados para poder generar patrones, los cuales usará posteriormente para clasificar datos nuevos [26]. La relación que se descubre se representa en una estructura denominada modelo. Normalmente, los modelos describen y explican fenómenos que están ocultos en el conjunto de datos y que se pueden usar para predecir el valor de la salida a partir de los valores de los atributos de entrada [26].

Existen diferentes algoritmos de aprendizaje supervisado, por ejemplo, los Árboles de Decisión que son un algoritmo que se puede usar para clasificar, ya que se estructuran de manera jerárquica tomando decisiones y sus consecuencias. El objetivo es crear un modelo que predice el valor de una variable objetivo mediante el aprendizaje de reglas de decisión definidas a partir de las características de los datos [27]. Un árbol se componen de nodos internos de decisión y hojas terminales las cuales están predefinidas: los nodos corresponden a los atributos de muestras y las hojas a los atributos de clases. Los valores posibles del atributo están representados en los arcos que salen del nodo correspondiente [27]. El proceso comienza en la raíz, cada nodo de decisión implementará una función de prueba y su elección indicará que rama seguir, este proceso se realiza de manera recursiva hasta llegar a las hojas para lograr una predicción a partir de un conjunto de reglas de decisión [26].

Los Árboles de Decisión muestran un modelo visual, por lo que no solamente se puede utilizar como modelo de aprendizaje supervisado, también podemos usarlos como un modelo que permita visualizar los atributos que están aportando información relevante en los modelos de aprendizaje y observar las jerarquías de importancia entre los atributos [27].

Otro ejemplo de algoritmo de aprendizaje supervisado es el Bosque Aleatorio (*Random Forest*) que construye un conjunto de árboles de decisión que se combinan para obtener una predicción más precisa. Cuantos más árboles se agreguen la aleatoriedad aumenta, en vez de

buscar la característica más importante al dividir un nodo, busca un árbol aleatorio con el mejor atributo, por lo que da un rango más amplio de elección. [28]

Las Máquinas de Vectores de Soporte (*Support vector machine*, SVM) son un algoritmo de clasificación supervisada que representa las variables en el espacio buscando márgenes que separen los atributos de las clases, estas separaciones son lineales y son llamadas hiperplanos. El algoritmo SVM busca un hiperplano óptimo por medio de los vectores de soporte, los cuales son creados a partir de los ejemplos de entrenamiento de cada clase, separando entre una clase y otra, para posteriormente hacer las correspondencias con los nuevos datos [29]. Como no todos los problemas son lineales, el algoritmo de SVM transforma el espacio en otras dimensiones donde sea posible separarlos formando así los hiperplanos que separan las clases. El algoritmo de SVM puede utilizarse como clasificador binario, sin embargo, también funciona como clasificador multiclases, eligiendo uno para compararlo con los demás, o bien, comparar uno a uno [30].

Regresión Logística es un algoritmo para la clasificación de dos clases y multiclase, que busca pronosticar la probabilidad de que ocurra o no un suceso determinado, modelando la relación que existe entre una variable dependiente y una o más variables independientes [26]. Por medio de función logística como la función *sigmoide*, o bien como una curva en forma de S, que toma como entrada cualquier número real y devuelve un número real comprendido entre 0 y 1, los cuales corresponden a la clasificación interpretado como la probabilidad de que el objeto de entrada pertenezca a una u otra clase [26].

Boosting es un método general que busca mejorar el rendimiento de cualquier algoritmo de aprendizaje, usando el refuerzo para reducir significativamente el error de cualquier modelo de aprendizaje débil [31]. *Boosting* es un algoritmo que combina clasificadores, requiriendo el rendimiento de cada uno por lo que se refiere a un método general y probablemente efectivo para producir una regla de predicción más precisa a la clasificación verdadera mediante la combinación de las reglas generales de los modelos débiles [32].

Para tener resultados óptimos de los métodos de aprendizaje supervisado se utilizan múltiples divisiones en los conjuntos de entrenamiento-prueba llamadas validación cruzada, donde los resultados se promedian en múltiples conjuntos de entrenamiento diferentes dando así resultados más confiables [33]. En la validación cruzada de K iteraciones (*K-fold cross-validation*)

los datos se dividen en K subconjuntos, donde uno de los subconjuntos se utiliza como datos de prueba y los subconjuntos restantes (K-1) como datos de entrenamiento [34]. El proceso se repite K veces, con cada uno de los posibles subconjuntos de datos de prueba. El error se calcula a partir de la media aritmética de los errores de cada iteración para obtener un único resultado [34]. Si MSE_i denota el error de la iteración $i-$ésima, entonces: El error de la validación cruzada se estima por

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

Una vez que se crea el modelo de aprendizaje automático, es necesario evaluarlo para conocer el rendimiento que tendrá y la exactitud [35]. La exactitud (*Accuracy*) muestra la precisión del subconjunto obtenido, si todo el conjunto de etiquetas predichas para una muestra coincide estrictamente con el conjunto de etiquetas verdaderas, entonces la precisión del subconjunto es 1.0, de lo contrario es 0.0 [36]. Si $\hat{y}$ es el valor predicho de la $i-th$ muestra y y_i es el valor verdadero correspondiente, entonces la fracción de predicciones correctas sobre $n_{samples}$ se define como [36]:

(*Accuracy*),

$$exactitud(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}} 1(\hat{y}_i = y_i)$$

3.4. Lenguaje de programación Python

Python es un lenguaje que está disponible en varias plataformas (Unix, Windows, OS/2, Mac), en las que se puede ejecutar y programar sin necesidad de configuraciones especiales [37]. La sintaxis del lenguaje Python es simple por lo que proporciona claridad en él código y se puede manejar con sencillez palabras o cadenas de caracteres [37]. Es usado en el área científica, en la implementación de inteligencia artificial, biotecnología, aprendizaje automático. Python cuenta con diversos módulos y herramientas para el desarrollo de sistemas basados en el procesamiento de lenguaje natural [37].

En la tabla 3.6 se muestra una comparación entre Python y otros lenguajes de programación que cuentan con módulos para realizar tareas de procesamiento de lenguaje natural [38].

Tabla 3.6: Comparación entre distintos lenguajes de programación

Característica	**Python**	**Java**	**C++**
Representación	POO	POO	Estructurado
Descripción	Tiene tipado dinámico, por lo que no es necesario declarar el tipo de dato que va a contener una variable. Es un lenguaje con un tiempo de ejecución alto, requiere menos líneas de código que Java en su programación y cuenta con diversos módulos que facilitan la realización de tareas.	Tiene tipado estático, por lo que es obligatorio indicar el tipo variable. Es un lenguaje con un tiempo de ejecución más lento que Python, requiere, a pesar de contar con funciones y algunos módulos su programación requiere gran cantidad de líneas de código.	Tiene tipado estático al igual que Java. Es un lenguaje con un tiempo de ejecución alto, pero requiere de una cantidad excesiva de líneas de código al ya que se trata de un lenguaje estructurado.

En la tabla 3.7, se muestra las herramientas con las que cuentan cada uno de los módulos con los que cuentan los lenguajes de programación mencionados en la tabla 3.6.

Tabla 3.7: Comparación entre herramientas para aplicación de PNL

Lenguaje	**Python**	**Python**	**Java**	**C++**
Herramienta	spaCy	NLTK	coreNLP	syntaxNET
Soporte multilingüe	x	x	x	x
Tokenización	x	x	x	x
Etiquetado del discurso	x	x	x	x
Reconocimiento de entidad	x	x	x	-

A continuación se detallan las diferencias que existen entre cada uno de los módulos. En la tabla 3.8 se describen cada una de las características de los módulos, notando que el módulo spaCy contiene una cantidad mayor de características que lo hacen una herramienta funcional para las tareas de procesamiento de lenguaje natural [38] [39].

Como parte de las herramientas que nos proporciona el módulo de spaCy, cuenta con un modelo en lenguaje español pre-entrenado para la realización de procesamiento de lenguaje natural de texto de este idioma.

Tabla 3.8: Diferencias entre herramientas para PNL

Lenguaje	**Python**	**Python**	**Java**	**C++**
Herramienta	spaCy	NLTK	coreNLP	syntaxNET
Diferencias	Cuenta con un procesamiento rápido, por lo que su implementación es muy útil, realiza métodos de PNL como lemantización y cuenta con una cantidad alta de etiquetas, utiliza un enfoque orientado a objetos lo que la hace compatible con la forma de trabajar en Python.	Cuenta con un buen procesamiento, realiza métodos de PNL, como *steaming* y cuenta con una cantidad menor de etiquetas a diferencia de spaCy, para su uso requiere que el usuario explore la documentación para descubrir las funciones que necesita.	Constituye el preámbulo sobre las técnicas tradicionales de Procesamiento del Lenguaje Natural, a pesar de estar desarrollada en Java, posee una interface con Python, aunque su procesamiento es más lento que spaCy.	SyntaxNet es una estructura de procesamiento de lenguaje natural para TensorFlow, sin embargo, sus herramientas no son tan funcionales como spaCy ya que este proporciona eficiencia y funcionalidad para la extracción de características.

En la tabla 3.9 se muestran las diferentes versiones que este modelo en español pre-entrenado se distribuye con spaCy. Las evaluaciones que se realizaron se llevaron a cabo a partir de texto sin formato y sin procesamiento previo, con la herramienta de etiquetado POS con la que cuenta esta herramienta [38].

Tabla 3.9: Comparación de versiones del modelo español de spaCy

Modelo	**spaCy**	**Tipo**	**POS**	**Tamaño**
es-core-news-sm 2.0.0	2.x	neural	96.9	35MB
es-core-news-md 2.0.0	2.x	neural	97.8	93MB
es-core-Web-md 1.1.0	1.x	lineal	96.7	377MB

Para la parte de preprocesamiento spaCy cuenta con una herramienta llamada Textacy [40], la cual se enfoca en el preprocesamiento de textos y que se encarga de procedimientos como: tokenización, POS etiquetado, estadísticas de legibilidad, análisis de sentimientos y citas de atribución.

Cuenta con herramientas que se encargan de transmitir datos en varios formatos, proporciona funciones para limpiar y normalizar el texto. Además, vincula contenido de documentos, utiliza metadatos para una mejor contabilidad, transforma documentos en bolsas de palabras, redes semánticas y listas de términos.

Esta herramienta cuenta con funciones que modifican el texto sin formato, reemplazando URL, correos electrónicos, números de teléfono y símbolos de moneda con el fin de mejorar y hacer más rápido el análisis del texto, (Ver anexo B.1). Aplica modelos a documentos analizados, y facilita la visualización e interpretación de los resultados del análisis.

Para la parte de aprendizaje se utilizará el módulo Scikit-learn [36] disponible para el lenguaje de programación de Python, cuenta con herramientas que permiten implementar algoritmos de aprendizaje automático supervisado. La biblioteca incluye algoritmos como máquinas de vectores de soporte, bosques aleatorios, k- medias, entre otros.

Existe otro módulo para aprendizaje automático llamado TensorFlow, este módulo permite a los desarrolladores comenzar a utilizar el aprendizaje profundo en la nube. El marco tiene un amplio respaldo en la industria y se ha convertido en una opción válida para la investigación

de aprendizaje profundo y el desarrollo de aplicaciones, especialmente en ámbitos como la visión artificial, la comprensión de lenguaje natural y la traducción de voz.

El motivo de utilizar Scikit-learn [36], es ofrece algoritmos estándar, por ejemplo, algoritmos de clasificación como SVM, bosques aleatorios, regresión logística, a diferencia de TensorFlow [41], ya que este módulo esta diseñado para algoritmos de aprendizaje profundo, ya que al implementar este tipo de tareas, le permite aprovechar las GPU para una capacitación más eficiente.

Capítulo 4

Descripción de método propuesto

En este capítulo se propone un método para la compilación de un corpus de noticias falsas en español y un prototipo de sistema para la detección de noticias falsas que utilice un algoritmo de aprendizaje supervisado y el corpus creado. El prototipo realiza un procesamiento de lenguaje natural para para obtener características lingüísticas que ayuden a creación de un modelo para la detección de noticias falsas, además del uso de herramientas especializadas en el tratamiento de lenguaje natural como spaCy, Textacy y Scikit-learn.

4.1. Descripción del método propuesto

El método para la detección de noticias falsas se define en dos etapas. La primera corresponde a la obtención de un modelo que permite identificar aquellas noticias que son falsas, el modelo se obtiene mediante un algoritmo de aprendizaje automático y un módulo que realiza procesamiento de texto a partir de las acciones de lematización, tokenización y normalización de las palabras.

Para la creación del modelo se utilizan como datos de entrada la frecuencia de ocurrencia de algunos descriptores lingüísticos como lo son los n-gramas de palabras, n-gramas de caracteres y n-gramas de etiquetas POS. Posteriormente se realizan experimentos que permitan evaluar el desempeño de los descriptores de manera individual y combinándolos.

La segunda etapa hace uso del modelo creado para realizar la clasificación de noticias e identificación de noticias falsas, es decir, recibe como entrada una noticia que se sospecha es falsa y como salida se le asigna una etiqueta en la que se determina si es falsa o no. Al igual que los datos de entrenamiento, la noticia a verificar se somete a su procesamiento previo a la implementación del clasificador, donde el resultado obtenido a la salida corresponde a la clasificación asignada por el modelo.

En la figura 4.1 y la figura 4.2 se muestran los diagramas a bloques de la arquitectura del prototipo de sistema para la detección de noticias falsas para la fase de entrenamiento y la fase de prueba respectivamente. En las secciones subsecuentes se describe el método para la compilación del corpus de noticias falsas en español y se detallan los elementos del prototipo para su detección.

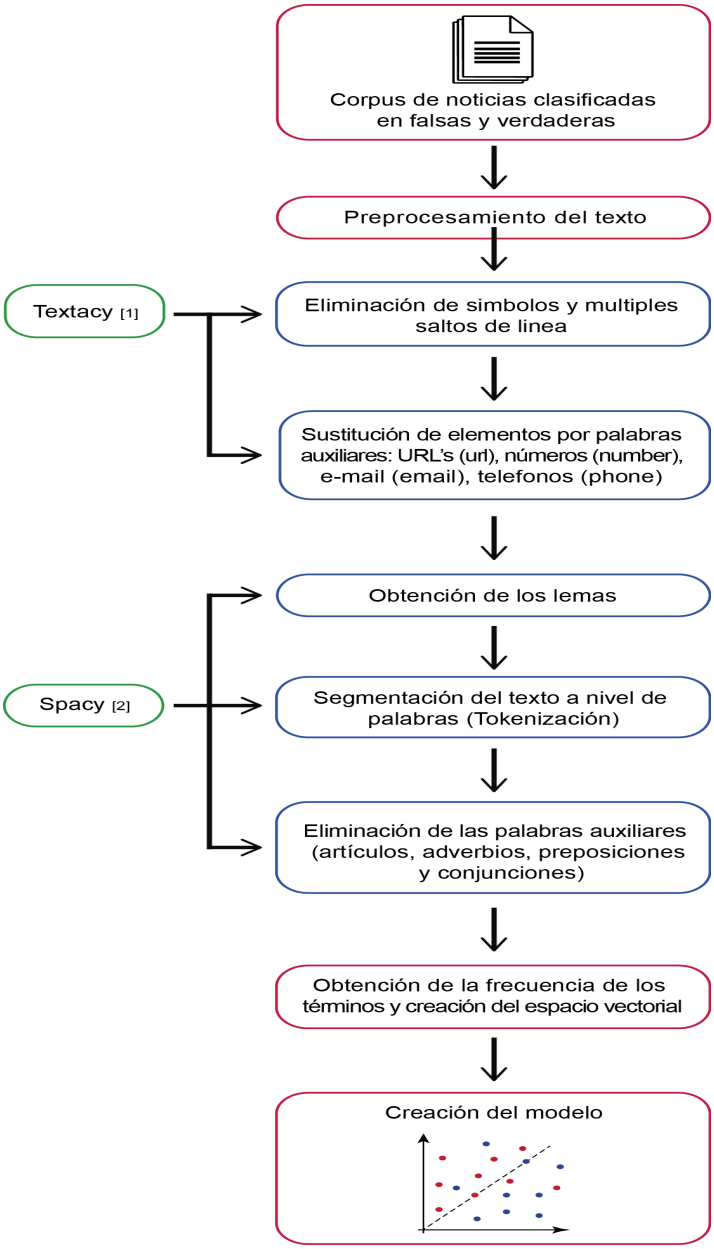
4.2. Método de compilación de corpus de noticias falsas en español

4.2.1. Búsqueda y clasificación de noticias

La utilización de un corpus es básico para cualquier estudio relacionado con la lingüística computacional y el procesamiento de lenguaje natural. El uso de un corpus como recurso permite obtener gran cantidad de información sobre el comportamiento real de la lengua y para ello es necesario su correcto diseño con bases estadísticas apropiadas que aseguren que se esté representando efectivamente el modelo de la realidad.

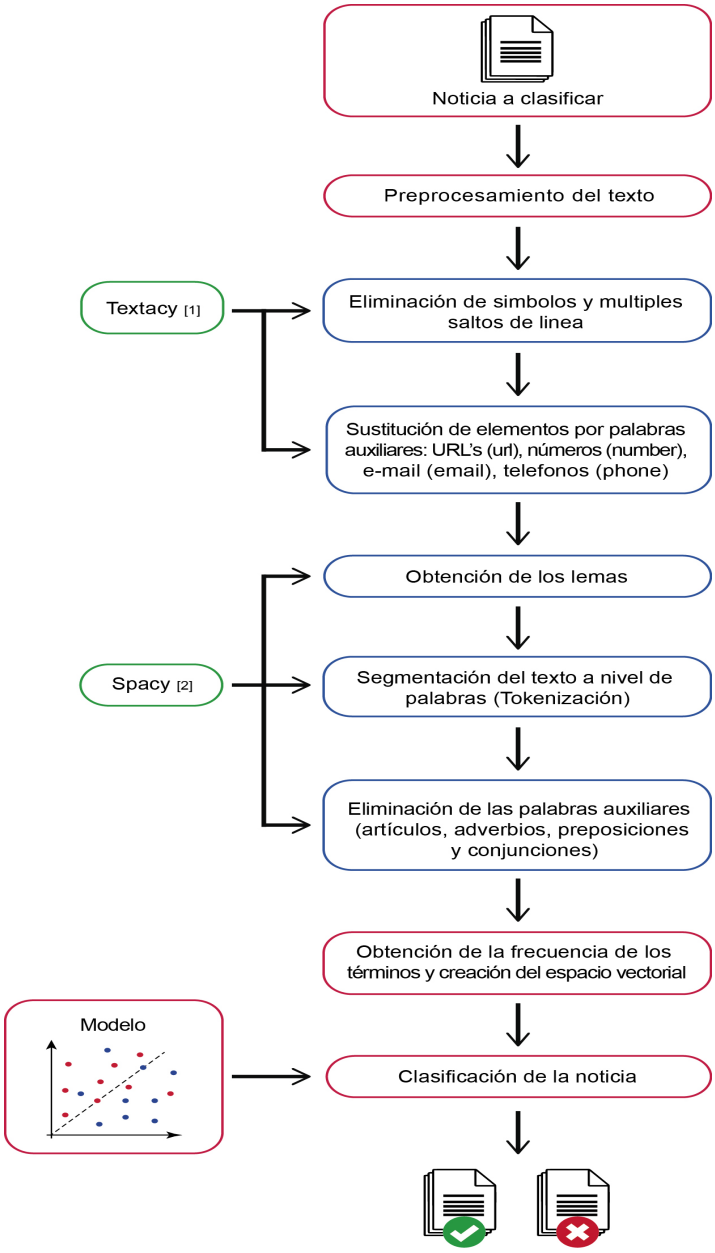
Existen numerosos trabajos relacionados con la compilación de corpus de noticias falsas en inglés, de varios tipos y tamaños. Sin embargo, para la lengua española la disponibilidad de corpus es escasa y esto limita la cantidad de trabajos de investigación realizada para este idioma.

En el presente proyecto se busca generar un corpus de noticias falsas en español, etiquetando de forma manual las noticias (verdadera o falsa). Adicionalmente se realiza un etiquetado de las noticias dependiendo de la temática principal que tienen las noticias. Para la creación del corpus se consideraron los siguientes aspectos:



[1] https://chartbeat-labs.github.io/textacy/
[2] https://spacy.io

Figura 4.1: Fase de entrenamiento



[1] https://chartbeat-labs.github.io/textacy/
[2] https://spacy.io

Figura 4.2: Fase de implementación

- **Variación del español**: Incluimos noticias de sitios Web de noticias de México, España y Chile. Con esta etiqueta se pretende propiciar la exploración del uso de la lengua española en diferentes países de habla hispana.

- **Sitio Web**: Identificamos las noticias escritas por sitios que se encargan de realizar noticias falsas sobre diferentes temas, de manera que se pretende motivar la investigación de atribución de autoría por parte de las páginas Web.

- **Clasificación**: Indicamos además el tipo de noticia de la que se está hablando (verdadero y falso) con el objetivo de hacer posible el uso de este corpus para investigación sobre identificación automática de noticias falsas.

La recopilación de noticias para la creación del corpus se realizó con base a una clasificación de verdadero y falso, a partir de una búsqueda en diversos sitios Web por medio del buscador de *Google*[1], realizada de la siguiente manera:

1. A través de plataformas que se encargan de desmentir manualmente noticias que se propagan en la Web como *VerificadoMX*[2], *Maldito Bulo*[3] y *Hoaxy*[4], se selecciona una noticia y se clasifica como falsa. Se eligieron estos sitios Web, ya que desmienten noticias con base en hechos que en realidad acontecieron y comprobando con información verídica las partes en las que una noticia fue manipulada (ver tabla 4.1).

2. A partir de sitios Web encargados de realizar noticias falsas en forma de sátira con el fin de entretener como: *Hay noticia*[8], *Mundo Today*[9] y *El Dizque*[10] (ver tabla 4.2), se selecciona una noticia y se clasifica como falsa. Se utilizó este tipo de plataformas, ya que sus noticias son creadas únicamente por ellos, tomando información destacada de una noticia verdadera y manipulando su contenido, agregando información ficticia, con el fin de burlarse de tal acontecimiento.

[1] `https://www.google.com`
[2] `https://verificado.mx`
[3] `https://maldita.es/malditobulo`
[4] `https://hoaxy.iuni.iu.edu`
[8] `https://haynoticia.es`
[9] `https://www.elmundotoday.com`
[10] `https://www.eldizque.com`

Tabla 4.1: Plataformas usadas para recopilación de noticias falsas

Sitio WEB	***VerificadoMX*** [5]	***Maldito Bulo*** [6]	***Hoaxy*** [7]
Objetivo	Si el equipo de Verificado 2018 detecta un sitio que sólo crea y difunde información falsa, desmentirá y exhibirá a este sitio para advertir a los lectores.Actualmente este sitio Web está inactivo, ya que su utilidad fue durante las elecciones pasadas.	Encontrar la información que circula en redes sociales y analizar el mensaje aplicando técnicas del periodismo de datos para su verificación.	Analizan los factores clave, incluidas las comunidades de redes, los intereses de los usuarios, la competencia, la atención finita, el sentimiento y las interacciones mutuas entre el tráfico y la estructura de la red.
¿Qué busca?	Información viralizada en medios y redes sociales, que cuente con más de mil interacciones o que se haya convertido en noticia para la opinión pública.	Proyecto periodístico independiente cuyo fin es dotar a los ciudadanos de "herramientas para que no te la cuelen".	Que se usen las herramientas para explorar cómo las ideas se propagan a través de las redes sociales en línea.
Clasificación	Revisan fuente de publicación, comparan la información con datos y hechos verídicos y los confirman con el protagonista de la noticia y testigos, anfitriones o asistentes a los eventos.	Desmienten las noticias por medio de una publicación en Twitter y captura de imagen de la noticia falsa que se esté hablando, para comparar la información.	Recopilan datos de blogs públicos y analizan el intercambio de información que brindan a los usuarios tendencias en línea y visualizar patrones temporales, de propagación de memes y actividad viral en las redes sociales.
Publican enlace	No	No	No
Categorías	Falso, Engañoso, No se puede probar y Verdadero	No cuentan con categorías solo desmienten la noticia	No cuentan con clasificación pues se dedican a factores que afectan la manera en que se dispersa la información.
Tipo de organismo	Público	Público	Privado

Tabla 4.2: Sitios que crean noticias falsas

Sitio WEB	**Descripción**	**Fecha y Lugar de creación**
Hay noticia[11]	Se denomina como un sitio de humor con la finalidad de entretener, mencionando que las referencias, nombres, marcas entre otros se usan como elementos contextuales, donde el contenido es ficción.	España
Mundo Today[12]	Es un sitio cuyo contenido es ficticio y humorístico, el cual usa el contexto de una noticia tradicional para crear una parodia o sátira. Es creado por Xavi Puig y Kike García.	España - 2009
Retroceso[13]	Es un sitio de noticias falsas principalmente elaboradas contra AMLO. El dominio de la página está registrado en Oregón, Estados Unidos, bajo un servicio de privacidad y proxy, que ayuda a camuflar la identidad de quien lo contrata. La identidad del supuesto autor fue suplantada al igual que la foto que acompaña la información de éste.	México - 2018
El Ruinaversal[14]	Sitio que presenta noticias con contenido falso a excepción de la categoría de "¡Increíblemente real!" que suponen ser verdaderas. La mayoría de las noticias se desarrollan en un contexto sobre México.	México - 2018
Argumento Político[15]	Es un sitio que supuestamente elabora noticias verdaderas, sin embargo, Verificado 2018 ha dado a conocer esta página como un sitio que reiteradamente comparte información falsa.	México - 2017
Dizque[16]	Es un sitio de noticias falsas que crea contenido absurdo y peculiar, pertenece a Kol.mx una agencia digital especializada en eLearning.	México - 2016
Censura 0[17]	Sitio que en año 2018 elaboró noticias falsas o ficticias, hoy en día la página se ha dado de baja.	Chile

3. A partir de plataformas que se encargan de realizar noticias falsas con el fin de desprestigiar la imagen de alguien o de algún otro sitio Web, así como noticias verdaderas para pasar desapercibidos como: *Retroceso*[18], *El Ruinaversal*[19], *Argumento Político*[20] y textitCensura 0[21] (Ver tabla 4.2), se selecciona una noticia. Ya que estos sitios Web no solo publican noticias falsas, es necesario hacer una verificación de que la noticia seleccionada es apta para la clasificación (ver figura 4.3):

3.1 Se realiza una búsqueda de referencias del título de la noticia seleccionada en diferentes motores de búsqueda: *DuckDuckGo*[22], *Bing*[23], *Yahoo!*[24]. Se eligieron estos buscadores ya que junto con *Google*[25], son conocidos y utilizados en la Web.

3.2 Se hace una revisión en los primeros 5 sitios Web que aparecen al realizar la búsqueda en cada uno de los buscadores mencionados anteriormente. Si por lo menos en uno de estos buscadores aparecen sitios Web como: *El Universal*[26], *BBC*[27], *Proceso*[28], *Animal Político*[29], *Aristegui Noticias*[30], *Forbes*[31], *Reporte Indigo*[32], *CNN en Español*[33], *HUFFPOST*[34], se clasifica la noticia seleccionada como verdadera. Se eligieron estos sitios Web, ya que se trata de sitios y periodistas profesionales que tienen una amplia trayectoria dentro de las comunicaciones, cuidando su reputación.

3.3 Si dentro de los primeros 5 sitios Web que aparecen al realizar la búsqueda en cada uno de los buscadores utilizados, no se encuentran ninguno de los sitios mencionados anteriormente, entonces:

[18] `https://retroceso.com`
[19] `https://www.elruinaversal.com`
[20] `https://www.argumentopolitico.com`
[21] `https://www.censura0.com`
[22] `https://duckduckgo.com`
[23] `https://www.bing.com`
[24] `https://espanol.yahoo.com`
[25] `https://www.google.com`
[26] `https://www.eluniversal.com.mx`
[27] `https://www.bbc.com/news`
[28] `https://www.proceso.com.mx`
[29] `https://www.animalpolitico.com`
[30] `https://aristeguinoticias.com`
[31] `https://www.forbes.com.mx`
[32] `https://www.reporteindigo.com`
[33] `https://cnnespanol.cnn.com`
[34] `https://www.huffingtonpost.com.mx`

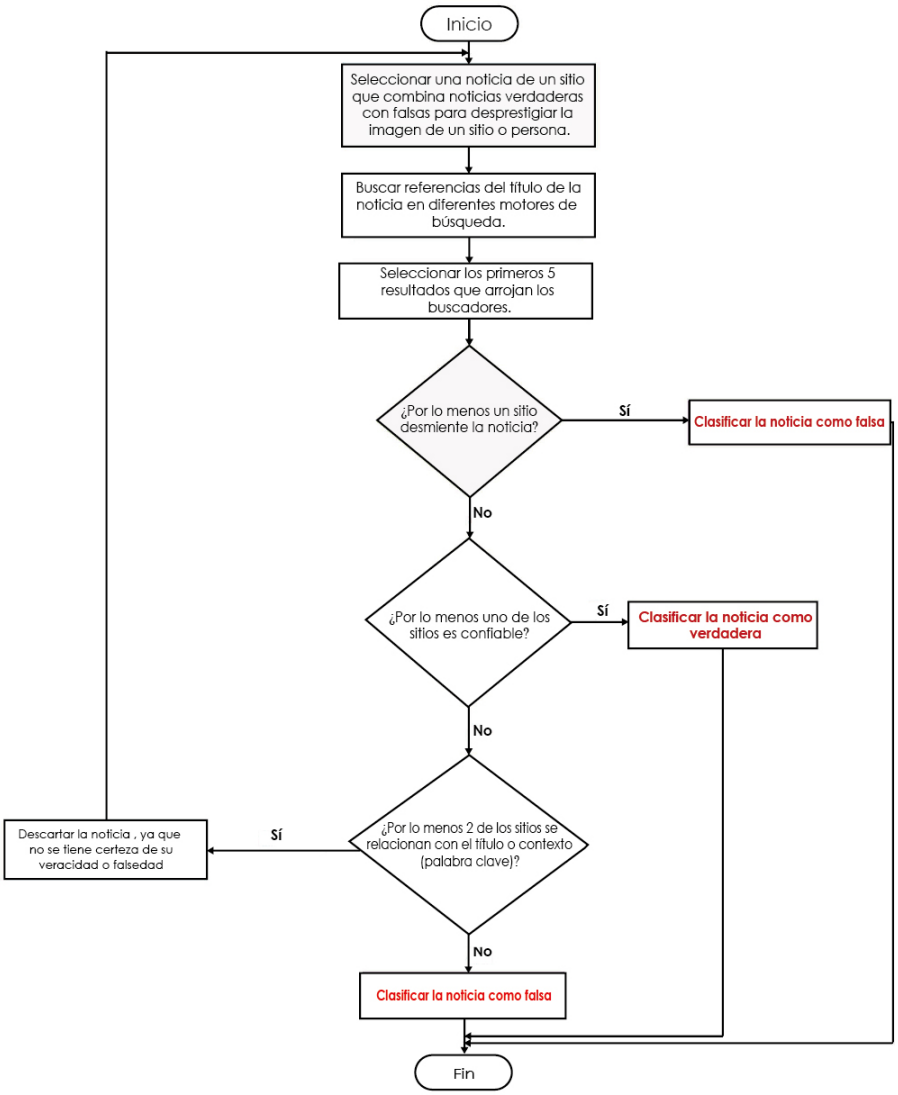


Figura 4.3: Búsqueda y clasificación en sitios Web que se encargan de realizar noticias falsas y verdaderas

3.3.1 Se hace una revisión en los primeros 5 sitios Web que aparecen al realizar la búsqueda en cada uno de los buscadores mencionados anteriormente , si por lo menos uno de estos se trata de sitios Web como: *VerificadoMX*[35], *Maldito Bulo*[36] y *Hoaxy*[37], se clasifica la noticia seleccionada como falsa.

3.3.2 Se hace una revisión en los primeros 5 sitios Web que aparecen al realizar la búsqueda en cada uno de los buscadores mencionados anteriormente, si al revisar su contenido, ninguno de estos se relaciona con el contexto y el título de la noticia seleccionada anteriormente, entonces se clasifica la noticia seleccionada como falsa, ya que es única y no se menciona sobre ella .

3.3.3 Se hace una revisión en los primeros 5 sitios Web que aparecen al realizar la búsqueda y si por lo menos 2 de los sitios que nos arrojan cada uno de los buscadores seleccionados se relacionan con el título o contexto (palabras clave) de la noticia seleccionada en la página dedicada a realizar noticias tanto falsas como verdaderas entonces esta noticia se descarta de la clasificación, pues no se tiene certeza de su veracidad o falsedad.

Una vez seleccionada la noticia falsa, para la creación del corpus se extrae de ella, el título, su contenido, la URL y se coloca la clasificación que se le asignó. Realizándolo de la siguiente manera:

1. En caso de seleccionar una noticia en sitios como: *VerificadoMX*[38], *Maldito Bulo*[39] y *Hoaxy*[40], se necesita buscar el enlace de la página que público dicha noticia, esto con el fin de obtener la información necesaria para anexarla al corpus, (Ver Figura 4.4).

2. En algunas publicaciones de las páginas que desmienten noticias, mencionan el nombre del sitio Web que se encargó de realizar dicha noticia falsa, si esto sucede, se realiza la búsqueda de la noticia dentro de la página proporcionada y se agrega la información extraída al corpus.

[35] `https://verificado.mx`
[36] `https://maldita.es/malditobulo`
[37] `https://hoaxy.iuni.iu.edu`
[38] `https://verificado.mx`
[39] `https://maldita.es/malditobulo`
[40] `https://hoaxy.iuni.iu.edu`

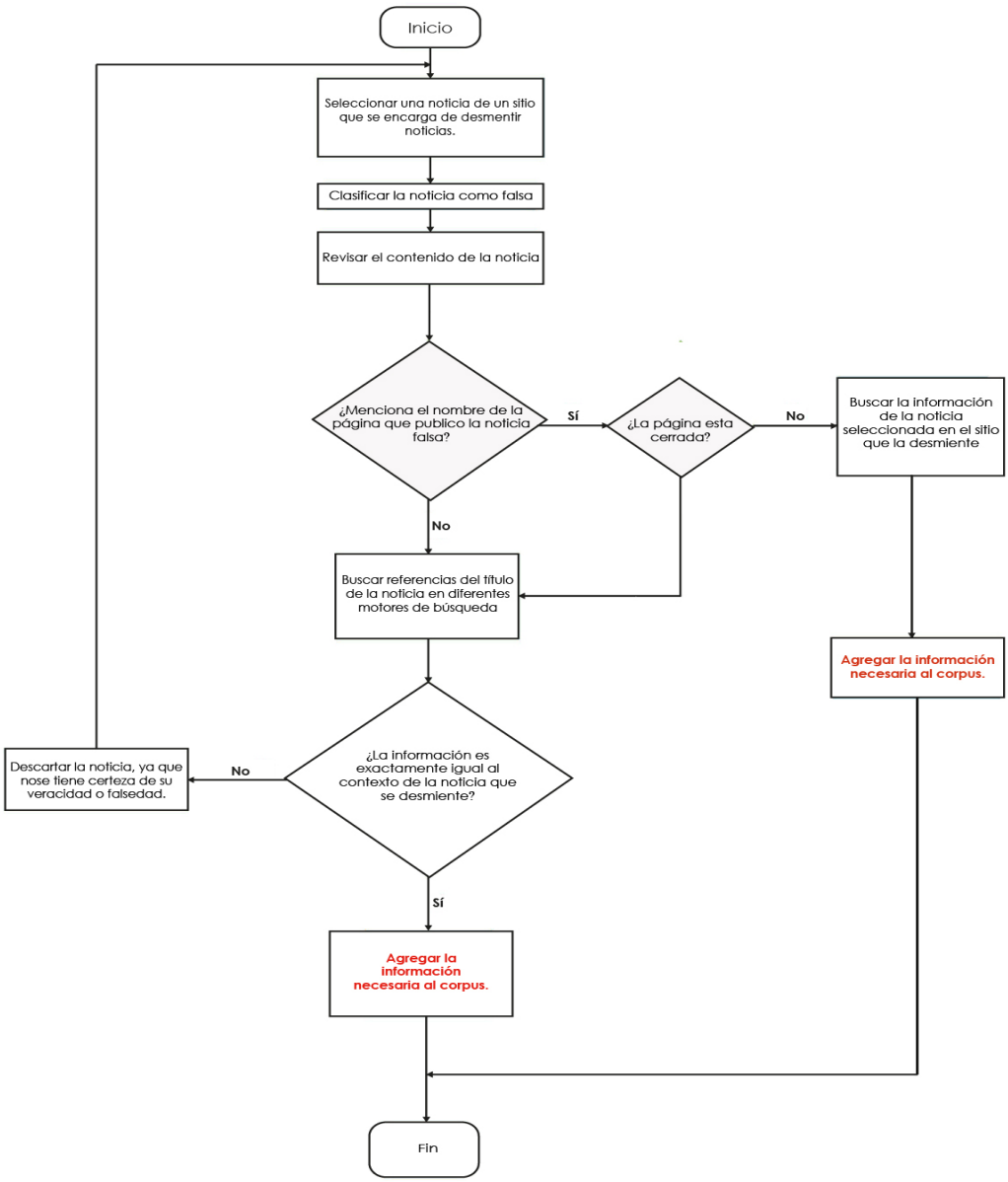


Figura 4.4: Búsqueda, clasificación y extracción en sitios Web que se encargan de desmentir noticias

3. Si la página que escribió la noticia falsa, no está incluida dentro de la información en la publicación que desmiente la noticia o cerraron dicha página, se realiza una búsqueda del título de la noticia seleccionada en diferentes motores de búsqueda: *DuckDuckGo*[41], *Bing*[42], *Yahoo!*[43], con el fin de hallar el sitio original que publicó la noticia y poder obtener la información necesaria para anexarla al corpus. Para este procedimiento se pueden obtener dos resultados:

3.1 Si al buscar la noticia en cada uno de los buscadores mencionados anteriormente, nos arroja como resultado una página que contenga dicha información exactamente con la noticia falsa descrita en la publicación que se encarga de desmentirla se extrae la información necesaria y se añade al corpus.

3.2 Si al buscar la noticia en cada uno de los buscadores seleccionados, ninguno de los sitios que nos arroja la búsqueda contiene la información exacta de la noticia que se está desmintiendo, entonces esta noticia se descarta.

4. En caso de haber seleccionado una noticia de sitios que se encargan de realizar noticias falsas en forma de sátira con el fin de entretener como: *Hay noticia*[44], *Mundo Today*[45] y *El Dizque*[46], se procede a obtener la información correspondiente para integrarla en el corpus.

5. En caso de haber seleccionado una noticia a partir de plataformas que se encargan de realizar noticias falsas con el fin de desprestigiar la imagen de alguien o de algún otro sitio Web, así como noticias verdaderas para pasar desapercibidos como: *Retroceso*[47], *El Ruinaversal*[48], *Argumento Político*[49] y *Censura 0*[50], y después de verificar que esta noticia es falsa, entonces se procede a extraer la información y se anexa al corpus.

[41] `https://duckduckgo.com`
[42] `https://www.bing.com`
[43] `https://espanol.yahoo.com`
[44] `https://haynoticia.es`
[45] `https://www.elmundotoday.com`
[46] `https://www.eldizque.com`
[47] `https://retroceso.com`
[48] `https://www.elruinaversal.com`
[49] `https://www.argumentopolitico.com`
[50] `https://www.censura0.com`

Para la creación del corpus de noticias verdaderas se toma como referencia la información dentro del corpus de noticias falsas, con el fin de nivelar las ocurrencias de los temas que se encuentran al realizar dicha clasificación por lo que se realizaron los pasos siguientes:

6. Se selecciona el título de una de las noticias dentro del corpus clasificada como falsa y se realiza una búsqueda en diferentes motores de búsqueda como: *DuckDuckGo*[51], *Bing*[52], *Yahoo!*[53], *Google*[54], esto con el fin de hallar una noticia que sea verdadera con un tema lo más relacionado posible a la noticia falsa seleccionada.

7. Si en los resultados de la búsqueda en cada uno de los buscadores seleccionados, se encuentran sitios como: *El Universal*[55], *BBC*[56], *Proceso*[57], *Animal Político*[58], *Aristegui Noticias*[59], *Forbes*[60], *Reporte Indigo*[61], *CNN en Español*[62], *HUFFPOST*[63], entonces esta noticia se clasifica como verdadera y se agrega la información necesaria al corpus.

8. En caso de que no encontrar la noticia verdadera de la noticia falsa seleccionada, se realiza una revisión del contexto (palabras claves) dentro de la información de la noticia falsa y del título para seleccionar una noticia verdadera que tenga relación con los temas tratados dentro de los sitios de confianza mencionados en el punto anterior, con el fin de nivelar ambos corpus y tener la mayor relación posible en el contenido de las noticias de cada uno de ellos.

4.2.2. Etiquetado temático de las noticias

Se realiza el etiquetado de noticias con el fin de dividir los temas que se mencionan en las diferentes noticias encontradas y evaluar el impacto que tienen las palabras para la realización

[51] `https://duckduckgo.com`
[52] `https://www.bing.com`
[53] `https://espanol.yahoo.com`
[54] `https://www.google.com`
[55] `https://www.eluniversal.com.mx`
[56] `https://www.bbc.com/news`
[57] `https://www.proceso.com.mx`
[58] `https://www.animalpolitico.com`
[59] `https://aristeguinoticias.com`
[60] `https://www.forbes.com.mx`
[61] `https://www.reporteindigo.com`
[62] `https://cnnespanol.cnn.com`
[63] `https://www.huffingtonpost.com.mx`

de la tarea de identificación de noticias falsas. La elección de estas etiquetas se atribuye a la clasificación realizada por los sitios de donde se obtuvieron la noticias.

En caso de que la noticia seleccionada no tuviera un etiquetado previo, entonces, se realiza un etiquetado manual donde 3 personas se encargarán de leer la información completa de la noticia en cuestión y definirán la etiqueta que se adecue más con la información que esta noticia contenga.

En la Tabla 4.3 se observa el registro de etiquetas que se eligieron de acuerdo a los temas de las noticias seleccionadas, así como el total de noticias compiladas dentro del corpus.

Tabla 4.3: Registro por categorías de noticias

Etiqueta	**Falsa**	**Verdadera**
Ciencia y Tecnología	43	46
Deporte	58	66
Economía	19	24
Educación	12	10
Espectáculos	78	70
Política	148	175
Salud	23	23
Seguridad	25	17
Sociedad	74	60
Total	480	491

Para determinadas combinaciones de etiquetas, se observó que las noticias de cada una de las categorías presentaban una correlación temática baja debido a que los temas eran ajenos entre sí. Sin embargo se observó que al utilizar noticias publicadas en el periodo electoral se observo que algunas combinaciones de categorías que nos esperaban estar relacionadas mostraron afinidades, tal es el caso de las categoría de Política y Espectáculos, que se pueden explicar como consecuencia de alguna eventualidad como lo fueron las elecciones. En México hubo noticias que hablaban de candidaturas de actores y actrices para ocupar cargos de

presidentes municipales, otro ejemplo es que existen noticias falsas que hablan de rumores sobre la primera dama Angélica Rivera de Peña.

Una vez terminado el proceso de etiquetado se realizó una división de noticias en una proporción de 70/30, esto se debe a que para cada categoría o etiqueta dada en el corpus, es necesario realizar un proceso aleatorio en donde se eligen noticias al azar para pertenecer a cada sector y realizarse pruebas y experimentos descritos en los puntos siguientes, evitando con esto un porcentaje alto en el sobre ajuste.

4.3. Procesamiento del texto

Para la creación del modelo de espacio vectorial es necesario realizar un preprocesamiento al corpus de noticias con el fin de obtener una representación homogénea de las palabras que componen los textos. En este módulo se plantea el uso de herramientas que permitan simplificar la forma de las palabras sin afectar su estructura semántica y sintáctica, por ejemplo, la segmentación de los textos a nivel de palabras, así como su debida representación a partir de su forma canónica (lema). También se implementa el enmascaramiento de algunas palabras y elementos textuales como URL's, números, e-mails, símbolos de moneda, así como la eliminación de los signos de puntuación y los múltiples saltos de línea implícitos en las noticias. A continuación, se describen de manera particular cada una de las tareas correspondientes al preprocesamiento de las noticias.

Debido a que la obtención de las noticias del corpus se realizó de medios digitales, es común encontrar datos relacionados con los sitios de divulgación, así como metadatos de archivos multimedia o enlaces a fuentes externas en el interior de los textos a analizar. Por esta razón se propone un módulo de normalización con la finalidad de eliminar información innecesaria para la tarea de clasificación. Se propone el desarrollo de un módulo para la normalización de los textos haciendo uso de la herramienta Textacy.

La normalización de los textos refiere a la identificación de aquellos elementos en el texto de las noticias del corpus que al no estar en su forma estándar no aportan información alguna para el proceso de clasificación, estos elementos son las URL's, los números, símbolos y los signos de puntuación, así como los múltiples espacios en blanco y saltos de línea. Se

propone como proceso de normalización del texto realizar el enmascaramiento (sustitución) o eliminación de estos términos, de forma que al final de este proceso sean detectados mediante una misma nomenclatura. A continuación, se ejemplifica la tarea de normalización para algunos de los elementos mencionados previamente:

Texto no normalizado
Precio del metro aumentará a $5.50 para financiar la reintegración de los "vagoneros". Corroborar si estás afiliado a algún partido político te toma menos de tres minutos y aquí te explicamos el paso a paso. @VerificadoMx pic.twitter.com/2iGNhcDwFx

Texto normalizado
Precio del metro aumentará a USD NUMBER para financiar la reintegración de los vagoneros Corroborar si estas afiliado a algún partido político te toma menos de tres minutos y aquí te explicamos el paso a paso VerificadoMx URL

El proceso de enmascaramiento considera los siguientes elementos:

1. **URL's**: Cualquier cadena de caracteres con el formato de una página Web será sustituida por la palabra URL.
2. **E-mails**: Cualquier cadena de caracteres con el formato de un email, es sustituida por la palabra EMAIL.
3. **Números telefónicos**: Se reemplaza cualquier número con formato telefónico por la cadena PHONE.
4. **Números**: Se reemplaza cualquier número sin formato alguno por la cadena NUMBER (Considera aquellos con punto flotante, y enteros).
5. **Símbolos de moneda**: Reemplaza los símbolos de moneda existentes en el texto con sus abreviaciones estándar de 3 letras, por ejemplo, USD, EUR, entre otros.
6. **Signos de puntuación**: Todos los signos de puntuación (excepto acentos) son reemplazados por un espacio en blanco.

7. **Espacios en blanco**: Se reemplazan los múltiples saltos de línea y espacios en blanco por un espacio/salto de línea simple.

Mediante el proceso de enmascaramiento se logra obtener una representación uniforme de las palabras. Las palabras acentuadas no se ven afectadas por este proceso de filtrado.

La tokenización realiza la segmentación de los textos en sus unidades léxicas independientes más pequeñas, es decir, en palabras, debido a que a partir de estos elementos básicos es posible obtener características para modelar las noticias. La segmentación o división del texto a nivel de palabras se realiza mediante el uso del módulo spaCy y su modelo de lenguaje para el idioma español mediante el uso de los espacios entre las palabras para realizar la tarea.

La segmentación se realiza a partir de los textos normalizados. El módulo spaCy es capaz de identificar *tokens* complejos como abreviaciones y palabras que dependen de elementos textuales como los prefijos, sufijos e infijos, en algunas noticias del corpus los símbolos utilizados para representar estos elementos no corresponden a los símbolos estándar que spaCy utiliza. A continuación, se enlistan algunos de los prefijos, sufijos e infijos utilizados por el módulo spaCy:

1. **Prefijos**: carácter(es) al inicio de las palabras, por ejemplo $, (, “, ¿.

2. **Sufijos**: carácter(es) al final de las palabras, por ejemplo km,),”,!.

3. **Infijos**: carácter(es) entre las palabras, por ejemplo -, –, /.

En ocasiones se utilizan caracteres alternos a los caracteres estándar en la composición del texto. Algunos ejemplos comunes referentes a la codificación de los símbolos en la extracción de las noticias del corpus se mencionan a continuación:

1. Uso de la prima (′) y doble prima (′′) para representar las comillas simples (‘) y comillas dobles (“).

2. Uso de la prima (′) para representar los apostrofes (‘)

3. Uso del acento grave (‘), esto considerando que en el idioma español se utiliza exclusivamente el acento agudo (´)

Después de realizar el proceso de copiado de información de un sitio Web al ordenador, se debe aplicar la tokenización al texto para evitar representaciones diferentes entre las palabras de los textos.

A continuación, se ejemplifica el proceso de tokenización a partir de un texto previamente normalizado:

Texto normalizado
Precio del metro aumentara a USD NUMBER para financiar la reintegración de los vagoneros

Texto tokenizado
Precio, del, metro, aumentará, a, USD, NUMBER, para, financiar, la, reintegración, de, los, vagoneros

Por lo tanto, con la tokenización se obtienen los términos representados de forma homogénea, que serán usados posteriormente para obtener una representación vectorial en términos de la frecuencia de ocurrencia de las características lingüísticas de los textos.

A pesar de que con la normalización y tokenización se obtiene una representación uniforme y generalizada de los textos a partir de los elementos textuales y su enmascaramiento, es necesario realizar la tarea de lematización para extender esta homogeneidad a nivel de las palabras mediante la extracción de su forma canónica.

Como se ha mencionado anteriormente, con el proceso de lematización se obtiene la forma canónica (lema) de las palabras flexionadas, como lo son las palabras escritas en plural, en femenino, los verbos conjugados, entre otras. El lema representa la raíz o lexema de una palabra, y con estos se puede obtener la relación lingüística entre palabras que implícitamente poseen una misma estructura morfológica, de esta manera es posible unificar la variación gramatical de las palabras y llevarlas a una sola nomenclatura.

Al igual que en el módulo correspondiente a la tokenización, para esta tarea se usó la herramienta spaCy. El modelo utilizado por spaCy para el proceso de lematización está basado en

una combinación de tablas de búsqueda y reglas gramaticales, las cuales permiten identificar las diferencias sintácticas a partir de la forma de las palabras.

El algoritmo para la lematización se define de la siguiente manera:

1. Se extrae la etiqueta gramatical (POS) correspondiente a cada *token* y se realiza una primera comparación, donde si la etiqueta POS no representa a elementos que varíen su morfología como los sustantivos, verbos, adjetivos, entre otros, el *token* no tiene la necesidad de ser lematizado.

2. Se verifica que las palabras no están flexionadas, en caso de que no lo estén se retorna la misma palabra.

3. Se verifica que las palabras no se encuentren clasificadas como irregulares dentro de los diccionarios de palabras de spaCy, es decir, que su lema tenga una morfología diferente, por ejemplo, las palabras derivadas de la conjugación del verbo *ser*. En caso de que sea una palabra irregular, se realiza la búsqueda de su lema a partir de su etiqueta POS en el listado correspondiente a este tipo de palabras.

4. Si los *tokens* contienen sufijos gramaticales que se encuentren enlistados en los metadatos de spaCy, se realiza una segunda comparación. Para esto se hace una búsqueda de los *tokens* en los diccionarios de palabras de spaCy asociados a las etiquetas POS, omitiendo el segmento del sufijo, si la palabra coincide con alguno de los elementos en el diccionario se realiza la lematización. En caso de no coincidir se realiza nuevamente la búsqueda, añadiendo uno por uno los caracteres pertenecientes al sufijo, si no se encuentra coincidencia alguna, se define que el *token* ya está en su forma canónica.

5. Si no es ninguno de los casos anteriores, se considera que los *tokens* están en su forma canónica.

Por ejemplo, la palabra “aumentará” tiene asociada la etiqueta gramatical referente a un verbo (VERB). Debido a que el verbo esta conjugado, se considera que esta flexionada, y a su vez, de manera empírica se infiere que no es una palabra irregular. Aplicando las reglas gramaticales definidas para los verbos en el módulo spaCy, se identifica a la partícula “ará” como sufijo para el verbo, por lo que la búsqueda del lema se hace a partir del segmento “aument”, sin embargo, esta palabra no existe en el idioma español, por lo que se le añade

el primer carácter del sufijo, siendo la palabra “aumenta” el resultado. Si bien esta palabra ya es existente en el léxico del idioma, esta nueva representación sigue estando flexionada, por lo que se le añade el siguiente carácter del sufijo. De esta manera se obtiene la palabra “aumentar”, la cual es la forma canónica del verbo original.

Al aplicar el algoritmo de lematización a un texto previamente tokenizado el resultado sería el siguiente:

Texto tokenizado
Mexicanos, del, norte, apoyan, dar, trabajo, a, migrantes, los, rechazan, en, el, Occidente, y, bajío, Mitofsky

Texto lematizado
***Mexicano**, del, norte, **apoyar**, dar, **trabajar**, a, **migrante**, **lo**, **rechazar**, en, el, Occidente, y, bajío, Mitofsky*

Posterior a la lematización, se realizó la conversión de cada una de las letras que conforman las palabras a minúsculas, utilizando el método *lower ()* de la biblioteca estándar de Python. Este proceso se realizó, ya que, al obtener la frecuencia de los términos, si una misma palabra varia su escritura incluso en función de una sola letra, esta es considerada como un elemento completamente diferente al compararla con las otras palabras. Este proceso se plantea posterior a la lematización y se consideran casos como el de las primera palabras de una oración, que por reglas sintácticas su primera letra debe de ser mayúscula y que la misma palabra aparece en minúsculas al ser usada en otras partes de la oración, sin embargo, corresponde a la misma palabra. En el caso de los nombres propios el realizar la conversión a minúsculas antes de la lematizaión puede ocasionar mala detección del lema de la palabra.

Al final de este proceso se obtiene un vocabulario homogéneo, sin que estos cambios representen una modificación a la estructura sintáctica de los textos contenidos en el corpus. Entonces el resultado final de esta tarea seria el siguiente:

Las palabras auxiliares o también denominadas *stopwords*, son aquellas palabras que no proporcionan información relevante acerca del contenido de los textos, estas palabras dependen del idioma, por ejemplo, en español se consideran los artículos, conjunciones, preposiciones,

Texto tokenizado
Mexicanos, del, norte, apoyan, dar, trabajo, a, migrantes, los, rechazan, en, el, Occidente, y, bajío, Mitofsky

Texto lematizado
***Mexicano**, del, norte, **apoyar**, dar, **trabajar**, a, **migrante**, **lo**, **rechazar**, en, el, Occidente, y, bajío, Mitofsky*

Texto procesado
mexicano**, del, norte, apoyar, dar, trabajar, a, migrante, lo, rechazar, en, el, **occidente**, y, bajío, **mitofsky

adverbios y algunos verbos que estadísticamente son muy comunes. Es decir, las palabras auxiliares son aquellas que generalmente tienen una alta ocurrencia en los textos.

Existen módulos en Python como NLTK y spaCy que permiten identificar las palabras auxiliares. Estas herramientas se basan en listas de palabras, sin embargo, NLTK no se encuentra totalmente optimizada para el idioma español, por lo que el número de palabras vacías que contiene se encuentran relativamente limitadas, siendo un total de 313. Por otro lado, spaCy cuenta con 551 palabras auxiliares, lo cual representa un filtro de mayor alcance. Debido a esta característica, se hace uso del módulo de spaCy para la extracción de las palabras auxiliares. Las palabras para cada uno de estos dos módulos se localizan en el anexo X.

Aunque el conjunto de palabras que contiene spaCy permite considerar una cantidad mayor de palabras, al intentar procesar los textos con esta herramienta se encontraron palabras que no se consideraban como auxiliares. Para solucionar este problema se plantea la elaboración de una lista complementaria de palabras auxiliares para suprimir las excepciones de spaCy.

En la tabla 4.4 se muestra una lista de las palabras auxiliares propuestas para complementar el listado proporcionada por spaCy.

El filtrado de las palabras auxiliares se realiza a partir de comparaciones, donde se verifica que cada palabra o token del texto no se encuentre en el conjunto de palabras etiquetadas

Tabla 4.4: Lista complementaria de palabras auxiliares

Palabras auxiliares propuestas			
a	ante	bajo	con
de	desde	durante	en
entre	excepto	hacia	hasta
mediante	para	por	salvo
según	sin	sobre	tras

como palabras auxiliares por spaCy, en caso de que se encuentre una similitud directa, el token correspondiente se suprime del texto.

Texto procesado
mexicano*, del, norte, apoyar, dar, trabajar, a, migrante, lo, rechazar, en, el,* ***occidente****, y, bajío,* ***mitofsky***

Texto sin palabras auxiliares
mexicano, norte, apoyar, migrante,
rechazar, occidente, bajío, mitofsky

De acuerdo a la información expuesta marco teórico, después de eliminar las palabras auxiliares se obtienen solo las palabras que aporten información referente al contenido de las noticias, sin embargo, a partir de los datos experimentales se observa que en aquellos experimentos donde no se suprimen las palabras auxiliares para el modelado del sistema, los clasificadores tienen un mayor porcentaje de precisión. A partir del análisis manual de las noticias en el corpus, se logra determinar que este comportamiento en los clasificadores se debe a que en la redacción de las noticias falsas se encuentran patrones comunes, como lo son el uso de las palabras auxiliares como conectores gramaticales, así como el uso de números.

Considerando los datos obtenidos, es posible argumentar que las palabras auxiliares son de ayuda en el proceso de clasificación, por esta razón no se realiza su eliminación en el módulo correspondiente al procesamiento del texto.

4.4. Extracción de características

Previo al procesamiento del texto, es necesario definir las características que representaran a los documentos dentro de la tarea de clasificación. Si se considera la definición de un texto como un conjunto de palabras, se puede inferir que las características que describen el contenido de un documento son precisamente las palabras, ya que como se ha mencionado previamente en el módulo de preprocesamiento, las palabras son los elementos básicos que definen a los textos.

Las palabras también se pueden representar en función de herramientas lingüísticas como lo son los n-gramas de palabras, n-gramas de caracteres, así como de sus etiquetas gramaticales también denominadas etiquetas POS, por lo que el número de características se extiende. En esta sección se propone el uso de estas herramientas lingüísticas de manera individual, así como la combinación de ellas, para la posterior representación vectorial de los textos de las noticias en función de estos parámetros.

En el esquema más básico, cada una de las palabras en los textos se considera una característica, donde los valores que representan a estas características se definen a partir su frecuencia de ocurrencia, también denominada como ponderación *tf* (frecuencia de termino), es decir, las características toman como valor el número de veces que determinada palabra está presente en determinado texto, siendo, siendo relativamente las palabras más recurrentes las que se consideran de mayor importancia para el proceso de clasificación.

Debido a que el proceso de clasificación se basa en la identificación de palabras relevantes para cada clase presente en el corpus (noticias verdaderas y noticias falsas), el hecho de que algunas palabras sean recurrentes en ambas clases puede afectar negativamente al sistema de clasificación, de acuerdo con el valor *idf*, se considera que estas palabras no aportan información relevante para ninguna de las clases. En estos casos, las palabras tendrían un comportamiento similar al de las palabras auxiliares. De la misma manera, si una palabra es común solo en alguna clase específica, esta palabra resulta ser de utilidad para realizar la clasificación.

Un patrón que se identificó en el corpus referente a las noticias falsas hace alusión al caso previamente expuesto, donde en las noticias extraídas del sitio Web *Dizque*[64], los autores hacen repetida mención de la palabra *dizque* para hacer énfasis en que la información es exclusiva de este sitio. Otra pauta encontrada refiere a la cantidad de palabras auxiliares utilizadas en la redacción de las noticias falsas, donde estas palabras son mas recurrentes en esta categoría en comparación con las noticias etiquetadas como verdaderas. Estas características permiten aumentar la precisión de los clasificadores utilizados.

Patrón textual encontrado en las noticias falsas

"Estaba platicando con Eli Roth", confesó
Tarantino en entrevista para El ***Dizque****.*
"Y me reclamó que en mis últimas cintas
le he bajado el tono a la violencia."

Los patrones se identificaron a partir de la implementación de dendogramas como herramientas gráficas, los cuales contienen información sobre aquellas palabras que predominan al momento de realizar la tarea de clasificación.

En la figura 4.5 se muestra un ejemplo de los dendogramas obtenidos utilizando el algoritmo de árboles de decisiones del módulo *Sci-Kit Learn* de Python, el cual identifica las características más significativas y su valor que permite identificar sus relaciones con otras características, con las cuales se obtienen los mejores conjuntos de información para el proceso de clasificación. Es decir, permite obtener nuevas características con mayor poder para clasificar los datos de prueba. El índice X[NUM] es el identificador de las palabras respecto a su posición en la matriz de características de los documentos igualado a su valor en el espacio vectorial. Las ramas indican las relaciones entre las características.

Para el caso particular de la palabra dizque, esta palabra permitía clasificar un alto porcentaje de las noticias de prueba. Sin embargo, debido a que esta palabra solo estaba presente en las

[64] `https://www.eldizque.com`

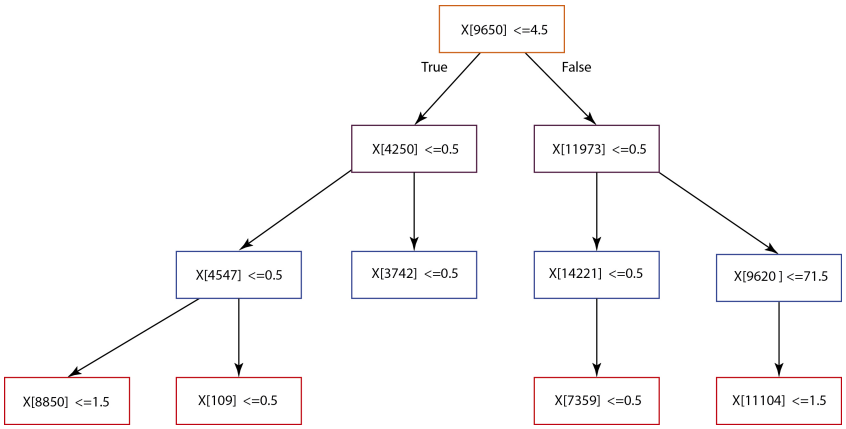


Figura 4.5: Ejemplo de dendograma

publicaciones relacionadas con el sitio Web *Dizque*[65] se tuvo que suprimir, esto considerando que si se introducía algún texto que no tuviera relación con este sitio Web, pero que en su contenido se incluía esta palabra, automáticamente se le clasificaba como falso debido al peso de la palabra *dizque* en las noticias falsas.

Así como las palabras son consideradas como las características básicas de los textos, también las representaciones obtenidas a partir de la implementación de las herramientas lingüísticas previamente mencionadas pueden ser consideradas como tal.

Las etiquetas de clase gramatical, son aquellas etiquetas que permiten definir la función gramatical o categoría léxica de las palabras. Haciendo uso de estas etiquetas se puede identificar la coherencia gramatical entre las palabras a partir de la probabilidad de que un tipo de palabra sea precedida por otra, usando como identificador la etiqueta gramatical asignada a cada una de ellas. Por ejemplo, la probabilidad de que un verbo sea precedido por un sustantivo es mayor que la probabilidad de que un verbo sea precedido por adverbio. Debido a esta coherencia gramatical, se propone como experimento el modelado del sistema haciendo uso de las etiquetas gramaticales de cada una de las palabras en los textos del corpus, al igual que en el caso donde se utilizaron las palabras como características los valores están en función de la frecuencia de los términos.

[65] `https://www.eldizque.com`

Otro tipo de características que se pueden considerar son las combinaciones de palabras, así como las de sus etiquetas gramaticales de acuerdo al orden de aparición de las palabras en los documentos. Estas secuencias se obtienen a partir de la implementación de los n-gramas de palabras, donde la letra "n" representa la cantidad de palabras en las combinaciones. Con el uso de los n-gramas se obtienen aquellas combinaciones de palabras, que estadísticamente son más frecuentes en la composición de los textos del corpus. Para la creación del espacio vectorial, se propone la implementación de bigramas, trigramas, 4-gramas y 5-gramas como características de los textos.

Unigramas de palabras
mexicano, del, norte, apoyar, dar, trabajar, a, migrante, lo, rechazar, en, el, occidente, y, bajío, mitofsky

Bigramas de palabras
mexicano del, del norte, norte apoyar, apoyar dar, dar trabajar, trabajar a, a migrante,migrante lo, lo rechazar, rechazar en, en el, el occidente, occidente y, y bajío, bajío mitofsky

Unigramas de palabras utilizando las etiquetas POS
ADJ, ADP, NOUN, AUX, VERB, VERB, ADP, ADV, DET, VERB, ADP, DET, NOUN, CONJ, NOUN, NOUN

Bigramas de palabras utilizando las etiquetas POS
ADJ ADP, ADP NOUN, NOUN AUX, AUX VERB, VERB VERB, VERB ADP, ADP ADV, ADV DET, DET VERB, VERB ADP, ADP DET, DET NOUN, NOUN CONJ, CONJ NOUN, NOUN NOUN

Haciendo uso de los n-gramas de palabras como características, es posible definir patrones textuales y de autoría referentes al uso de expresiones y la frecuencia de estas, por ejemplo, al hacer uso de la expresión "sin embargo" de manera particular como conector gramatical de contraste al redactar los textos, en lugar de algunas otras expresiones con el mismo significado, como "no obstante". Entonces se puede inferir que este tipo de características permite identificar patrones de redacción.

La frecuencia de ocurrencia de los n-gramas de palabras se ve limitada en comparación con la de los unigramas (palabras), por lo que el clasificador recibe menos valores para la asignación de clases, sin embargo, esto no afecta el proceso debido a que el léxico utilizado en la redacción de las noticias falsas difiere en cierto grado con el de las noticias verdaderas, por lo que los n-gramas permiten obtener resultados óptimos en los clasificadores.

Algunos sitios Web difusores de noticias falsas como *Retroceso* [66], *El Ruinaversal*[67], se identificaron como sitios en los que un mismo autor es quien redacta el contenido de la mayoría de las publicaciones. Por ejemplo, en el caso de *Retroceso* [68], *El Ruinaversal*[69], *VerificadoMX*[70], lo identifico como un sitio en el que un mismo autor es el encargado de la redacción de las noticias publicadas en el sitio Web. Debido a esta observación, se propone el uso de herramientas lingüísticas que permitan inferir patrones para identificar este tipo de situaciones en las noticias del corpus.

Los n-gramas de caracteres son una herramienta que propone hacer uso de las secuencias de caracteres de las palabras que conforman los textos, como características que los representen en el espacio vectorial. Esta herramienta lingüística es utilizada en tareas de atribución de autoría, la cual se basa en la identificación de los prefijos y sufijos mas comunes que utiliza un autor al momento de redactar, y de esta manera realizar comparaciones para determinar el estilo de escritura.

Texto lematizado
mexicano, del, norte

Tigramas de caracteres
*mex, exi,xic,ica,can,ano, no_, o_d, del, el_, l_n, _no, nor, ort, rte, te_*

Al igual que en el caso de los n-gramas de palabras, se utilizan n-gramas de tamaño 2, 3, 4 y 5. Debido a que en los n-gramas de caracteres los espacios entre las palabras son importantes

[66] `https://retroceso.com`
[67] `https://www.elruinaversal.com`
[68] `https://retroceso.com`
[69] `https://www.elruinaversal.com`
[70] `https://verificado.mx`

para determinar los prefijos y sufijos, se añadió el símbolo "_"para representar este espacio entre los *tokens* del texto.

La implementación de los n-gramas de caracteres arrojo buenos resultados al entrenar el clasificador, permitiendo inferir la existencia de patrones referentes al estilo de redacción en las noticias falsas además de los mencionados previamente.

De manera general, al considerar las características de manera individual se obtuvieron resultados óptimos al evaluar la precisión del sistema. A partir de estos resultados se plantea evaluar el sistema con el conjunto de características consideradas previamente, esto es crear el espacio vectorial a partir de la concatenación de las características individuales (etiquetas POS, palabras, n-gramas de palabras y n-gramas de caracteres).

En primera instancia se plantea la creación del espacio vectorial a partir de las siguientes consideraciones:

1. Realizando la concatenación de los n-gramas de palabras $(n = 2, 3, 4, 5)$.
2. Realizando la concatenación de los n-gramas de palabras utilizado etiquetas POS $(n = 2, 3, 4, 5)$.
3. Realizando la concatenación de los n-gramas de caracteres $(n = 2, 3, 4, 5)$.

Sin embargo los resultados experimentales de estas pruebas resultaron ser poco favorables para la clasificación de las noticias.

Considerando todas las características individuales es posible incluir cada uno de los casos posibles al entrenar el modelo, esto es, se considera la frecuencia individual de los términos, la frecuencia de las combinaciones de palabras, así como de las relaciones gramaticales al usar las etiquetas POS, así como los patrones relacionados con el estilo de redacción de los documentos. A partir de la suma de las características el modelo presenta mejoras, por lo que se infiere que los patrones encontrados en cada una de las clases también son identificados al realizar estas pruebas.

Si bien las características individuales proporcionan información necesaria para realizar la clasificación correcta de las noticias, es importante realizar la evaluación del modelo considerando cada una de las características extraídas de los textos. Las características en su

totalidad permiten que el modelo se base en la estructura interna de los documentos y se logre extraer la mayor cantidad de información posible relacionada con la elaboración de las noticias falsas, de forma que se logre maximizar la precisión del clasificar de manera objetiva.

4.5. Creación del modelo

Algunas tareas implicadas en el área de procesamiento del lenguaje natural se basan en los modelos de aprendizaje automático para dar solución a las problemáticas planteadas. El proceso de clasificación binario propuesto, tiene como fundamento el uso de herramientas de aprendizaje supervisado, donde los valores de las características se modelan para su representación en forma de espacio vectorial.

Para el caso particular de este proyecto, los datos para el entrenamiento y prueba corresponden al corpus de noticias. El sistema analizara el conjunto de datos de prueba para inferir estadísticamente la clase a asignar a nuevos datos de entrada, es decir si la noticia es falsa o verdadera.

De acuerdo a la información del marco teórico, en los modelos de aprendizaje supervisado los documentos deben estar etiquetados con sus respectivas clases, en este caso, al ser un proceso de clasificación binario, se utilizan las siguientes etiquetas:

- 0 para identificar las noticias falsas.
- 1 para identificar las noticias verdaderas.

Al tener los textos procesados y etiquetados, se divide el corpus en subconjuntos de entrenamiento y de prueba esto es a partir de la relación 70-30 propuesta en el estado del arte, esto para evitar el sobreajuste de los datos en la etapa experimental. Posteriormente se entrena el modelo haciendo uso de los atributos de los documentos del conjunto de entrenamiento, mientras que la evaluación respecto a la precisión del sistema se realiza utilizando el conjunto de pruebas.

El estado del arte proporciona información respecto a que algoritmos de clasificación tienen mayor índice de efectividad en tareas referentes al PLN. A partir de esta información se definen los siguientes clasificadores para el desarrollo del sistema:

1. Regresión Logística
2. *Support Vector Machines (SVM)*
3. *Random Forest*
4. *Boosting*

Para la creación del modelo se hizo uso del módulo *Sci-Kit Learn*, el cual provee una colección de algoritmos para el desarrollo de herramientas basadas en aprendizaje automático supervisado para realizar tareas de clasificación. Además de la colección de algoritmos, *Sci-Kit Learn* permite la transformación de los datos a su representación vectorial.

Como se plantea en la sección de extracción de características, los valores de las características se definen en función de la frecuencia de aparición de los términos, esto a su vez permite representar los textos del corpus en el dominio numérico para su posterior transformación a un formato vectorial. El resultado de representar los documentos como vectores es la matriz denominada matriz termino-documento. Debido a la variación en el léxico utilizado para la redacción de las noticias, así como en la extensión de los textos, la matriz termino-documento obtenida tiene el formato de una matriz dispersa.

El modelo de bolsa de palabras (BoW- Bag of Words) es el método utilizado para representar los documentos en el proceso de clasificación. El modelo de bolsa de palabras usa la frecuencia de los términos sin importar su orden de aparición como base para la tarea de clasificación, es decir, no lo hace a partir de la morfología y estructura sintáctica de los textos.

4.6. Costos de Proyecto

Para la realización del proyecto se consideraron elementos como el material usado, horas hombre , tiempo de experimentación , entre otros para evaluar su costo, en la tabla 4.5 se describe de forma mas detallada los parámetros que se utilizaron así como el resultado de la evaluación de los mismos.

Tabla 4.5: Costo de proyecto

Elemento	**Tipo de recurso**	**Tipo de unidad**	**Unidades**	**Precio por unidad**	**Costo**
Personal	Sueldo de Data Scientist Jr.	Jornada semanal	5	$ 1,000.00	$ 5,000.00
Personal	Beca realización corpus de noticias	Jornada semanal	5	$ 333.33	$ 1,666.66
Computadora	Laptop Hp-15	Pieza	1	$ 15,779.00	$ 15,779.00
Computadora	Laptop Acer	Pieza	1	$ 15,998.00	$ 15,998.00
Computadora	Laptop Dell Inspiron 14	Pieza	1	$ 15,899.00	$ 15,899.00
Total					**$ 54,342.66**

Capítulo 5

Resultados

En este capítulo se presentan los resultados obtenidos en la etapa de experimentación, en ellos se mostrará el balance del corpus, en cuanto a la cantidad de palabras y las temáticas, así como la evolución particular de cada uno de los métodos utilizados para el procesamiento de lenguaje natural y la comparación de resultados obtenidos en la etapa de entrenamiento con los modelos de aprendizaje automático.

5.1. Intersección del vocabulario lematizado del corpus

En la tabla 5.1 se muestran los porcentajes que indican la intersección del vocabulario aplicando el proceso de lematización, entre todas las categorías utilizando únicamente las noticias falsas. De forma similar la tabla 5.2 contiene los porcentajes de la intersección del vocabulario lematizado, pero aplicado al conjunto de noticias verdaderas.

Dos de las categorías que tienen una intersección alta en el vocabulario son Política-Espectáculos y Política-Sociedad ya que el periodo de recopilación de las noticias coincidió con la época electoral en México, en la cual varias temáticas se vieron relacionadas. En la tabla 5.3 se muestra un ejemplo de noticias para el par Política-Espectáculos.

Tabla 5.1: Intersección de vocabulario lematizado entre categorías de noticias falsas.

Etiqueta	CyT	Deporte	Economía	Educación	Espect.	Política	Salud	Seguridad	Sociedad
CyT		14.51 %	12.39 %	9.81 %	15.45 %	15.57 %	12.96 %	10.64 %	16.31 %
Deporte	14.51 %		10.74 %	9.21 %	16.61 %	15.40 %	11.13 %	11.05 %	15.48 %
Economía	12.39 %	10.74 %		8.91 %	11.64 %	11.47 %	9.67 %	9.66 %	13.42 %
Educación	9.81 %	9.21 %	8.91 %		9.37 %	8.12 %	9.14 %	8.97 %	9.51 %
Espect.	15.45 %	16.61 %	11.64 %	9.37 %		20.13 %	10.26 %	10.93 %	18.84 %
Política	15.57 %	15.40 %	11.47 %	8.12 %	20.13 %		10.01 %	10.22 %	20.05 %
Salud	12.96 %	11.13 %	9.67 %	9.14 %	10.26 %	10.01 %		9.46 %	12.75 %
Seguridad	10.64 %	11.05 %	9.66 %	8.97 %	10.93 %	10.22 %	9.46 %		11.23 %
Sociedad	16.31 %	15.48 %	13.42 %	9.51 %	18.84 %	20.05 %	12.75 %	11.23 %	

CyT: Ciencia y Tecnología.

Tabla 5.2: Intersección de vocabulario lematizado entre categorías de noticias verdaderas.

Etiqueta	CyT	Deporte	Economía	Educación	Espect.	Política	Salud	Seguridad	Sociedad
CyT		15.40 %	16.94 %	9.14 %	15.89 %	17.46 %	16.40 %	10.90 %	18.74 %
Deporte	15.40 %		15.72 %	8.86 %	18.47 %	17.51 %	13.74 %	11.69 %	17.31 %
Economía	16.94 %	15.72 %		9.85 %	13.72 %	15.30 %	13.96 %	11.32 %	16.28 %
Educación	9.14 %	8.86 %	9.85 %		8.42 %	7.34 %	9.65 %	7.86 %	8.62 %
Espect.	15.89 %	18.47 %	13.72 %	8.42 %		19.47 %	13.64 %	12.56 %	19.23 %
Política	17.46 %	17.51 %	15.30 %	7.34 %	19.47 %		13.40 %	11.57 %	23.63 %
Salud	16.40 %	13.74 %	13.96 %	9.65 %	13.64 %	13.40 %		11.27 %	15.91 %
Seguridad	10.90 %	11.69 %	11.32 %	7.86 %	12.56 %	11.57 %	11.27 %		12.13 %
Sociedad	18.74 %	17.31 %	16.28 %	8.62 %	19.23 %	23.63 %	15.91 %	12.13 %	

CyT: Ciencia y Tecnología.

En las noticias anteriores se involucra a un personaje político con alguien perteneciente al espectáculo.

Por otro lado, categorías como política-educación y espectáculos-CyT tienen una intersección alta en el vocabulario por la diferencia en la cantidad de noticias en cada categoría del corpus, como se muestra en la tabla 4.3.

Tabla 5.3: Ejemplo de n-gramas

Falsa:	Propone AMLO a Belinda para la Secretaría de Cultura" ¡YA LE TIRÓ SU HUESO!
Verdadera:	Cierra AMLO campaña en el Azteca con Belinda, Susana Harp...

Con los ejemplos previos y en comparación con la tabla 5.4 que muestra los porcentajes de la intersección del vocabulario lematizado, tomando en cuenta las noticias falsas y verdaderas de cada categoría. Se puede observar que al unir las dos clases (falsas y verdaderas) hubo un ligero incremento, el cual se debe a que en la elaboración del corpus se buscó una similitud a nivel vocabulario y temático, sin embargo, no se encontraron noticias que fueran completamente opuestas y a su vez hablaran del mismo tema.

Tabla 5.4: Intersección de vocabulario lematizado entre categorías de noticias falsas y verdaderas.

Etiqueta	CyT	Deporte	Economía	Educación	Espect.	Política	Salud	Seguridad	Sociedad
CyT		18.62 %	18.78 %	12.64 %	19.66 %	19.83 %	18.46 %	14.27 %	21.52 %
Deporte	18.62 %		17.35 %	12.03 %	21.86 %	20.19 %	16.01 %	14.88 %	20.12 %
Economía	18.78 %	17.35 %		12.71 %	16.20 %	16.88 %	15.54 %	14.57 %	18.49 %
Educación	12.64 %	12.03 %	12.71 %		11.45 %	9.85 %	13.07 %	11.07 %	11.69 %
Espect.	19.66 %	21.86 %	16.20 %	11.45 %		23.46 %	15.48 %	15.18 %	22.87 %
Política	19.83 %	20.19 %	16.88 %	9.85 %	23.46 %		14.65 %	13.66 %	25.81 %
Salud	18.46 %	16.01 %	15.54 %	13.07 %	15.48 %	14.65 %		14.03 %	17.75 %
Seguridad	14.27 %	14.88 %	14.57 %	11.07 %	15.18 %	13.66 %	14.03 %		14.58 %
Sociedad	21.52 %	20.12 %	18.49 %	11.69 %	22.87 %	25.81 %	17.75 %	14.58 %	

CyT: Ciencia y Tecnología.

5.2. Linea base

La tabla 5.5 contiene en la segunda columna los resultados del clasificador al entrenarlo con la frecuencia de las palabras eliminando las palabras auxiliares. La tercera columna contiene los resultados del clasificador al entrenarlo únicamente con la frecuencia de las palabras auxiliares. La cuarta columna muestra los resultados del clasificador al entrenarlo solamente con las etiquetas gramaticales de las palabras.

Tabla 5.5: Resultados de linea base

	Sin palabras auxiliares	Palabras auxiliares	Etiquetas de clases gramaticales (POS)
Regresión Logística	72,20 %	67,45 %	67,11 %
SVM	71,52 %	**68,81 %**	**68,13 %**
Bosques Aleatorios	**76,27 %**	68,13 %	63,72 %
Boosting	72,54 %	66,44 %	61,01 %

Los resultados entrenando solo con palabras auxiliares muestran que la clasificación no proporciona resultados óptimos, sin embargo, se obtienen valores significativos que muestran que las palabras auxiliares proporcionan información de ayuda en el proceso de clasificación. Mientras que las etiquetas POS son características que proporcionan menor información del contenido de los textos en comparación a los anteriores.

5.3. Identificación de características relevantes

En esta sección se muestran las primeras diez palabras más relevantes para la clasificación de los modelos.

La tabla 5.6 muestran los resultados del clasificador al entrenarlo únicamente con palabras auxiliares, para este caso se utilizaron n-gramas de palabras.

Tabla 5.6: Resultados de n-gramas utilizando unicamente palabras auxiliares.

N-gramas utilizando unicamente palabras auxiliares			
	3-gramas	**4-gramas**	**5-gramas**
No.	Palabra	Palabra	Palabra
1	en el de	de de lo de	lo de de lo de
2	en el en	en el de lo	de lo el en el
3	lo de de	lo de lo de	de lo el de lo
4	de de lo	el en el lo	lo en lo de lo
5	de los el	lo de lo en	durante el los de desde
6	lo de lo	de en lo en	lo de lo por el
7	los lo lo	el lo por lo	el de lo con el
8	en en lo	en de los de	en lo de en lo
9	el de lo	el el de en	de el de lo lo
10	en el el	de lo el en	tras de de de lo
Regresión Logística	62,37 %	61,35 %	59,66 %
SVM	62,03 %	58,98 %	62,03 %
Bosques Aleatorios	**67,79 %**	**65,76 %**	**67,11 %**
Boosting	60,67 %	56,61 %	57,62 %

En la tabla 5.5 se muestra la relevancia de las palabras auxiliares, por lo que se probó con n-gramas y palabras auxiliares, dando mejores resultados utilizando 3-gramas (tabla 5.6). A

partir del uso de árboles de decisión se logra observar que la palabra auxiliar más relevante ha sido *de*.

En la tabla 5.7 se muestran los resultados de la implementación de n-gramas de caracteres eliminando las palabras auxiliares, así como las palabras que tienen un mayor peso para la clasificación.

Tabla 5.7: Resultados de n-gramas de caracteres sin palabras auxiliares.

N-gramas de caracteres sin palabras auxiliares			
	3-gramas	**4-gramas**	**5-gramas**
No.	Palabra	Palabra	Palabra
1	num	numb	numbe
2	or_	ción	ción_
3	al_	evel	evela
4	min	alic	alici
5	nco	aliz	aliza
6	sir	salv	salva
7	io_	_rec	r_ele
8	_pú	aje_	ngo_p
9	anc	ia_e	ante_
10	ién	pega	lar_h
Resultados del clasificador			
Regresión Logística	67,11 %	69,15 %	70,16 %
SVM	68,81 %	**72,20 %**	**73,55 %**
Bosques Aleatorios	69,83 %	71,18 %	**73,55 %**
Boosting	**70,16 %**	71,52 %	73,22 %

En la tabla 5.7 se observa que los n-gramas de caracteres sin palabras auxiliares, muestran que los sufijos y prefijos son las principales características en las palabras que favorecen el proceso de clasificación.

En la tabla 5.8 se muestran los resultados de la implementación de n-gramas de caracteres utilizando las palabras auxiliares, así como las palabras que tienen un mayor peso para la clasificación. Se observa que los n-gramas de caracteres con palabras auxiliares tienen mejores resultados, además los sufijos y prefijos siguen siendo características relevantes, incluso algunos n-gramas son palabras auxiliares. Los resultados coinciden con la teoría de

que el uso de n-gramas de caracteres ayudan a identificar patrones de autoría relacionados con las palabras auxiliares, prefijos y sufijos.

Tabla 5.8: Resultados de n-gramas de caracteres con palabras auxiliares.

N-gramas de caracteres con palabras auxiliares			
	3-gramas	**4-gramas**	**5-gramas**
No.	Palabra	Palabra	Palabra
1	num	numb	_numb
2	_de	_de_	ción_
3	_en	en_e	_en_e
4	_ya	oder	avés_
5	ord	ta_q	_ya_q
6	ame	tras	r_hab
7	en_	evel	revel
8	cir	iden	aje_d
9	io_	cuat	l_que
10	pet	y_sa	_fin_
Resultados del clasificador			
Regresión Logística	71,18 %	76,61 %	75,93 %
SVM	72,54 %	75,59 %	75,59 %
Bosques Aleatorios	74,57 %	75,25 %	**76,27 %**
Boosting	**75,25 %**	**77,28 %**	**76,27 %**

La tabla 5.9 contiene los resultados del clasificador al entrenarlo con las etiquetas de clases gramaticales de las palabras y eliminando las palabras auxiliares. Se observa que usar únicamente las etiquetas de clases gramaticales proporciona información importante para el proceso de clasificación.

La tabla 5.10 muestra los resultados de los modelos al entrenarlo con la concatenación de las matrices correspondientes a las palabras y a sus etiquetas de clases gramaticales, así como eliminando las palabras auxiliares. Este experimento muestra mejores resultados que utilizar únicamente las características individuales. También se puede observar en esta sección que el uso de números y palabras auxiliares en las noticias son patrones principales para el proceso de clasificación.

Tabla 5.9: Resultados de n-gramas de etiquetas de clases gramaticales sin palabras auxiliares

Resultados de n-gramas de etiquetas de clases gramaticales sin palabras auxiliares			
	3-gramas	**4-gramas**	**5-gramas**
No.	Palabra	Palabra	Palabra
1	verb-noun-adj	verb-noun-noun-adj	verb-noun-verb-noun-adj
2	verb-noun-noun	verb-noun-adj-adj	adj-aux-verb-verb-noun
3	adj-aux-verb	verb-adp-verb-verb	noun-adj-aux-verb-verb
4	verb-adj-noun	noun-verb-noun-adj	verb-adj-noun-verb-noun
5	adj-adv-aux	verb-noun-verb-adj	noun-noun-verb-noun-adj
6	adv_noun_adj	noun-adj-verb-noun	verb-noun-verb-noun-verb
7	noun-noun-verb	noun-noun-verb-verb	aux-adv-noun-aux-verb
8	adj-verb-adj	noun-noun-adj-adj	noun-noun-verb-verb-noun
9	adp_noun_noun	noun-verb-noun-noun	aux-verb-verb-verb-noun
10	adj-noun-adj	adv-noun-verb-verb	adj-adj-noun-adj-verb
Resultados del clasificador			
Regresión Logística	63,38 %	61,01 %	56,27 %
SVM	60,67 %	58,30 %	57,62 %
Bosques Aleatorios	**67,11 %**	**66,10 %**	**63,05 %**
Boosting	62,37 %	59,32 %	61,01 %

5.4. Resultados de clasificadores a partir de la concatenación de características

La tabla 5.11 contiene en la segunda columna los resultados del clasificador al entrenarlo con la suma de n-gramas de tamaño 3, 4 y 5 usando solo las etiquetas gramaticales de las palabras (POS). La tercera columna muestra los resultados del clasificador al entrenarlo con la suma de n-gramas de caracteres de tamaño 3, 4 y 5. La cuarta columna contiene los resultados del clasificador al entrenarlo con la suma de n-gramas de tamaño 3, 4 y 5 usando n-gramas de palabras. La quinta columna contiene los resultados del clasificador al entrenarlo con la suma de los tres experimentos anteriores, es decir, al concatenar las matrices de n-gramas de palabras, de caracteres y categorías de clases gramaticales; cada una de estas matrices contiene la suma de n-gramas de tamaño 3,4 y 5 de sus respectivas características.

Los resultados concatenando las características no difieren mucho debido a la extensión de los n-gramas, por lo que se propone la inclusión de bigramas en los siguientes experimentos.

Tabla 5.10: Resultados de etiquetas de clases gramaticales + palabras.

Etiquetas de clases gramaticales + Palabras	
No.	Palabra
1	number
2	detallar
3	revelar
4	domingo
5	dama
6	vital
7	noun
8	mensaje
9	acabar
10	inminente
Resultados del clasificador	
Regresión Logística	73,55 %
SVM	70,84 %
Bosques Aleatorios	**76,94 %**
Boosting	72,20 %

5.5. Entrenamientos y pruebas con una sola categoría

En las siguientes tablas se muestran los resultados de entrenar con una categoría y con esa misma probar la clasificación. Esto se lleva acabo con cada categoría y diferentes tipos de procesamientos.

Tabla 5.11: Resultados de suma de n-gramas (3 + 4 + 5).

	Suma de n-gramas (3 + 4 + 5)			
	Etiquetas de clases gramaticales (POS)	de caracteres	de palabras	de caracteres + POS + palabras
Regresión Logística	60,00 %	66,10 %	53,89 %	64,06 %
SVM	60,33 %	**73,22 %**	51,18 %	**73,22 %**
Bosques Aleatorios	**64,74 %**	71,18 %	50,84 %	70,84 %
Boosting	61,01 %	**73,22 %**	**58,98 %**	72,88 %

La tabla 5.12 contiene los resultados de entrenar con y sin palabras auxiliares. Los mejores resultados siguen siendo con palabras auxiliares, aunque SVM tiene resultados altos sin palabras auxiliares.

Tabla 5.12: Resultados con y sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	Sin palabras auxiliares				**Con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	42.85 %	57.14 %	42.85 %	71.42 %	57.14 %	57.14 %	42.85 %	42.85 %
Deporte	60.52 %	60.52 %	60.52 %	52.63 %	71.05 %	71.05 %	71.05 %	55.26 %
CyT	59.25 %	59.25 %	66.66 %	55.55 %	44.44 %	55.55 %	66.66 %	59.25 %
Seguridad	53.84 %	**92.30 %**	69.23 %	69.23 %	53.84 %	84.61 %	76.92 %	**100.00 %**
Salud	78.57 %	71.42 %	64.28 %	85.71 %	71.42 %	78.57 %	78.57 %	85.71 %
Economía	76.92 %	76.92 %	61.53 %	76.92 %	61.53 %	61.53 %	69.23 %	61.53 %
Sociedad	68.29 %	78.04 %	75.60 %	78.04 %	73.17 %	75.17 %	78.04 %	78.04 %
Política	74.22 %	77.31 %	81.44 %	70.10 %	78.35 %	79.38 %	84.53 %	80.41 %
Espect.	68.88 %	62.22 %	60.00 %	73.33 %	71.11 %	64.44 %	66.66 %	71.11 %

En la tabla 5.13 se muestran los resultados de entrenar y probar sin palabras auxiliares, y utilizando 2-gramas, 3-grmas, 4-gramas y 5-gramas. Del mismo modo la tabla 5.14 contiene los resultados n-gramas, pero en este caso dejando las palabras auxiliares.

Los n-gramas de palabras con palabras auxiliares de 2-gramas y 3-gramas obtuvieron mejores resultados que los n-gramas de 4 y 5.

En la tabla 5.15 se muestran los resultados de sumar los n-gramas (2+3+4+5) de palabras con y sin palabras auxiliares.

Tabla 5.13: Resultados de n-gramas sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	N-gramas de palabras sin palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	**71.42 %**	42.85 %	42.58 %	28.57 %	57.14 %	42.85 %	42.85 %	57.14 %
Deporte	**63.15 %**	55.26 %	57.89 %	57.89 %	47.36 %	52.63 %	55.26 %	55.26 %
CyT	**55.55 %**	51.85 %	**55.55 %**	**55.55 %**	51.85 %	48.14 %	51.85 %	**55.55 %**
Seguridad	53.84 %	53.84 %	53.84 %	**69.23 %**	53.84 %	53.84 %	53.84 %	53.84 %
Salud	**71.42 %**	57.14 %	57.14 %	64.28 %	64.28 %	50.00 %	50.00 %	**71.42 %**
Economía	**76.92 %**	**76.92 %**	46.15 %	69.23 %	**76.92 %**	61.53 %	46.15 %	**76.92 %**
Sociedad	**70.73 %**	60.97 %	56.09 %	65.85 %	56.09 %	53.65 %	53.65 %	53.65 %
Política	**64.94 %**	59.79 %	61.85 %	56.70 %	49.48 %	49.48 %	56.70 %	57.73 %
Espect.	**57.77 %**	55.55 %	51.11 %	**57.77 %**	48.88 %	51.11 %	51.11 %	53.33 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	42.85 %	42.85 %	57.14 %	42.85 %	42.85 %	42.85 %	57.14 %
Deporte	52.63 %	47.36 %	55.26 %	52.63 %	44.73 %	44.73 %	55.26 %	55.26 %
CyT	48.14 %	48.14 %	51.85 %	51.85 %	48.14 %	48.14 %	51.85 %	51.85 %
Seguridad	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %
Salud	50.00 %	50.00 %	50.00 %	50.00 %	50.00 %	50.00 %	50.00 %	50.00 %
Economía	69.23 %	53.84 %	46.15 %	38.46 %	61.53 %	53.84 %	46.15 %	46.15 %
Sociedad	53.65 %	53.65 %	53.65 %	51.21 %	53.65 %	53.65 %	53.65 %	53.65 %
Política	49.48 %	48.45 %	55.67 %	57.73 %	46.39 %	45.36 %	55.67 %	51.54 %
Espect.	51.11 %	51.11 %	51.11 %	51.11 %	51.11 %	51.11 %	51.11 %	51.11 %

Tabla 5.14: Resultados de n-gramas con palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	N-gramas de palabras con palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	**57.14 %**	**57.14 %**	42.85 %	**57.14 %**	**57.14 %**	42.85 %	42.85 %	**57.14 %**
Deporte	**76.31 %**	73.68 %	65.78 %	63.15 %	60.52 %	57.89 %	55.26 %	47.36 %
CyT	55.55 %	59.25 %	**66.66 %**	51.85 %	62.96 %	55.55 %	59.25 %	55.55 %
Seguridad	61.53 %	**92.30 %**	53.84 %	**92.30 %**	46.15 %	53.84 %	53.84 %	69.23 %
Salud	78.57 %	71.42 %	57.14 %	**85.71 %**	78.57 %	50.00 %	50.00 %	57.14 %
Economía	61.53 %	53.84 %	46.15 %	61.53 %	61.53 %	61.53 %	46.15 %	61.53 %
Sociedad	73.17 %	**75.60 %**	70.73 %	68.29 %	65.85 %	56.09 %	56.09 %	**75.60 %**
Política	75.25 %	**81.44 %**	78.35 %	68.04 %	76.28 %	64.94 %	71.13 %	68.04 %
Espect.	57.77 %	60.00 %	**62.22 %**	**62.22 %**	55.55 %	55.55 %	53.33 %	46.66 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	**57.14 %**	42.85 %	42.85 %	**57.14 %**	**57.14 %**	42.85 %	42.85 %	42.85 %
Deporte	50.00 %	52.63 %	55.26 %	57.89 %	55.26 %	57.89 %	55.26 %	57.89 %
CyT	59.25 %	48.14 %	51.85 %	51.85 %	48.14 %	48.14 %	51.85 %	44.44 %
Seguridad	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %	53.84 %
Salud	71.42 %	50.00 %	50.00 %	50.00 %	64.28 %	50.00 %	57.14 %	50.00 %
Economía	**76.92 %**	61.53 %	46.15 %	53.84 %	61.53 %	53.84 %	46.15 %	38.46 %
Sociedad	51.21 %	53.65 %	53.65 %	63.41 %	53.65 %	53.65 %	53.65 %	60.97 %
Política	59.79 %	51.54 %	63.91 %	68.04 %	52.57 %	45.36 %	56.70 %	65.97 %
Espect.	48.88 %	51.11 %	51.11 %	55.55 %	51.11 %	51.11 %	51.11 %	55.55 %

Tabla 5.15: Resultados de suma de n-gramas de palabras con y sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	Suma de N-gramas (2+3+4+5) de palabras sin palabras auxiliares				**Suma de N-gramas (2+3+4+5) de palabras con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	71.42 %	42.85 %	42.85 %	28.57 %	57.14 %	42.85 %	42.85 %	42.85 %
Deporte	52.63 %	55.26 %	55.26 %	76.31 %	68.42 %	71.05 %	60.52 %	78.94 %
CyT	55.55 %	48.14 %	48.14 %	48.14 %	59.25 %	62.96 %	55.55 %	55.55 %
Seguridad	53.84 %	53.84 %	53.84 %	61.53 %	46.15 %	76.92 %	53.84 %	**92.30 %**
Salud	**78.57 %**	50.00 %	64.28 %	57.14 %	85.71 %	64.28 %	57.14 %	64.28 %
Economía	76.92 %	69.23 %	46.15 %	69.23 %	53.84 %	53.84 %	46.15 %	61.53 %
Sociedad	73.17 %	53.65 %	53.65 %	68.25 %	73.17 %	70.73 %	63.41 %	75.60 %
Política	63.91 %	51.54 %	65.97 %	62.88 %	74.22 %	79.38 %	79.38 %	79.38 %
Espect.	55.55 %	53.33 %	51.11 %	60.00 %	57.77 %	57.77 %	55.55 %	57.77 %

En la tabla 5.15 se puede ver que el mejor resultado concatenando n-gramas fue con palabras auxiliares, principalmente las categorías de política y sociedad dan mejores resultados en los cuatro modelos, estas categorías son de las que tienen mayor cantidad de noticias.

Tabla 5.16: Resultados de n-gramas de caracteres sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	N-gramas de caracteres sin palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	28.57 %	42.85 %	42.85 %	**57.14 %**	42.85 %	42.85 %	42.85 %	42.85 %
Deporte	60.52 %	57.89 %	52.63 %	55.26 %	63.15 %	60.52 %	55.26 %	63.15 %
CyT	51.85 %	55.55 %	**74.07 %**	66.66 %	51.85 %	59.25 %	62.96 %	59.25 %
Seguridad	30.76 %	69.23 %	**92.30 %**	76.92 %	38.46 %	76.92 %	84.61 %	61.53 %
Salud	71.42 %	**85.71 %**	**85.71 %**	78.57 %	71.42 %	78.57 %	**85.71 %**	78.57 %
Economía	69.23 %	69.23 %	61.53 %	61.53 %	**76.92 %**	**76.92 %**	69.23 %	53.84 %
Sociedad	70.73 %	68.29 %	75.60 %	78.04 %	73.17 %	75.60 %	75.60 %	63.41 %
Política	64.94 %	67.01 %	72.16 %	72.16 %	72.16 %	76.28 %	73.19 %	76.28 %
Espect.	44.44 %	62.22 %	62.22 %	60.00 %	62.22 %	66.66 %	60.00 %	64.44 %
	4				5			
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	**57.14 %**	42.85 %	42.85 %	14.28 %	**57.14 %**	42.85 %	42.85 %	42.85 %
Deporte	63.15 %	55.26 %	60.52 %	57.89 %	63.15 %	55.26 %	**68.42 %**	65.78 %
CyT	55.55 %	59.25 %	**74.07 %**	51.85 %	48.14 %	66.66 %	62.96 %	59.25 %
Seguridad	38.46 %	84.61 %	76.92 %	61.53 %	46.15 %	84.61 %	69.23 %	76.92 %
Salud	71.42 %	71.42 %	78.57 %	71.42 %	71.42 %	71.42 %	71.42 %	78.57 %
Economía	**76.92 %**	69.23 %	69.23 %	53.84 %	69.23 %	61.53 %	69.23 %	46.15 %
Sociedad	68.29 %	78.04 %	80.48 %	**82.92 %**	68.29 %	78.04 %	73.17 %	70.73 %
Política	72.16 %	77.31 %	76.28 %	76.28 %	71.13 %	77.31 %	**79.38 %**	74.22 %
Espect.	64.44 %	62.22 %	60.00 %	62.22 %	64.44 %	66.66 %	62.22 %	**71.11 %**

Comparando los resultados entre n-gramas de caracteres sin y con palabras auxiliares, tablas 5.16 y 5.17, se puede observar que algunos valores se mantienen constantes, sin embargo, categorías como sociedad, política y espectáculos, mejoraron los resultados utilizando n-gramas de caracteres con palabras auxiliares.

La concatenación de n-gramas de caracteres tuvo menores resultados que estos mismos de manera individual como se muestra en la tabla 5.18

Tabla 5.17: Resultados de n-gramas de caracteres con palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	N-gramas de caracteres con palabrabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	28.57 %	28.57 %	42.85 %	**57.14 %**	42.85 %	42.85 %	42.85 %	28.57 %
Deporte	65.78 %	65.78 %	68.42 %	63.15 %	63.15 %	60.52 %	71.05 %	60.52 %
CyT	59.25 %	62.96 %	**74.07 %**	66.66 %	62.96 %	62.96 %	**74.07 %**	70.37 %
Seguridad	53.84 %	69.23 %	**92.30 %**	84.61 %	38.46 %	84.61 %	**92.30 %**	76.92 %
Salud	78.57 %	78.57 %	85.71 %	85.71 %	85.71 %	78.57 %	85.71 %	78.57 %
Economía	61.53 %	**69.23 %**	46.15 %	38.46 %	61.53 %	**69.23 %**	61.53 %	61.53 %
Sociedad	78.04 %	70.73 %	78.04 %	75.60 %	**82.92 %**	80.48 %	73.17 %	70.73 %
Política	72.16 %	71.13 %	**81.44 %**	69.07 %	77.31 %	74.22 %	77.31 %	77.31 %
Espect.	62.22 %	57.77 %	57.77 %	57.77 %	**73.33 %**	71.11 %	62.22 %	60.00 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	**57.14 %**	**57.14 %**	42.85 %	28.57 %	42.85 %	42.85 %	42.85 %	42.85 %
Deporte	63.15 %	63.15 %	73.68 %	65.78 %	68.42 %	68.42 %	**78.94 %**	57.89 %
CyT	62.96 %	62.96 %	**74.07 %**	66.66 %	66.66 %	62.96 %	70.37 %	62.96 %
Seguridad	46.15 %	**92.30 %**	84.61 %	69.23 %	46.15 %	**92.30 %**	84.61 %	**92.30 %**
Salud	78.57 %	78.57 %	85.71 %	**92.85 %**	78.57 %	78.57 %	85.71 %	71.42 %
Economía	**69.23 %**	61.53 %	**69.23 %**	61.53 %	**69.23 %**	61.53 %	**69.23 %**	53.84 %
Sociedad	78.04 %	80.48 %	80.48 %	80.48 %	70.73 %	80.48 %	75.60 %	80.48 %
Política	80.41 %	80.41 %	77.31 %	78.35 %	80.41 %	78.35 %	78.35 %	72.16 %
Espect.	**73.33 %**	68.88 %	64.44 %	71.11 %	71.11 %	64.44 %	60.00 %	71.11 %

Tabla 5.18: Resultados de suma de n-gramas de caracteres con y sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	Suma de n-gramas (2+3+4+5) de caracteres sin palabras auxiliares				**Suma de n-gramas (2+3+4+5) de caracteres con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	42.85 %	42.85 %	42.85 %	28.57 %	42.85 %	42.85 %	42.85 %	28.57 %
Deporte	65.78 %	63.15 %	60.52 %	68.42 %	65.78 %	68.42 %	71.05 %	60.52 %
CyT	48.14 %	59.25 %	70.37 %	55.55 %	66.66 %	66.66 %	70.37 %	66.66 %
Seguridad	46.15 %	84.61 %	84.61 %	76.92 %	38.46 %	84.61 %	84.61 %	84.61 %
Salud	71.42 %	78.57 %	**85.71 %**	**85.71 %**	78.57 %	78.57 %	**85.71 %**	78.57 %
Economía	76.92 %	69.23 %	69.23 %	53.84 %	69.23 %	69.23 %	61.53 %	46.15 %
Sociedad	70.73 %	75.60 %	75.60 %	70.73 %	80.48 %	78.04 %	80.48 %	80.48 %
Política	68.04 %	75.25 %	77.31 %	75.25 %	72.16 %	76.28 %	75.25 %	78.35 %
Espect.	57.77 %	55.55 %	64.44 %	60.00 %	73.33 %	66.66 %	66.66 %	73.33 %

En la tabla 5.19 se muestran los resultados de n-gramas de POS sin palabras auxiliares donde los mejores resultados se presentaron en 2-gramas y 3-gramas.

Tabla 5.19: Resultados de n-gramas de POS sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	N-gramas de POS sin palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	28.57 %	28.57 %	28.57 %	42.85 %	42.85 %	42.85 %	28.57 %	28.57 %
Deporte	55.26 %	52.63 %	60.52 %	63.15 %	52.63 %	55.26 %	57.89 %	**68.42 %**
CyT	44.44 %	55.55 %	74.07 %	**81.48 %**	62.96 %	62.96 %	70.37 %	77.77 %
Seguridad	61.53 %	69.23 %	**92.30 %**	84.61 %	61.53 %	84.61 %	84.61 %	84.61 %
Salud	57.14 %	85.71 %	85.71 %	78.57 %	57.14 %	85.71 %	85.71 %	85.71 %
Economía	53.84 %	53.84 %	**61.53 %**	46.15 %	30.76 %	38.46 %	53.84 %	46.15 %
Sociedad	65.85 %	63.41 %	**75.60 %**	63.41 %	65.85 %	**75.60 %**	70.73 %	68.29 %
Política	68.04 %	64.94 %	68.04 %	61.85 %	60.82 %	62.88 %	68.04 %	52.57 %
Espect.	53.33 %	44.44 %	**62.22 %**	51.11 %	57.77 %	48.88 %	51.11 %	57.77 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	28.57 %	42.85 %	42.85 %	42.85 %	42.85 %	42.85 %	42.85 %	14.28 %
Deporte	42.10 %	50.00 %	57.89 %	55.26 %	52.63 %	57.89 %	65.78 %	47.36 %
CyT	59.25 %	55.55 %	66.66 %	59.25 %	62.96 %	74.07 %	66.66 %	70.37 %
Seguridad	53.84 %	84.61 %	76.92 %	69.23 %	46.15 %	76.92 %	69.23 %	61.53 %
Salud	57.14 %	85.71 %	85.71 %	57.14 %	64.28 %	78.57 %	71.42 %	85.71 %
Economía	30.76 %	53.84 %	61.53 %	46.15 %	38.46 %	61.53 %	38.46 %	38.46 %
Sociedad	60.97 %	60.97 %	70.73 %	60.97 %	58.53 %	70.73 %	68.29 %	56.09 %
Política	56.70 %	59.79 %	**70.10 %**	65.97 %	56.70 %	57.73 %	64.94 %	62.88 %
Espect.	44.44 %	42.22 %	42.22 %	46.66 %	51.11 %	51.11 %	48.88 %	46.66 %

Comparando los resultados de las tablas 5.19 y 5.20, los mejores resultados han sido los n-gramas de POS con palabras auxiliares, aunque de la misma forma que la tabla 5.19 los resultados son más altos en 2-gramas y 3-gramas.

Tabla 5.20: Resultados de n-gramas de POS con palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	N-gramas de POS con palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	57.14 %	57.14 %	42.85 %	57.14 %	57.14 %	42.85 %	57.14 %
Deporte	71.05 %	65.78 %	73.68 %	71.05 %	68.42 %	68.42 %	**78.94 %**	65.78 %
CyT	44.44 %	44.44 %	74.07 %	**77.77 %**	66.66 %	62.96 %	66.66 %	62.96 %
Seguridad	84.61 %	84.61 %	76.92 %	69.23 %	53.84 %	76.92 %	76.92 %	69.23 %
Salud	64.28 %	78.57 %	85.71 %	57.14 %	57.14 %	78.57 %	85.71 %	**92.85 %**
Economía	61.53 %	69.23 %	69.23 %	53.84 %	69.23 %	69.23 %	69.23 %	61.53 %
Sociedad	65.85 %	68.29 %	75.60 %	70.73 %	68.29 %	68.29 %	70.73 %	80.48 %
Política	70.10 %	68.04 %	**83.50 %**	70.10 %	70.10 %	72.16 %	76.28 %	70.10 %
Espect.	53.33 %	60.00 %	62.22 %	55.55 %	**73.33 %**	66.66 %	53.33 %	62.22 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	57.14 %	42.85 %	71.42 %	28.57 %	57.14 %	42.85 %	14.28 %
Deporte	57.89 %	63.15 %	57.89 %	65.78 %	57.89 %	55.26 %	60.52 %	57.89 %
CyT	62.96 %	59.25 %	66.66 %	55.55 %	62.96 %	66.66 %	62.96 %	70.37 %
Seguridad	53.84 %	**92.30 %**	69.23 %	46.15 %	53.84 %	76.92 %	69.23 %	69.23 %
Salud	57.14 %	85.71 %	78.57 %	85.71 %	64.28 %	71.42 %	71.42 %	64.28 %
Economía	69.23 %	61.53 %	69.23 %	38.46 %	61.53 %	53.84 %	69.23 %	46.15 %
Sociedad	65.85 %	68.29 %	78.04 %	73.17 %	68.29 %	75.60 %	75.60 %	73.17 %
Política	68.04 %	72.16 %	74.22 %	68.04 %	65.97 %	72.16 %	76.28 %	75.25 %
Espect.	62.22 %	60.00 %	53.33 %	64.44 %	55.55 %	55.55 %	60.00 %	55.5 %

Tabla 5.21: Resultados de suma de n-gramas de POS con y sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	Suma de n-gramas (2+3+4+5) POS sin palabras auxiliares				**Suma de n-gramas (2+3+4+5) con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	28.57 %	28.57 %	42.85 %	42.85 %	57.14 %	57.14 %	42.85 %	71.42 %
Deporte	60.52 %	52.63 %	57.89 %	60.52 %	68.42 %	68.42 %	73.68 %	71.05 %
CyT	66.66 %	74.07 %	70.37 %	74.04 %	48.14 %	62.96 %	74.07 %	66.66 %
Seguridad	46.15 %	**92.30 %**	84.61 %	84.61 %	69.23 %	**92.30 %**	84.61 %	76.92 %
Salud	50.00 %	78.57 %	85.71 %	57.14 %	64.28 %	78.57 %	85.71 %	64.28 %
Economía	30.76 %	53.84 %	53.84 %	30.76 %	61.53 %	69.23 %	69.23 %	69.23 %
Sociedad	58.53 %	70.73 %	68.29 %	65.85 %	70.73 %	68.29 %	73.17 %	73.10 %
Política	56.70 %	58.76 %	67.01 %	60.82 %	71.13 %	74.22 %	76.28 %	72.16 %
Espect.	53.33 %	46.66 %	46.66 %	55.55 %	73.33 %	66.66 %	57.77 %	62.22 %

La concatenación de POS con y sin palabras auxiliares tabla 5.21 tiene mejores resultados que de manera individual, específicamente los resultados más altos fueron los n-gramas de POS con palabras auxiliares.

La tabla 5.22 muestra mejores resultados en la concatenación de n-gramas (2+3+4+5) de POS, palabras, caracteres y palabras auxiliares.

Tabla 5.22: Resultados de suma de n-gramas de POS, caracteres, con y sin palabras auxiliares, utilizando una misma categoría de entrenamiento y prueba.

	Suma de n-gramas (2+3+4+5) POS + caracteres + palabras				**Suma de n-gramas (2+3+4+5) POS + caracteres + palabras + palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	42.85 %	42.85 %	28.57 %	57.14 %	57.14 %	42.85 %	28.57 %
Deporte	60.52 %	68.42 %	63.15 %	76.31 %	65.78 %	68.42 %	73.68 %	60.52 %
CyT	48.14 %	59.25 %	66.66 %	74.07 %	59.25 %	66.66 %	70.37 %	70.37 %
Seguridad	46.15 %	**92.30 %**	76.92 %	84.61 %	46.15 %	**84.61 %**	**84.61 %**	**84.61 %**
Salud	71.42 %	78.57 %	85.71 %	85.71 %	78.57 %	78.57 %	85.71 %	78.57 %
Economía	69.23 %	69.23 %	69.23 %	46.15 %	69.23 %	69.23 %	69.23 %	46.15 %
Sociedad	70.73 %	75.60 %	73.17 %	68.29 %	78.04 %	82.92 %	78.04 %	78.04 %
Política	71.13 %	75.25 %	75.25 %	81.44 %	74.22 %	76.28 %	76.28 %	78.35 %
Espect.	57.77 %	60.00 %	62.22 %	60.00 %	75.55 %	68.88 %	64.44 %	71.11 %

5.6. Entrenamientos sin utilizar la categoría de prueba

Esta sección contiene los resultados de entrenar seleccionando una categoría entre las demás consideradas en el corpus (Política, Sociedad, etc.). La categoría seleccionada se retira del conjunto de noticias de entrenamiento, sin embargo, esa categoría es la única que se considera para las pruebas. Por ejemplo, de las categorías:

Educación, Deporte, Seguridad, Salud

Si la categoría seleccionada es *Deporte*, en el entrenamiento se considerará todas las noticias de las demás categorías,

Educación, Seguridad, Salud

y para pruebas se consideran únicamente las noticias de *Deporte* .

Tabla 5.23: Resultados de entrenamientos con y sin palabras auxiliares, sin utilizar la categoría de prueba.

	Sin Palabras auxiliares				**Con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	71.42 %	71.42 %	**100.00 %**	71.42 %	71.42 %	71.42 %	**100.00 %**	85.71 %
Deporte	57.89 %	60.52 %	57.89 %	60.52 %	65.78 %	68.42 %	57.89 %	60.52 %
CyT	70.37 %	74.07 %	74.07 %	70.37 %	77.77 %	77.77 %	74.07 %	74.04 %
Seguridad	84.61 %	92.30 %	84.61 %	92.30 %	84.61 %	84.61 %	84.61 %	92.30 %
Salud	85.71 %	85.71 %	85.71 %	92.85 %	85.71 %	85.71 %	85.71 %	85.71 %
Economía	69.23 %	69.23 %	69.23 %	61.53 %	53.84 %	53.84 %	69.23 %	61.53 %
Sociedad	82.92 %	85.36 %	87.80 %	78.04 %	78.04 %	85.36 %	85.36 %	80.48 %
Política	73.19 %	72.16 %	75.25 %	73.19 %	78.35 %	75.25 %	76.28 %	70.10 %
Espect.	55.55 %	64.44 %	64.44 %	64.44 %	73.33 %	64.44 %	66.66 %	66.66 %

Tabla 5.24: Resultados de entrenamientos de n-gramas de palabras eliminando palabras auxiliares, sin utilizar la categoría de prueba.

	N-gramas de palabras sin palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	42.85 %	57.14 %	42.85 %	57.14 %	42.85 %	42.85 %	42.85 %	42.85 %
Deporte	50.00 %	50.00 %	47.36 %	**52.63 %**	47.36 %	47.36 %	47.36 %	47.36 %
CyT	**59.25 %**	44.44 %	51.85 %	55.55 %	51.85 %	48.14 %	48.14 %	48.14 %
Seguridad	**76.92 %**	**76.92 %**	69.23 %	**76.92 %**	61.53 %	53.84 %	61.53 %	69.23 %
Salud	78.57 %	78.57 %	78.57 %	**85.71 %**	64.28 %	64.28 %	64.28 %	78.57 %
Economía	76.92 %	76.92 %	**84.61 %**	69.23 %	61.53 %	61.53 %	61.53 %	69.23 %
Sociedad	**82.92 %**	73.17 %	68.29 %	65.85 %	56.09 %	53.65 %	56.09 %	65.85 %
Política	64.94 %	62.88 %	52.57 %	**67.01 %**	48.45 %	46.39 %	44.32 %	51.54 %
Espect.	55.55 %	53.33 %	**57.77 %**	53.33 %	53.33 %	55.55 %	53.33 %	51.11 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	42.85 %	42.85 %	42.85 %	57.14 %	42.85 %	42.85 %	42.85 %	57.14 %
Deporte	44.73 %	44.73 %	44.73 %	47.36 %	44.73 %	44.73 %	44.73 %	42.10 %
CyT	48.14 %	48.14 %	48.14 %	44.44 %	48.14 %	48.14 %	48.14 %	44.44 %
Seguridad	53.84 %	53.84 %	53.84 %	46.15 %	53.84 %	53.84 %	46.15 %	46.15 %
Salud	50.00 %	50.00 %	50.00 %	71.42 %	50.00 %	50.00 %	50.00 %	57.14 %
Economía	61.53 %	61.53 %	53.84 %	69.23 %	53.84 %	53.84 %	53.84 %	53.84 %
Sociedad	53.65 %	53.65 %	56.09 %	58.53 %	53.65 %	53.65 %	48.78 %	53.65 %
Política	48.45 %	47.42 %	45.36 %	47.42 %	46.39 %	46.39 %	45.36 %	46.39 %
Espect.	53.33 %	53.33 %	53.33 %	53.33 %	53.33 %	53.33 %	53.33 %	55.55 %

Tabla 5.25: Resultados de entrenamientos de n-gramas de palabras con palabras auxiliares, sin utilizar la categoría de prueba.

	N-gramas de palabras con palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	**100.00 %**	85.71 %	85.71 %	**100.00 %**	71.42 %	42.85 %	42.85 %	57.14 %
Deporte	**68.42 %**	63.15 %	55.26 %	57.84 %	52.63 %	50.00 %	47.36 %	47.36 %
CyT	70.37 %	**74.07 %**	70.37 %	66.66 %	59.25 %	59.25 %	51.85 %	66.66 %
Seguridad	**92.30 %**	**92.30 %**	84.61 %	84.61 %	**92.30 %**	**92.30 %**	69.23 %	69.23 %
Salud	71.42 %	78.57 %	**92.58 %**	71.42 %	85.71 %	78.57 %	85.71 %	85.71 %
Economía	61.53 %	61.53 %	61.53 %	69.23 %	76.92 %	**84.61 %**	76.92 %	69.23 %
Sociedad	82.92 %	82.92 %	87.80 %	78.04 %	80.48 %	70.73 %	70.73 %	**87.80 %**
Política	**77.31 %**	71.13 %	69.07 %	73.19 %	64.94 %	62.88 %	55.67 %	67.01 %
Espect.	**73.33 %**	64.44 %	66.66 %	66.66 %	60.00 %	62.22 %	62.22 %	57.77 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	42.85 %	42.85 %	42.85 %	28.57 %	42.85 %	42.85 %	42.85 %	57.14 %
Deporte	47.36 %	47.36 %	47.36 %	42.10 %	44.73 %	44.73 %	44.73 %	42.10 %
CyT	51.85 %	48.14 %	48.14 %	59.25 %	48.14 %	48.14 %	48.14 %	44.44 %
Seguridad	61.53 %	53.84 %	53.84 %	61.53 %	53.84 %	53.84 %	46.15 %	46.15 %
Salud	64.28 %	64.28 %	64.28 %	57.14 %	50.00 %	50.00 %	50.00 %	64.28 %
Economía	69.23 %	61.53 %	61.53 %	**84.61 %**	61.53 %	61.53 %	61.53 %	76.92 %
Sociedad	65.85 %	58.53 %	60.97 %	68.29 %	53.65 %	53.65 %	60.97 %	63.41 %
Política	52.57 %	51.54 %	47.42 %	61.85 %	46.39 %	46.39 %	45.36 %	53.60 %
Espect.	57.77 %	55.55 %	55.55 %	55.55 %	53.33 %	53.33 %	53.33 %	55.55 %

Tabla 5.26: Resultados de entrenamientos concatenando n-gramas (2+3+4+5) de palabras con y sin palabras auxiliares, sin utilizar la categoría de prueba.

	Resultados de suma de n-gramas (2+3+4+5) de palabras sin palabras auxiliares				**Resultados de suma de n-gramas (2+3+4+5) de palabras con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	42.85 %	42.85 %	42.85 %	71.42 %	**100.00 %**	**100.00 %**	57.14 %	85.71 %
Deporte	50.00 %	52.63 %	47.36 %	50.00 %	65.78 %	57.89 %	52.63 %	57.89 %
CyT	55.55 %	44.44 %	48.14 %	51.85 %	62.96 %	66.66 %	70.37 %	66.66 %
Seguridad	76.92 %	61.53 %	69.23 %	**84.61 %**	92.30 %	92.30 %	84.61 %	84.61 %
Salud	78.57 %	78.57 %	71.42 %	71.42 %	71.42 %	78.57 %	92.85 %	85.71 %
Economía	76.92 %	61.53 %	76.92 %	76.92 %	61.53 %	61.53 %	69.23 %	61.53 %
Sociedad	80.48 %	65.85 %	63.41 %	68.29 %	80.48 %	85.36 %	85.36 %	82.92 %
Política	63.91 %	55.67 %	47.42 %	69.07 %	79.38 %	72.16 %	67.01 %	76.28 %
Espect.	60.00 %	57.77 %	55.55 %	53.33 %	68.88 %	62.22 %	64.44 %	68.88 %

Tabla 5.27: Resultados de entrenamientos de n-gramas de caracteres sin palabras auxiliares, sin utilizar la categoría de prueba.

	N-gramas de caracteres sin palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	42.85 %	71.42 %	57.14 %	71.42 %	71.42 %	71.42 %	**85.71 %**
Deporte	36.84 %	44.73 %	63.15 %	50.00 %	47.36 %	52.63 %	63.15 %	**65.70 %**
CyT	62.96 %	62.96 %	66.66 %	62.96 %	74.07 %	66.66 %	70.37 %	66.66 %
Seguridad	69.23 %	76.92 %	84.61 %	92.30 %	69.23 %	84.61 %	92.30 %	84.61 %
Salud	71.42 %	57.14 %	85.71 %	85.71 %	64.28 %	78.57 %	85.71 %	85.71 %
Economía	53.84 %	53.84 %	69.23 %	46.15 %	**76.92 %**	69.23 %	61.53 %	53.84 %
Sociedad	60.97 %	70.73 %	73.17 %	65.85 %	68.29 %	73.17 %	73.17 %	80.48 %
Política	64.94 %	69.07 %	64.94 %	64.94 %	70.10 %	77.31 %	70.10 %	73.19 %
Espect.	68.88 %	55.55 %	51.11 %	51.11 %	71.11 %	53.33 %	55.55 %	46.66 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	71.42 %	71.42 %	57.14 %	**85.71 %**	**85.71 %**	71.42 %	57.14 %	71.42 %
Deporte	60.52 %	57.89 %	60.52 %	**65.78 %**	57.89 %	60.52 %	57.89 %	50.00 %
CyT	**77.77 %**	**77.77 %**	66.66 %	**77.77 %**	70.37 %	74.07 %	70.37 %	66.66 %
Seguridad	76.92 %	92.30 %	92.30 %	76.92 %	84.61 %	92.30 %	92.30 %	84.61 %
Salud	74.42 %	85.71 %	85.71 %	85.71 %	78.57 %	85.71 %	85.71 %	78.57 %
Economía	69.23 %	69.23 %	61.53 %	61.53 %	69.23 %	69.23 %	69.23 %	69.23 %
Sociedad	68.29 %	82.92 %	82.92 %	80.48 %	68.29 %	**87.80 %**	82.92 %	75.60 %
Política	73.19 %	**80.41 %**	72.16 %	74.22 %	72.16 %	72.16 %	73.19 %	76.28 %
Espect.	68.88 %	64.44 %	51.11 %	**71.11 %**	66.66 %	68.88 %	62.22 %	64.44 %

La tabla 5.23 muestra mejores resultados usando palabras auxiliares, sin embargo algunos resultados permanecieron con valores similares a los experimentos sin palabras auxiliares como educación y economía. La tabla 5.24 muestra los resultados de n-gramas de palabras sin palabras auxiliares, aunque en la tabla 5.25 se observan mejores resultados del mismo experimento pero con palabras auxiliares, en particular en 2-gramas y 3-gramas.

En la tabla 5.26 se puede observar que la suma de n-gramas (2+3+4+5) de palabras con palabras auxiliares tiene mejores resultados, a diferencia de no utilizar palabras auxiliares.

Los experimentos con n-gramas de caracteres sin palabras auxiliares, tabla 5.27 en general dan mejores resultados en los 3-gramas. En la tabla 5.28 se muestran los resultados de n-gramas de caracteres con palabras auxiliares, en la cual se observan mejores resultados en 3-gramas y 5-gramas

Tabla 5.28: Resultados de entrenamientos de n-gramas de caracteres con palabras auxiliares, sin utilizar la categoría de prueba.

	N-gramas de caracteres con palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	57.14 %	71.42 %	57.14 %	71.42 %	71.42 %	71.42 %	85.71 %
Deporte	63.15 %	63.15 %	63.15 %	57.84 %	60.52 %	65.78 %	63.15 %	**73.68 %**
CyT	70.37 %	66.66 %	70.37 %	70.37 %	70.37 %	74.07 %	77.77 %	81.48 %
Seguridad	46.15 %	46.15 %	84.61 %	76.92 %	76.92 %	92.30 %	92.30 %	84.61 %
Salud	85.71 %	85.71 %	85.71 %	**92.85 %**	**92.85 %**	85.71 %	85.71 %	85.71 %
Economía	61.53 %	61.53 %	69.23 %	53.84 %	76.92 %	69.23 %	69.23 %	69.23 %
Sociedad	63.41 %	60.97 %	78.04 %	65.85 %	60.97 %	68.29 %	**85.36 %**	73.17 %
Política	70.10 %	68.04 %	72.16 %	72.16 %	71.13 %	74.22 %	72.16 %	78.35 %
Espect.	68.88 %	66.66 %	62.22 %	60.00 %	71.11 %	68.88 %	53.33 %	68.88 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	71.42 %	71.42 %	57.14 %	85.71 %	71.42 %	85.71 %	85.71 %	**100.00 %**
Deporte	68.42 %	63.15 %	63.15 %	68.42 %	63.15 %	65.78 %	63.15 %	55.26 %
CyT	77.77 %	85.18 %	74.07 %	74.07 %	70.37 %	77.77 %	66.66 %	**92.59 %**
Seguridad	**100.00 %**	92.30 %	92.30 %	92.30 %	**100.00 %**	92.30 %	**100.00 %**	84.61 %
Salud	85.71 %	85.71 %	85.71 %	85.71 %	85.71 %	85.71 %	85.71 %	85.71 %
Economía	76.92 %	61.53 %	69.23 %	61.53 %	**79.62 %**	69.23 %	69.23 %	69.23 %
Sociedad	70.73 %	78.04 %	**85.36 %**	73.17 %	70.73 %	**85.36 %**	82.92 %	82.92 %
Política	74.22 %	74.22 %	75.25 %	71.44 %	74.22 %	75.25 %	75.25 %	**79.38 %**
Espect.	**73.33 %**	71.11 %	57.77 %	62.22 %	71.11 %	64.44 %	62.22 %	68.88 %

La tabla 5.30 muestra los resultados de n-gramas POS sin palabras auxiliares, los 2-gramas tienen mejores resultados, sin embargo en la tabla 5.31 se observa que con palabras auxiliares se obtienen en general mejores resultados, aunque los 2-gramas continúan teniendo siendo los más altos.

Tabla 5.29: Resultados de entrenamientos concatenando n-gramas (2+3+4+5) de caracteres con y sin palabras auxiliares, sin utilizar la categoría de prueba.

	Suma de n-gramas (2+3+4+5) de caracteres sin palabras auxiliares				**Suma de n-gramas (2+3+4+5) de caracteres con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	85.71 %	57.14 %	57.14 %	85.71 %	71.42 %	71.42 %	85.71 %	85.71 %
Deporte	52.63 %	55.26 %	63.15 %	60.52 %	60.52 %	65.78 %	63.15 %	68.42 %
CyT	74.07 %	70.37 %	74.07 %	77.77 %	81.48 %	81.48 %	77.77 %	70.37 %
Seguridad	76.92 %	**92.30 %**	**92.30 %**	**92.30 %**	**92.30 %**	**92.30 %**	84.61 %	**92.30 %**
Salud	64.28 %	71.42 %	85.71 %	78.57 %	85.71 %	85.71 %	85.71 %	85.71 %
Economía	69.23 %	69.23 %	61.53 %	61.53 %	69.23 %	69.23 %	69.23 %	61.53 %
Sociedad	63.41 %	75.60 %	80.48 %	70.73 %	70.73 %	78.04 %	87.80 %	80.48 %
Política	69.07 %	79.38 %	70.10 %	77.31 %	73.19 %	71.13 %	74.22 %	80.41 %
Espect.	68.88 %	55.55 %	53.33 %	71.11 %	71.11 %	66.66 %	55.55 %	71.11 %

Tabla 5.30: Resultados de entrenamientos de n-gramas de POS eliminando palabras auxiliares, sin utilizar la categoría de prueba.

	N-gramas POS sin palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	71.42 %	71.42 %	71.42 %	**85.71 %**	42.85 %	28.57 %	71.42 %	42.85 %
Deporte	50.00 %	55.26 %	57.89 %	50.00 %	68.42 %	**71.05 %**	63.15 %	50.00 %
CyT	51.85 %	48.14 %	**70.37 %**	51.85 %	**70.37 %**	70.37 %	66.66 %	59.25 %
Seguridad	**92.30 %**	84.61 %	76.92 %	53.84 %	**92.30 %**	**92.30 %**	84.61 %	69.23 %
Salud	78.57 %	78.57 %	78.57 %	64.28 %	42.85 %	42.85 %	85.71 %	78.47 %
Economía	61.53 %	38.46 %	53.84 %	53.84 %	30.76 %	30.76 %	69.23 %	53.84 %
Sociedad	70.73 %	75.60 %	**80.48 %**	70.73 %	63.41 %	65.85 %	70.73 %	63.41 %
Política	68.04 %	65.97 %	64.94 %	67.01 %	60.82 %	54.63 %	67.01 %	63.91 %
Espect.	55.55 %	48.88 %	46.66 %	53.33 %	48.88 %	44.44 %	37.77 %	55.55 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	57.14 %	57.14 %	85.71 %	57.14 %	42.85 %	71.42 %	57.14 %
Deporte	60.52 %	57.89 %	63.15 %	39.47 %	57.89 %	57.89 %	63.15 %	60.52 %
CyT	59.25 %	48.14 %	62.96 %	55.55 %	51.85 %	59.25 %	62.96 %	**70.37 %**
Seguridad	84.61 %	76.92 %	84.61 %	84.61 %	69.23 %	84.61 %	**92.30 %**	76.92 %
Salud	42.85 %	50.00 %	**85.71 %**	64.28 %	57.14 %	71.42 %	**85.71 %**	57.14 %
Economía	61.53 %	53.84 %	69.23 %	53.84 %	**76.92 %**	69.23 %	61.53 %	38.46 %
Sociedad	68.29 %	60.97 %	68.29 %	63.41 %	53.65 %	53.65 %	73.17 %	70.73 %
Política	65.97 %	62.88 %	64.94 %	57.73 %	60.82 %	61.85 %	**70.10 %**	57.73 %
Espect.	57.77 %	**62.22 %**	46.66 %	51.11 %	42.22 %	40.00 %	46.66 %	46.66 %

Tabla 5.31: Resultados de entrenamientos de n-gramas de POS con palabras auxiliares, sin utilizar la categoría de prueba.

	N-gramas POS con palabras auxiliares							
	2-gramas				**3-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	85.71 %	85.71 %	85.71 %	85.71 %	71.42 %	85.71 %	71.42 %	57.14 %
Deporte	57.89 %	60.52 %	65.78 %	55.26 %	52.63 %	52.63 %	65.78 %	55.26 %
CyT	**74.07 %**	66.66 %	66.66 %	55.55 %	62.96 %	66.66 %	66.66 %	59.25 %
Seguridad	76.92 %	76.92 %	**92.30 %**	76.92 %	84.61 %	84.61 %	92.30 %	76.92 %
Salud	78.57 %	85.71 %	85.71 %	71.42 %	64.28 %	71.42 %	85.71 %	85.71 %
Economía	46.15 %	46.15 %	53.84 %	53.84 %	**69.23 %**	**69.23 %**	53.84 %	61.53 %
Sociedad	63.41 %	73.17 %	**82.92 %**	73.17 %	60.97 %	73.17 %	80.48 %	60.97 %
Política	72.16 %	72.16 %	**75.25 %**	64.94 %	57.73 %	57.73 %	72.16 %	71.13 %
Espect.	**71.11 %**	66.66 %	64.44 %	60.00 %	57.77 %	66.66 %	55.55 %	68.88 %
	4-gramas				**5-gramas**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	71.42 %	71.42 %	71.42 %	71.42 %	71.42 %	57.14 %	57.14 %	71.42 %
Deporte	55.26 %	60.52 %	65.78 %	63.15 %	60.52 %	63.15 %	65.78 %	60.52 %
CyT	55.55 %	55.55 %	62.96 %	70.37 %	66.66 %	62.96 %	59.25 %	48.14 %
Seguridad	69.23 %	76.92 %	84.61 %	**92.30 %**	61.53 %	69.23 %	84.61 %	76.92 %
Salud	71.42 %	71.42 %	85.71 %	78.57 %	64.28 %	71.42 %	85.71 %	71.42 %
Economía	53.84 %	46.15 %	53.84 %	69.23 %	46.15 %	53.84 %	53.84 %	53.84 %
Sociedad	68.29 %	75.60 %	73.17 %	70.73 %	75.60 %	78.04 %	73.17 %	65.85 %
Política	71.13 %	70.10 %	64.94 %	65.97 %	69.07 %	72.16 %	62.88 %	64.94 %
Espect.	55.55 %	62.22 %	55.55 %	68.88 %	57.77 %	66.66 %	60.00 %	66.66 %

Tabla 5.32: Resultados de entrenamientos concatenando n-gramas (2+3+4+5) de POS con y sin palabras auxiliares, sin utilizar la categoría de prueba.

	Suma de n-gramas (2+3+4+5) POS sin palabras auxiliares				**Suma de n-gramas (2+3+4+5) POS con palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	57.14 %	57.14 %	57.14 %	57.14 %	57.14 %	71.42 %	71.42 %	85.71 %
Deporte	65.78 %	65.78 %	63.15 %	55.26 %	65.78 %	50.00 %	65.78 %	57.89 %
CyT	55.55 %	62.96 %	66.66 %	51.85 %	66.66 %	62.96 %	70.37 %	62.96 %
Seguridad	84.61 %	84.61 %	**92.30 %**	**92.30 %**	61.53 %	76.92 %	**92.30 %**	84.61 %
Salud	57.14 %	64.28 %	85.71 %	50.00 %	64.28 %	71.42 %	85.71 %	78.57 %
Economía	69.23 %	61.53 %	69.23 %	38.46 %	61.53 %	53.84 %	61.53 %	46.15 %
Sociedad	60.97 %	58.53 %	73.17 %	56.09 %	73.17 %	63.41 %	78.04 %	75.60 %
Política	64.94 %	65.97 %	67.01 %	62.88 %	62.88 %	72.16 %	67.01 %	69.07 %
Espect.	57.77 %	57.77 %	42.22 %	57.77 %	68.88 %	68.88 %	55.55 %	66.66 %

En la tablas 5.32 se pude observar que en general los valores mejoran cuando se ha concatenado también las palabras auxiliares, además del POS. De la misma manera en la tabla 5.33 al concatenar todas las características de esta sección, los resultados son mejor con palabras auxiliares, aunque en algunas categorías son más altos los resultados sin palabras auxiliares en la mayoría de las categorías aumenta el porcentaje cuando se utilizan las palabras auxiliares.

Tabla 5.33: Resultados de entrenamientos concatenando n-gramas (2+3+4+5) de POS, caracteres, palabras, con y sin palabras auxiliares, sin utilizar la categoría de prueba.

	Suma de n-gramas (2+3+4+5) POS + caracteres + palabras				**Suma de n-gramas (2+3+4+5) POS + caracteres + palabras + palabras auxiliares**			
	Regresión logística	SVM	Bosques aleatorios	Boosting	Regresión logística	SVM	Bosques aleatorios	Boosting
Educación	85.71 %	57.14 %	57.14 %	85.71 %	71.42 %	71.42 %	85.71 %	**100.00 %**
Deporte	55.26 %	57.89 %	63.15 %	73.68 %	55.26 %	60.52 %	63.15 %	65.78 %
CyT	70.37 %	70.37 %	74.07 %	70.37 %	81.48 %	85.18 %	74.07 %	70.37 %
Seguridad	76.92 %	**100.00 %**	92.30 %	92.30 %	84.61 %	92.30 %	85.71 %	85.71 %
Salud	57.14 %	71.42 %	85.71 %	78.57 %	85.71 %	78.57 %	85.71 %	85.71 %
Economía	69.23 %	69.23 %	69.23 %	61.53 %	76.92 %	69.23 %	69.23 %	76.92 %
Sociedad	68.29 %	78.04 %	73.17 %	82.92 %	68.29 %	75.60 %	80.48 %	82.92 %
Política	68.04 %	75.25 %	67.01 %	72.12 %	73.19 %	74.07 %	65.78 %	79.38 %
Espect.	66.66 %	53.33 %	51.11 %	68.88 %	75.55 %	68.88 %	53.33 %	71.11 %

Conclusiones

El objetivo del presente trabajo terminal refiere a la identificación y clasificación de noticias falsas en el idioma español, a partir de las características gramaticales de los textos publicados en los medios digitales. Para lograr dicho objetivo, inicialmente se propone la compilación de un corpus de noticias etiquetas de manera binaria, es decir, utilizando dos clases: verdadera y falsa, utilizando técnicas reportadas en el estado del arte de los sistemas de clasificación para la asignación de clases.

Debido a la cantidad de información encontrada en los medios digitales, para la compilación del corpus de entrenamiento se propone un algoritmo de clasificación manual, el cual permite realizar las tareas de identificación y extracción de las noticias. El algoritmo propuesto, permite definir que la detección de noticias falsas de forma manual es un proceso tardado y complejo, debido a la cantidad de variables a considerar, como lo son los textos satíricos, la identificación de sitios que divulgan las noticias falsas, incluso el contexto social que favorece a la creación y difusión de este tipo de noticias.

Al realizar la compilación del corpus, se logró detectar que la difusión de las noticias falsas no es una tarea exclusiva de sitios específicos. Como se ha mencionado previamente, debido a que la identificación de estas noticias es subjetiva, algunos medios como El Universal y Televisa, han difundido noticias falsas de forma no intencional, al compartir información popular en las redes sociales sin verificar las fuentes de información. A partir del análisis de resultados, se logró identificar patrones sobre la composición de las noticias falsas como el uso continuo de artículos, sustantivos, así como palabras características de cierto tipo de publicaciones como el caso de la palabra “dizque”. Estas características proporcionan

información para la correcta clasificación de las noticias falsas. En ocasiones estos patrones tienen implícita información temática.

A partir de los patrones identificados, se pueden inferir que el modelo (Bolsa de palabras) creado usando aprendizaje automático y herramientas lingüísticas como (n-gramas, palabras auxiliares, etiquetas POS, caracteres, lematización) y un clasificador (máquinas de soporte vectorial, bosques aleatorios, boosting, regresión logística), permite obtener un obtener un modelo para clasificación de noticias falsas, arrojando resultados con un porcentaje de precisión mayor al 70 %, siendo *Random Forest* y *SVM* los clasificadores que mejores resultados obtuvieron en el proceso de clasificación.

Los experimentos también permitieron definir que las características con mayor peso para la clasificación de noticias falsas son los n-gramas de caracteres, así como las etiquetas POS al ser concatenadas con las palabras de los textos.

Como trabajo futuro se propone hacer uso de los metadatos de los textos como características complementarias de las noticias, así como el uso algoritmos de aprendizaje profundo como las redes neuronales como herramienta para la creación del modelo de clasificación.

Las contribuciones del proyecto tienen dos vertientes principales, siendo la primera de la identificación de aquellas noticias potencialmente falsas para evitar la distribución de información cuyo propósito sea uno diferente a comunicar hechos de relevancia, como lo es el contenido satírico. La segunda se refiere a que el proyecto también es viable para implementarse en rede sociales como un mecanismo de seguridad para evitar la difusión de noticias falsas.

Apéndices

Anexo A

Interfaz gráfica

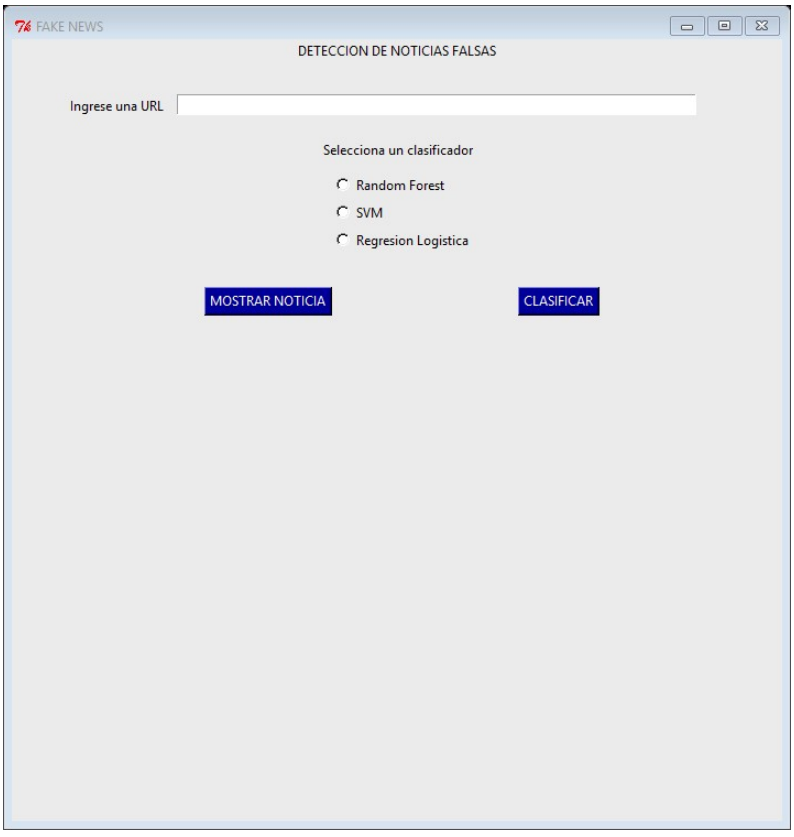


Figura A.1: Interfaz gráfica inicial

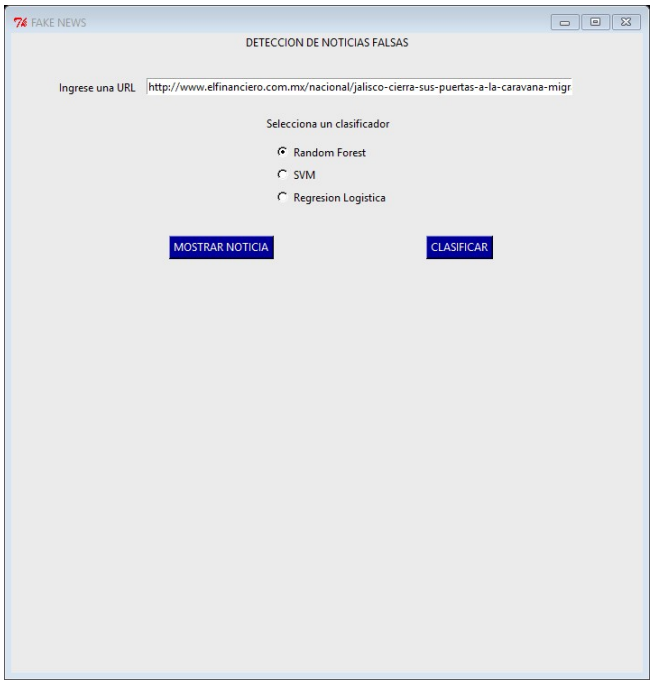


Figura A.2: Ingreso de la URL de la noticia a clasificar

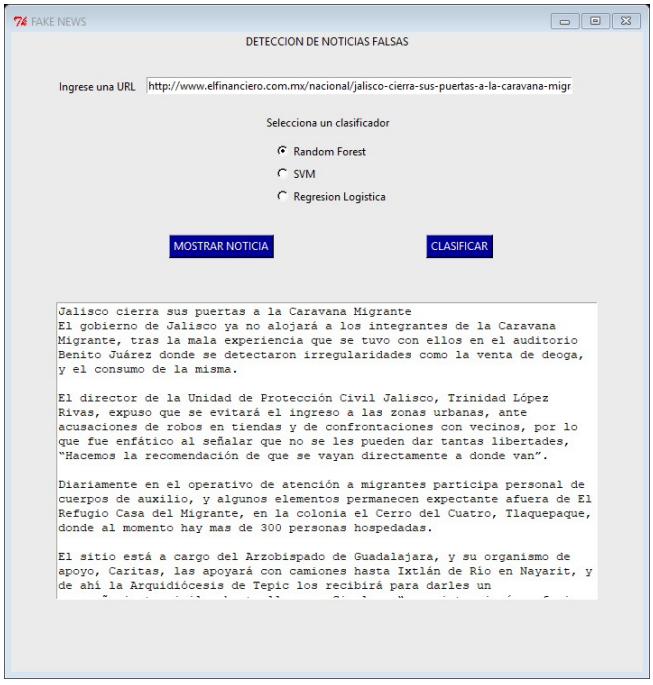


Figura A.3: Visualización del texto de la noticia

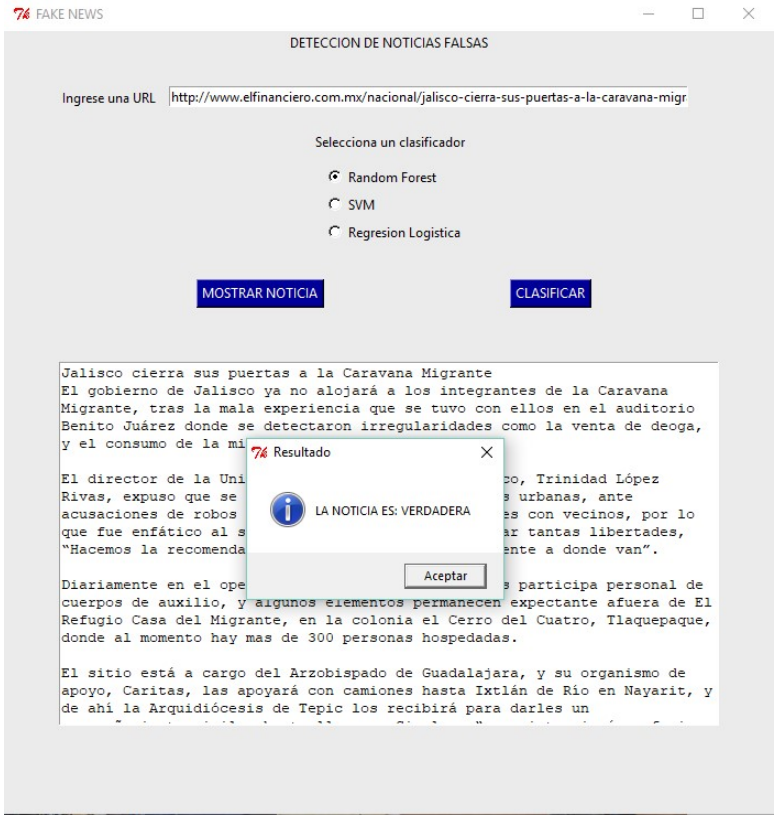


Figura A.4: Visualización del resultado del clasificador

```
from Tkinter import *
import tkMessageBox
from newspaper import Article
import spacy
import textacy
from spacy.lemmatizer import Lemmatizer
from spacy.lang.es.stop_words import STOP_WORDS
import os.path as path
import es_core_news_sm
import warnings
warnings.filterwarnings("ignore", message="numpy.dtype␣size␣changed")
warnings.filterwarnings("ignore", message="numpy.ufunc␣size␣changed")
```

```
import warnings
warnings.filterwarnings("ignore", category=FutureWarning)
import pickle


def get_text (titulo,texto):

    noticia=""
    noticia=titulo + "\n" + texto
    return noticia


def scrapping ():

    link=url.get()
    noticia=Article(link, language="es")
    noticia.download()
    noticia.parse()
    noticia.nlp()

    titulo= noticia.title
    texto=noticia.text
    noticia_= get_text(titulo, texto)

    return noticia_


def show_new ():

    noticia=scrapping()

    scrollbar = Scrollbar(root)
    scrollbar.pack(side=RIGHT, fill=Y)

    text = Text(root, wrap=WORD, yscrollcommand=scrollbar.set)
    text.insert(INSERT, noticia)
    text.config(state=DISABLED)
    text.pack()
```

```
    text.place(height=330, width=600, x=50,y=300)
    scrollbar.config(command=text.yview)
    scrollbar.pack_forget()



def preprocesar (noticia):

    item=textacy.preprocess.fix_bad_unicode(noticia, normalization='NFC')

    item=textacy.preprocess.normalize_whitespace(item)

    item=textacy.preprocess.preprocess_text(item, no_currency_symbols=
        True, no_punct=True, no_urls=True, no_emails=True,
        no_phone_numbers=True, no_numbers=True)

    return item

def lematizador (noticia,nlp):

    lema=[]
    doc=nlp(noticia)

    for token in doc:
        palabras=token.lemma_
        palabras=palabras.lower()
        lema.append(palabras)

    return lema


def prueba (noticia):

    ln=""
    lista=[]

    for word in noticia:
        if word!= "\n":
```

```
            lista.append(word)

    for w in lista:
        ln+=w+ "␣"
    ln=ln.rstrip()

    return ln


def test_new():

    nlp = es_core_news_sm.load()
    noticia=scrapping()
    lista=[]

    text_prep=preprocesar(noticia)
    lemas=lematizador(text_prep, nlp)
    test=prueba(lemas)
    lista.append(test)

    return lista

def clasificacion():

    test=test_new()
    ruta="Corpus/modelos.pickle"

    tr = open(ruta, 'r')
    data = pickle.load(tr)
    modelos=data.values()
    RF=modelos[0]
    Vector=modelos[1]
    SVM=modelos[2]
    LR=modelos[3]
    tr.close()


    X=Vector.transform(test)
```

```
    clasif=opcion.get()

    if clasif==1:
        clf= RF
    elif clasif==2:
        clf= SVM
    elif clasif==3:
        clf=LR

    prediccion = clf.predict(X)

    pred=str(prediccion[0])
    resultado="LA NOTICIA ES: " + pred.upper()
    tkMessageBox.showinfo("Resultado ", resultado)



if __name__ == "__main__":

    root = Tk()
    url=StringVar()
    opcion=IntVar()
    fondo="gray92"
    letra="black"

    '''_____ INTERFAZ DEFINICION_____'''

    root.title("FAKE NEWS")
    root.geometry("700x750")
    root.minsize(width=700, height=710)
    root.maxsize(width=700, height=710)
    root.configure(background = fondo)

    etiquetaTitulo = Label(root, text="DETECCION DE NOTICIAS FALSAS", bg=
        fondo, fg=letra).pack(anchor='center')

    box_url=Entry(root, textvariable=url, width=78 ,selectbackground='
        blue',cursor="hand2").place(x=150,y=50)
```

```
tagURL=Label(root, text="Ingrese una URL", bg=fondo, fg=letra).place(
    x=50, y=50)



'''__________CLASIFICADORES_________________'''

etiquetaClasif = Label(root, text="Selecciona un clasificador", bg=
    fondo, fg=letra).place(x=280,y=90)

random=Radiobutton(root, text="Random Forest",value=1,variable=opcion
    ,bg=fondo,cursor="hand2").place(x=290, y=120)

svm=Radiobutton(root, text="SVM",value=2,variable=opcion,bg=fondo,
    cursor="hand2").place(x=290, y=145)

rl=Radiobutton(root, text="Regresion Logistica",value=3,variable=
    opcion,bg=fondo,cursor="hand2").place(x=290, y=170)


'''__________BOTONES_________________'''

SHOW= Button(root, text="MOSTRAR NOTICIA", command=show_new, bg="#009
    ", fg="white",cursor="hand2").place(x=175, y=225)

clasificar= Button(root, text="CLASIFICAR", command=clasificacion, bg
    ="#009", fg="white", cursor="hand2").place(x=460, y=225)

root.mainloop()
```

Anexo B

Módulos

Tabla B.1: Funciones disponibles en la librería Textacy

Función	**Descripción**
textacy.preprocess.fix_bad_unicode ()	Repara el texto Unicode que está roto usando ftffy.
textacy.preprocess.normalize_whitespace ()	Remplaza uno o más espacios con uno solo y uno o más saltos de línea con una sola línea.
textacy.preprocess.preprocess_text ()	Aplica todas las funciones de preprocesamiento (Elimina acentos, puntuación, reemplaza símbolo de moneda por MON, correos electrónicos por EMAIL, números por NUM, números de teléfono por TEL y los enlaces por URL).

Funciones de módulo textacy para preprocesamiento de textos.

Anexo C

Etiquetado de clases gramaticales (POS)

Tabla C.1: Etiquetado de clases gramaticales (POS) de spaCy.

POS	Descripción	Ejemplos
ADJ	adjective	grande, viejo, verde
ADP	adposition	para, a, en
ADV	adverb	adelante, atrás, ahí
AUX	auxiliary	is, has (done), will (do), should (do)
CONJ	conjunction	y, e, o
CCONJ	coordinating conjunction	y, e, ni
DET	determiner	a, el
INTJ	interjection	psst, ouch
NOUN	noun	niña, gato, rama, aire
NUM	numeral	1, 2017, uno, veinte, IV
PART	particle	’s, not,
PRON	pronoun	yo, tú, él, ella, nosotros, alguien
PROPN	proper noun	Pablo, Juan, Londres, HBO
PUNCT	punctuation	., (,), ?
SCONJ	subordinating conjunction	porque, ya que, que
SYM	symbol	$, %, §, ©, +, , ×, ÷, =, :), XD
VERB	verb	correr, corrió, correrá, corriendo
X	other	sfpksdpsxmsa
SPACE	space	

Anexo D

Lista de palabras auxiliares de NLTK

Tabla D.1: Lista de palabras auxiliares de NLTK

Palabras auxiliares de NLTK				
a	ellas	estaban	estarían	estuvieseis
al	ellos	estabas	estarías	estuviésemos
algo	en	estad	estas	estuviesen
algunas	entre	estada	estás	estuvieses
algunos	era	estadas	este	estuvimos
ante	erais	estado	esté	estuviste
antes	éramos	estados	estéis	estuvisteis
como	eran	estáis	estemos	estuvo
con	eras	estamos	estén	fue
contra	eres	están	estés	fuera
cual	es	estando	esto	fuerais
cuando	esa	estar	estos	fuéramos
de	esas	estará	estoy	fueran
del	ese	estarán	estuve	fueras
desde	eso	estarás	estuviera	fueron
donde	esos	estaré	estuvierais	fuese
durante	esta	estaréis	estuviéramos	fueseis
e	está	estaremos	estuvieran	fuésemos
el	estaba	estaría	estuvieras	fuesen
él	estabais	estaríais	estuvieron	fueses

ella	estábamos	estaríamos	estuviese	fui
fuimos	habría	hubieras	mía	otras
fuiste	habríais	hubieron	mías	otro
fuisteis	habríamos	hubiese	mío	otros
ha	habrían	hubieseis	míos	para
habéis	habrías	hubiésemos	mis	pero
había	han	hubiesen	mucho	poco
habíais	has	hubieses	muchos	por
habíamos	hasta	hubimos	muy	porque
habían	hay	hubiste	nada	que
habías	haya	hubisteis	ni	qué
habida	hayáis	hubo	no	quien
habidas	hayamos	la	nos	quienes
habido	hayan	las	nosotras	se
habidos	hayas	le	nosotros	sea
habiendo	he	les	nuestra	seáis
habrá	hemos	lo	nuestras	seamos
habrán	hube	los	nuestro	sean
habrás	hubiera	más	nuestros	seas
habré	hubierais	me	o	sentid
habréis	hubiéramos	mi	os	sentida
habremos	hubieran	mí	otra	sentidas
sentido	soy	tened	tienen	tuvisteis
sentidos	su	tenéis	tienes	tuvo
será	sus	tenemos	todo	tuya
serán	suya	tenga	todos	tuyas
serás	suyas	tengáis	tu	tuyo
seré	suyo	tengamos	tú	tuyos
seréis	suyos	tengan	tus	un
seremos	también	tengas	tuve	una
sería	tanto	tengo	tuviera	uno
seríais	te	tenía	tuvierais	unos
seríamos	tendrá	teníais	tuviéramos	vosostras
serían	tendrán	teníamos	tuvieran	vosostros
vuestra	vuestras	vuestro	vuestros	y
ya	yo			

Anexo E

Lista de palabras auxiliares de spaCy

Tabla E.1: Lista de palabras auxiliares de spaCy

Palabras auxiliares de spaCy				
actualmente	acuerdo	adelante	ademas	además
adrede	afirmó	agregó	ahi	ahí
ahora	al	algo	algún	alguna
algunas	alguno	algunos	alli	allí
alrededor	ambos	ampleamos	antano	antaño
ante	anterior	antes	añadió	apenas
aproximadamente	aquel	aquél	aquella	aquélla
aquellas	aquéllas	aquello	aquellos	aquéllos
aqui	aquí	arriba	arribaabajo	aseguró
asi	así	atras	aun	aún
aunque	ayer	bajo	bastante	bien
breve	buen	buena	buenas	bueno
bueno	buenos	cada	casi	cerca
cierta	ciertas	cierto	ciertos	cinco
claro	comentó	como	cómo	con
conmigo	conocer	conseguimos	conseguir	considera

consideró	consigo	consigue	consiguen	consigues
contigo	contra	cosas	creo	cual
cuál	cuales	cuáles	cualquier	cuando
cuándo	cuanta	cuánta	cuantas	cuántas
cuanto	cuánto	cuantos	cuántos	cuatro
cuenta	da	dado	dan	dar
de	debajo	debe	deben	debido
decir	dejó	del	delante	demás
demasiado	dentro	deprisa	desde	despacio
despues	después	detras	detrás	dia
día	dias	días	dice	dicen
dicho	dieron	diferente	diferentes	dijeron
dijo	dio	donde	dónde	dos
durante	ejemplo	el	él	ella
ellas	ello	ellos	embargo	empleais
emplean	emplear	empleas	empleo	en
encima	encuentra	enfrente	enseguida	entonces
entre	era	eramos	eran	eras
eres	es	esa	ésa	esas
ésas	ese	ése	eso	esos
ésos	esta	está	ésta	estaba
estaban	estado	estados	estais	estamos
estan	estar	estará	estas	éstas
este	éste	esto	estos	éstos
estoy	estuvo	ex	excepto	existe
existen	explicó	expresó	expresó	fin
final	fue	fuera	fueron	fui
fuimos	general	gran	grandes	gueno
ha	haber	habia	había	habían
habla	hablan	habrá	hace	haceis
hacemos	hacen	hacer	hacerlo	haces
hacia	haciendo	hago	han	hasta
hay	haya	he	hecho	hemos

hicieron	hizo	horas	hoy	hubo
igual	incluso	indicó	informo	informó
intenta	intentais	intentamos	intentan	intentar
intentas	intento	ir	junto	la
lado	largo	las	le	lejos
les	llegó	lleva	llevar	lo
los	luego	lugar	mal	manera
manifestó	mas	más	mayor	me
mediante	medio	mejor	mencionó	menos
menudo	mi	mí	mia	mía
mias	mías	mientras	mio	mío
mios	míos	mis	misma	mismas
mismo	mismos	modo	momento	mucha
muchas	mucho	muchos	muy	nada
nadie	ni	ningún	ninguna	ningunas
ninguno	ningunos	no	nos	nosotras
nosotros	nuestra	nuestras	nuestro	nuestros
nueva	nuevas	nuevo	nuevos	nunca
ocho	os	otra	otras	otro
otros	pais	paìs	para	parece
parte	partir	pasada	pasado	peor
pero	pesar	poca	pocas	poco
pocos	podeis	podemos	poder	podrá
podrán	podria	podría	podriais	podriamos
podrian	podrían	podrias	poner	por
porque	posible	primer	primera	primero
primeros	principalmente	pronto	propia	propias
propio	propios	proximo	próximo	próximos
pudo	pueda	puede	pueden	puedo
pues	qeu	que	qué	quedó
queremos	quien	quién	quienes	quiénes
quiere	quiza	quizá	quizas	quizás
raras	realizado	realizar	realizó	repente
respecto	sabe	sabeis	sabemos	saben

saber	sabes	salvo	se	sé
sea	sean	segun	según	segunda
segundo	seis	señaló	ser	sera
será	serán	sería	si	sí
sido	siempre	siendo	siete	sigue
siguiente	sino	sobre	sois	sola
solamente	solas	solo	sólo	solos
somos	son	soy	soyos	su
supuesto	sus	suya	suyas	suyo
tal	tambien	también	tampoco	tan
tanto	tarde	te	temprano	tendrá
tendrán	teneis	tenemos	tener	tenga
tengo	tenía	tenido	tercera	ti
tiempo	tiene	tienen	toda	todas
todavia	todavía	todo	todos	total
trabaja	trabajais	trabajamos	trabajan	trabajar
trabajas	trabajo	tras	trata	través
tres	tu	tú	tus	tuvo
tuya	tuyas	tuyo	tuyos	última
últimas	ultimo	último	últimos	un
últimas	ultimo	último	últimos	un
una	unas	uno	unos	usa
usais	usamos	usan	usar	usas
uso	usted	ustedes	va	vais
valor	vamos	van	varias	varios
vaya	veces	ver	verdad	verdadera
vez	vosotras	vosotros	voy	vuestra
vuestras	vuestro	vuestros	ya	yo

Anexo F

Glosario

- Lematización: Proceso lingüístico que consiste en hallar el lema correspondiente, dada una forma de una palbra (plural,femenino,conjugada,entre otras).
- Lema: Es la forma que por convenio se acepta como representante de todas las formas de una misma palabra.
- Token: Es una cadena con un significado asignado y por lo tanto identificado.El nombre del token es una categoría de unidad léxica.
- N-gramas: son secuencias de elementos textuales (palabras, caracteres, lemas, etiquetas de clases gramaticales, etc.) según su orden de aparición en documentos.
- Palabras auxiliares o *stopwords*: también denominadas *stopwords*, son aquellas palabras que no proporcionan información relevante acerca del contenido de los textos.
- Etiquetado de clases gramaticales o POS: Proceso de asignar (o etiquetar) a cada una de las palabras de un texto su categoría gramatical.
- Web scraping: Técnica utilizada mediante programas de software para extraer información de sitios Web.
- Dendograma: Diagrama de árbol que muestra los grupos que se forman al crear uniones de las observaciones que se encuentran en cada paso y sus niveles de similitud.

- Afijo:Partícula que se une a una palabra o a una base para formar palabras derivadas; puede aparecer al principio, en medio o al final de la palabra.
- Prefijo: Afijo que se añade al comienzo de una palabra para formar una palabra derivada.
- Sufijo: Afijo que se añade al final de una palabra o de su raíz para formar una palabra derivada.
- Infijo: Afijo con el que se forman, en el interior de una palabra derivada o de su lexema o raíz, palabras derivadas.
- Matriz dispersa: En álgebra lineal numérica es una matriz de gran tamaño en la que la mayor parte de sus elementos son cero.
- Preprocesamiento: Tecnicismo utilizado para referirse a la etapa de normalización de textos para posteriormente ser procesados.

Bibliografía

[1] Roberto Rodríguez-Andrés. Fundamentos del concepto de desinformación como práctica manipuladora en la comunicación política y las relaciones internacionales. 2017.

[2] Roberto Rodríguez-Andrés. Trump 2016:¿ presidente gracias a las redes sociales? *Palabra Clave*, 21(3):831–859, 2018.

[3] Émile Durkheim. *La educación moral*. Ediciones Morata, 2002.

[4] Luis Jesús Galindo Cáceres. *Técnicas de investigación en sociedad, cultura y comunicación*. Pearson Educación, 1998.

[5] Álex Grijelmo. *El estilo del periodista*. Taurus, 2014.

[6] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[7] Peter Krejzl, Barbora Hourová, and Josef Steinberger. Stance detection in online discussions. *arXiv preprint arXiv:1701.00504*, 2017.

[8] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168, 2016.

[9] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[10] Tahora H Nazer, Guoliang Xue, Yusheng Ji, and Huan Liu. Intelligent disaster response via social media analysis a survey. *ACM SIGKDD Explorations Newsletter*, 19(1):46–59, 2017.

[11] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.

[12] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. *arXiv preprint arXiv:1703.06959*, 2017.

[13] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

[14] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[17] H Yu, C Ho, Y Juan, and C Lin. Libshorttext: A library for short-text classification and analysis. *Rapport interne, Department of Computer Science, National Taiwan University. Software available at http://www. csie. ntu. edu. tw/cjlin/libshorttext*, 2013.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[19] Augusto Cortez Vásquez, Jaime Pariona Quispe, Ana Maria Huayna, et al. Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2):45–54, 2009.

[20] Grigori Sidorov. Construcción no lineal de n-gramas en la lingüística computacional. *Sociedad Mexicana de Inteligencia Artificial*, 2013.

[21] De repesentaciones clásicas a representaciones avanzadas para nlp, 2018.

[22] María del Consuelo Justicia de la Torre et al. *Nuevas técnicas de minería de textos: Aplicaciones*. PhD thesis, Universidad de Granada.

[23] Luis Rodríguez Yunta. La lematización en español: una aplicación para la recuperación de información (r. gómez díaz). *Revista española de Documentación Científica*, 29(1):175–176, 2006.

[24] Stemming and lemmatization, 2008.

[25] Adaptación, optimización y expansión de ecode un sistema extractor de contextos definitorios, 2012.

[26] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.

[27] Lior Rokach and Oded Z Maimon. *Data mining with decision trees: theory and applications*, volume 69. World scientific, 2008.

[28] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[29] Sebastián Maldonado and Richard Weber. Modelos de selección de atributos para support vector machines. *Revista Ingeniería de Sistemas*, 26:49–70, 2012.

[30] Enrique J Carmona Suárez. Tutorial sobre máquinas de vectores soporte (svm). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*, 2014.

[31] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Citeseer, 1996.

[32] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[33] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147, 1974.

[34] Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 115(16):1058–1073, 2015.

[35] José Otero and Luciano Sánchez. Disenos experimentales y tests estadısticos, tendencias actuales en machine learning. In *V Congreso Espanol sobre Metaheurısticas, Algoritmos Evolutivos y Bioinspirados (MAEB'07). Universidad de La Laguna. Puerto de La Cruz (Espana, 2007)*, pages 295–302, 2007.

[36] *scikit-learn.org, Scikit-learn Machine Learning*. 2018-07-11.

[37] Guido Van Rossum et al. Python programming language. In *USENIX Annual Technical Conference*, volume 41, page 36, 2007.

[38] spacy.io , spacy procesamiento del lenguaje natural, 2018-07-15.

[39] www.nltk.org , natural language toolkit (nltk), 2018-07-01.

[40] Burton DeWilde. textacy documentation. 2017.

[41] *https://aws.amazon.com/es/tensorflow/,TensorFlow*. 2018-07-19.

[42] Tom M Mitchell and Machine Learning. Mcgraw-hill science. *Engineering/Math*, 1:27, 1997.

[43] Técnica de clasificación bayesiana para identificar posible plagio en información textual, 2014.

[44] An introduction to information retrieval, 2009.

[45] https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html, talosintelligence, 2018-07-11.

[46] *https://keras.io/,Keras*. 2018-07-19.