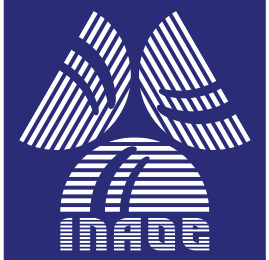

Enmascaramiento de la Información para la Detección Automática de Noticias Falsas

por

Jennifer Pérez Santiago

Disertación presentada en cumplimiento parcial
de los requisitos para
el grado de

MSc. en Ciencias Computacionales

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica *(INAOE)*
Tonantzintla, Puebla, Mexico
febrero, 2022

Tutores:

Dr. Luis Villaseñor Pineda
Coordinación de Ciencias Computacionales
INAOE, Mexico

Abstract

Currently, thanks to the availability of digital media, and in particular of social networks, users receive journalistic notes, opinions and information on a wide variety of topics on a daily basis. The same medium allows you to easily share and forward your own opinions, enriching the discussion and reflection on the topics of interest. Unfortunately, these circumstances have led to the increasingly frequent appearance of malicious fake news with the aim of misinformation. This phenomenon has reached enormous proportions, becoming a serious problem.

The objective of this thesis is to propose a method for the automatic detection of fake news by masking the textual information of the content and writing style of the news. From the experimentation carried out, it was possible to conclude that the proposed method locates and even improves the classification for the data sets used, by hiding the most general or the most specific words according to each model studied (style or content). The results achieved are encouraging, demonstrating the usefulness of the method. It should be noted that, to our knowledge, this work is the first to use information masking, not only of words and numbers, but also of punctuation marks and other symbols for the automatic detection of fake news

Keywords:. *Fake News*, SVM, CNN, masking, style, content, most frequent words

Resumen

Actualmente, gracias a la disponibilidad de los medios digitales, y en particular de las redes sociales, los usuarios reciben a diario notas periodísticas, opiniones e información de muy diversos temas. El mismo medio permite compartir y reenviar fácilmente opiniones propias enriqueciendo la discusión y reflexión de los temas de interés. Desafortunadamente, estas circunstancias han motivado la aparición, cada vez más frecuente, de noticias falsas malintencionadas con el objetivo de desinformar. Este fenómeno ha alcanzado enormes proporciones llegando a convertirse en un serio problema.

El objetivo de esta tesis es proponer un método para la detección automática de noticias falsas mediante el enmascarado de la información textual del contenido y estilo de escritura de la noticia. A partir de la experimentación realizada, fue posible concluir, que el método propuesto se sitúa e incluso mejora la clasificación para los conjuntos de datos utilizados, al ocultar las palabras más generales o las más específicas de acuerdo a cada modelo estudiado (estilo o contenido). Los resultados alcanzados son alentadores, demostrando la utilidad del método. Cabe resaltar que, según nuestro conocimiento, este trabajo es el primero en utilizar el enmascarado de la información, no solo de palabras y números sino también de signos de puntuación y otros símbolos para la detección automática de noticias falsas.

Palabras Claves: *Noticias Falsas*, SVM, CNN, enmasacarmiento, estilo, contenido, palabras más frecuentes

Agradecimientos

Esta investigación fue realizada gracias al apoyo otorgado por el Consejo Nacional de Ciencia y Tecnología (CONACYT), a través de la beca otorgada y del proyecto CB-2015-01-257383. De igual forma, agradecemos el apoyo otorgado por el CONACYT para acceder a los recursos computacionales proporcionados a través de la Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje del INAOE.

Al apoyo de mi asesor el doctor Luis Villaseñor Pineda y al doctor Manuel Montes y Gómez, por su colaboración. Por su entrega y sus conocimientos y experiencias transmitidos hacia mí, haciéndome sentir cómoda siempre con ellos en todas las consultas, pero sobre todo por su comprensión siempre y en cada etapa de mi vida como nueva madre de un bebé.

Agradezco al INAOE, a sus trabajadores, y especialmente a los profesores que compartieron su preparación y experiencias dentro del aula.

A los sinodales: Dr. Hugo Jair Escalante, Dr. Aurelio López López, y Dr. Carlos, les agradezco sus comentarios, sugerencias, y críticas.

Finalmente, y no menos importante a mi esposo, mi familia y amigos, muchas gracias por sus palabras y presencia.

Gracias a todos,
Jennifer.
Tonantzintla, Puebla, Mexico.
febrero, 2022.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

DV-MA	Vista Distorsionada Múltiples Asteriscos
DV-SA	Vista Distorsionada Simples Asteriscos
FCE	*Frequently Co-occurring Entropy*
SVM	*Support Vector Machine*
CNN	*Convolutional Neural Networks*

Capítulo 1

Introducción

Las noticias falsas han surgido como una tendencia mundial atrayendo la atención pública. Estudios recientes [Kumar and Shah, 2018] han demostrado, que las personas, obtienen cada vez más sus noticias de los medios digitales que de fuentes de noticias tradicionales, esto, debido a que los medios en línea son conocidos por su inmediatez y porque constituyen una alternativa gratuita para leer las noticias. Sin embargo, no se comprueba la autenticidad de las noticias que están presentes en línea [Vishwakarma and Jain, 2020].

Con el auge de las redes sociales, muchos usuarios reciben a diario sus noticias por este medio, que a su vez, rompe las barreras del distanciamiento físico y permite compartir, reenviar, revisar y opinar sobre todos los temas, amplificando y reforzando así la información sesgada [Kumar and Shah, 2018].

1.1 *Fake News.* Definición

La amplia difusión de noticias falsas trae consigo un grave impacto negativo en las personas y la sociedad. Las noticias falsas pueden romper el equilibrio de autenticidad del ecosistema de noticias, persuaden intencionalmente a los consumidores a aceptar creencias sesgadas o falsas y generalmente son manipuladas por propagandistas para transmitir mensajes políticos o de influencia. Las noticias falsas, cambian la forma en que las personas interpretan y responden a las noticias reales [Shu et al., 2017]. Se hace necesario, ayudar a mitigar los efectos negativos causados por las noticias falsas, desarrollar métodos que nos permitan identificar características discriminatorias y re-

levantes mediante un estudio de este tipo de noticias, así como la detección automática de las mismas.

El primer paso para realizar el análisis de las noticias falsas consiste en adoptar una definición clara y precisa. Sin embargo, esta tarea no es simple, pues el término *Fake News* ha estado ligado a otros términos. Tal como exponen Zhou and Zafarani [2020] y Koumouridis [2020] algunos de los términos asociados son: noticias falsas, noticias falsas maliciosas, noticias fabricadas, desinformación o engaño, *misinformation* (la cual podemos traducir como información errónea) y rumor. Los siguientes párrafos presentan la interrelación entre estos términos y se contrastan sus diferencias para finalmente presentar la definición de *Fake News* adoptada en este trabajo.

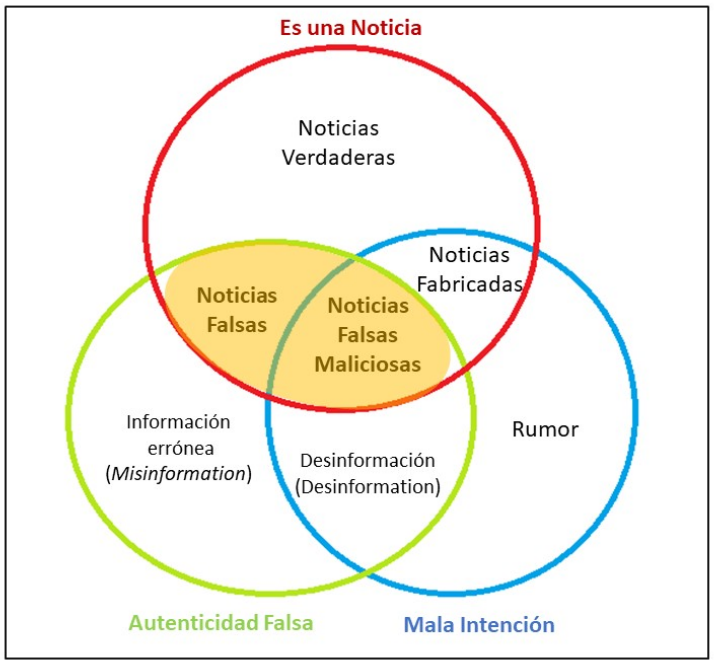

Figura 1.1: Términos Relacionados que tienden a confundirse con *Fake News* y su relación en cuanto a: autenticidad, intención y si la información es noticia o no.

En Zhou and Zafarani [2020] se distinguen los diferentes conceptos asociados al término *Fake News*, mediante tres características fundamentales: (i) autenticidad, (ii)

intención y (iii) si la información es noticia. Gracias a estas tres características es posible definir cual es el concepto detrás del término *Fake News*, teniendo en cuenta, sobre todo, las principales características y la naturaleza de las colecciones de datos con las cuales trabajamos en nuestra investigación.

	Autenticidad	**Intención**	**¿Es noticia?**
Noticias Falsas	Falsa	Desconocida	Si
Noticias Falsas Maliciosas	Falsa	Mala	Si
Noticias Fabricadas	Desconocida	Mala	Si
Desinformación	Falsa	Mala	Desconocida
Información Errónea (*Misinformation*)	Falsa	Desconocida	Desconocida
Rumor	Desconocida	Mala	Desconocida

Tabla 1.1: Tabla de comparación entre conceptos relacionados. Tomada como referencia de Zhou and Zafarani [2020]

Tomamos como referencia la Tabla 1.1 que describe las características para cada uno de estos términos teniendo en cuenta la autenticidad (falsa o no), intención (mala o no) y si la información es noticia o no. Esta tabla fue tomada de Zhou and Zafarani [2020] y modificada en algunos parámetros para ajustarla a nuestro enfoque y entendimiento. A partir de esta Tabla construimos un esquema, Figura 1.1 que nos permite apreciar mejor las relaciones entre los diferentes términos.

Tomando ambas, tabla y figura, como punto de partida, realizamos una búsqueda en la literatura, de la definición, para cada uno de estos términos. Cuando nos referimos, por ejemplo, al término **Rumor**, y comenzamos por él ya que lo consideramos como el término más alejado en lo que respecta a ***Fake News***. Este término, es definido por varios autores Buntain and Golbeck [2017], Koumouridis [2020] y Kumar et al. [2021] como *"un elemento de información no respaldado, una declaración no verificada y relevante que circula, y que podría luego confirmarse como verdadero, falso o no confirmado. Por lo tanto, su valor real no está resuelto"*. Lo único que se tiene claro, cuando lanzamos un rumor es que la intención no es buena, buscamos manipular y condicionar al lector por encima de la información objetiva, no necesariamente tiene

que estar redactado como noticia, de hecho, son muy escasos los rumores cuyo cuerpo informativo está estructurado como noticia y por lo que entendemos de la definición su autenticidad es desconocida.

Por su parte, Koumouridis [2020] hace una categorización más general de información falsa basada en su intención, la cual la subdivide en: mala información o información errónea (***misinformation***) y **desinformación** (***desinformation***).

Mala información o información errónea no es más que *"información falsa que puede ser el resultado de la tergiversación de una pieza original de información válida debido a sesgos cognitivos o falta de atención y se propaga de manera involuntaria"*. Sin embargo, la **Desinformación** *"es la creación y distribución de información falsa con la total intención de engañar y manipular a la audiencia"*. Lo común para ambos términos es que su autenticidad es falsa y no necesariamente la información es una noticia. La diferencia radica en que una se difunde con mala intención; es decir con la total intención de engañar y la otra simplemente desconoce su intención.

Entre los términos donde la información si es noticia, tenemos: las noticias fabricadas, las noticias falsas y las noticias falsas maliciosas. Según Koumouridis [2020] la **fabricación de noticias** se refiere a *"historias que no tienen una base fáctica pero que se publican al estilo de los artículos de noticias para crear legitimidad. La intención del productor es desinformar a los lectores. Las noticias inventadas generalmente se publican en sitios web, blogs o plataformas de redes sociales"*.

A diferencia de los anteriores, se maneja en la literatura otros dos términos: **noticias falsas maliciosas** y **noticias falsas**. Si nos referimos a las características resumidas de la Tabla 1.1 podemos ver que ambas son noticias, su autenticidad es falsa, sin embargo una tiene malas intenciones al buscar engañar al lector, como bien lo indica su nombre y para la segunda su intención no necesariamente tiene que ser mala.

En [Zhou et al., 2019], se describen estas dos definiciones, para el caso muy particular

de **noticias falsas maliciosas**: *"son noticias falsas intencionales y no verificables, publicadas por un medio de comunicación"*. Con esta definición se enfatiza tanto la autenticidad como en la intención de la noticia.

Por otro lado, para el término **noticias falsas** argumentan que *"son noticias que incluyen en general afirmaciones, declaraciones, discursos, publicaciones, entre otros tipos de información publicadas por un medio de comunicación y su autenticidad no es verificable"*. Esta definición, enfatiza en la autenticidad de la información y debilita la exigencia de intenciones de información. Sin embargo, esta definición es compatible con la mayoría de las noticias falsas existentes, estudios y conjuntos de datos.

Los conjuntos de datos de noticias falsas actuales (dentro de ellos con los que trabajamos en nuestra investigación) a menudo proporcionan una verdad fundamentada para la autenticidad de las afirmaciones, declaraciones, discursos o publicaciones, mientras que no se proporciona información sobre las intenciones.

Diversas son las definiciones que podemos encontrar en la literatura que intentan caracterizar el término ***Fake News*** [Allcott and Gentzkow, 2017], [Golbeck et al., 2018],[Sharma et al., 2019]. En el presente trabajo de investigación se tomará en cuenta la definición propuesta por Zhou and Zafarani [2020] y Abonizio et al. [2020], que relaciona el término ***Fake News*** simplemente con **Noticias Falsas**. Se selecciona esta definición por su compatibilidad, principalmente con los conjuntos de datos utilizados en nuestra investigación, en ella se enfatiza en la autenticidad de la noticia, se desconoce la intención de la misma y es una de las más aceptadas en la literatura consultada:

Fake News: *"son noticias publicadas por un medio de comunicación, que incluyen: afirmaciones, declaraciones, discursos, publicaciones, entre otros tipos de información y su autenticidad no es verificable (falsa)".*

1.1.1 Perspectivas de Estudio de las Noticias Falsas

Cuando estudiamos las noticias falsas, la información que extraemos o identificamos para distinguir las características puede estar relacionada con la propia noticia (titular, texto del cuerpo de la noticia, creador, editor) o relacionada con el contexto social en el que se desenvuelve (comentarios, red de propagación y difusores) cubriendo todo el ciclo de vida de la noticia, desde el momento en que se crea, hasta cuando se publica o se difunde [Zhou and Zafarani, 2020].

Dentro de este ciclo de vida de la noticia podemos distinguir el estudio de cuatro principales perspectivas (ver Figura 1.2): **(1) basadas en el conocimiento** [Zhou and Zafarani, 2020], no es más que detectar o analizar las noticias falsas mediante un proceso conocido como verificación de hechos (*fact-checking*); **(2) basada en la propagación**, que a diferencia de las perspectivas basadas en el conocimiento y el estilo que estudian las noticias falsas en función de su contenido, cuando se estudian las noticias falsas desde una perspectiva basada en la propagación [Zhou and Zafarani, 2020; Koumouridis, 2020], se aprovecha la información relacionada con la difusión de la noticia; **(3) basada en la credibilidad**, se estudian las noticias falsas basado en información relacionada con la noticia y el contexto social que la rodea (fuente de noticias [Ma et al., 2018], los usuarios [Zhang et al., 2020] y los difusores) y; **(4) basada en el estilo** donde se analiza la información del contenido de la noticia e incluye el estilo de escritura [Koumouridis, 2020].

Los estudios basados en el estilo, tienen como objetivo evaluar la intención de las noticias y guardan mucha relación con el análisis y la detección del engaño (centrado en el estilo general de contenido engañoso) [Zhou and Zafarani, 2020]. Nuestro enfoque estará encaminado hacia la detección de noticias falsas siguiendo una **perspectiva basada en estilo**.

La mayoría de estas perspectivas identifican las noticias falsas después de su propa-

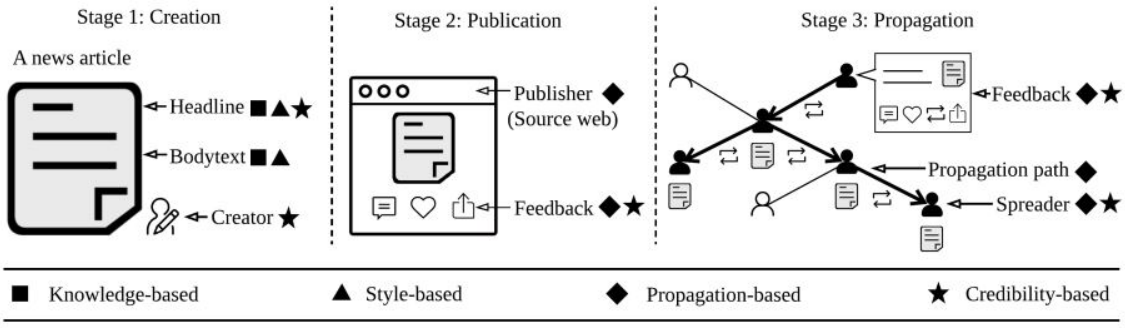

Figura 1.2: Ciclo de vida de la noticia y sus conexiones entre las cuatro perspectivas de estudio de la misma. Figura tomada de [Zhou and Zafarani, 2020].

gación. Por tanto, para la detección de noticias falsas, comprender el estilo de escritura de las noticias podría apoyar la tarea de detección temprana. Sin embargo, para detectar noticias falsas en una etapa temprana, es decir, cuando se publican en un medio de comunicación pero aún no se difunden en las redes, es de vital importancia desarrollar enfoques que puedan detectar noticias falsas centrándose en el contenido y estilo de escritura de las noticias.

1.2 Planteamiento del Problema

Los enfoques basados en estilo intentan detectar noticias falsas a través del estilo de la escritura de los contenidos de la noticia. Las características que han representado bien el estilo de las noticias falsas en el pasado, pueden no ser útiles en un futuro, debido a la constante evolución del estilo de escritura [Yang et al., 2018].

Para enfrentarnos a este desafío es importante la compatibilidad de las características, específicamente, características que pueden capturar la generalidad del estilo de escritura engañoso en temas, dominios y lenguajes, así como la evolución del estilo de escritura engañoso [Zhou and Zafarani, 2020]. De hecho, estas características lingüísticas de las noticias falsas son técnicamente muy exigentes para ser entendidas,

interpretadas, extraídas y analizadas [Shu et al., 2017], [Ruchansky et al., 2017] lo que supone, más que un desafío, una motivación.

En la detección automática de noticias falsas, desde una perspectiva basada en el estilo, el uso de las características lingüísticas ha sido explorado por diversos autores y se pueden clasificar en base a información léxica, sintáctica, morfológica y de legibilidad. Las características lingüísticas más comunes que se utilizan para representar información textual son características léxicas y morfológicas que incluyen características a nivel de palabra [Yang et al., 2018; Hardalov et al., 2016] y caracter [Potthast et al., 2018], como el número total de palabras y caracteres o la frecuencia de las palabras.

Las características sintácticas son otra categoría ampliamente conocida de características lingüísticas que busca cierta gramática dentro de la estructura de una oración [Yang et al., 2018; Potthast et al., 2018]. Algunos ejemplos son el etiquetado de partes del discurso (*part-of-speech*, POS).

Las características psicolingüísticas (o psicológicas) cuentan la proporción de palabras que se correlacionan con diversos procesos psicológicos y análisis de sentimientos básicos [Yang et al., 2018], [Hardalov et al., 2016]. Estas características generalmente se extraen con los diccionarios *Linguistic Inquiry and Word Count* (LIWC) que son básicamente grandes léxicos de categorías de palabras que representan emociones, procesos de percepción, etc.

El uso de la técnica de enmascaramiento del texto ha sido abordado en otras áreas de investigación (Agrupamiento de Texto [Granados et al., 2011], Atribución de Autoría [Stamatatos, 2017] [Sánchez-Junquera et al., 2020], Perfilado de Autor[Bacciu et al., 2019], [Jimenez-Villar et al., 2019], Detección de reclamos y Detección de engaño [Ghanem, 2018], [Ghanem et al., 2018], en la Detección de Hiperpartidismo en noticias [Sánchez-Junquera et al., 2019]) y ahora mucho más reciente en la detección de estereotipos sobre los inmigrantes [Sánchez-Junquera et al., 2021],sin embargo; no ha sido

abordado para la detección de noticias falsas, por lo que constituye una oportunidad de investigación.

En esta tesis nos encontramos con el problema de detección de noticias falsas utilizando solo datos textuales. Aunque este enfoque ha sido abordado en la literatura, la mayoría de los trabajos que lo abordan incorporan al análisis de texto otras características (metadatos, análisis de imágenes, uso del contexto en redes sociales, etc.) para poder elevar la exactitud de los resultados y los pocos que utilizan solo el enfoque textual no han obtenido del todo los resultados esperados.

Mediante técnicas de enmascarado del texto, ya sea del estilo o el contenido en las noticias, pretendemos encontrar patrones relevantes que nos permitan distinguir el engaño o la veracidad de la noticia con independencia del idioma (español e inglés), el dominio o el tema. Nuestra principal contribución computacional radica en definir qué se debe enmascarar y qué no entre las características antes mencionadas para discriminar entre noticias reales o falsas, para así obtener una mejor representación del texto. Siguiendo como principal motivación, el hecho de que el estilo de las noticias falsas apela a las emociones y creencias del lector y no a una argumentación objetiva. De ahí, que se espera una diferencia de estilo tanto en la forma de escritura, como en el uso de ciertos términos.

1.3 Pregunta de Investigación

Teniendo en cuenta la problemática definida en la sección anterior y rasgos característicos que definen las noticias falsas, tales como: suelen ser anónimas, no suelen tener fuentes para ser verificadas, apelan a la emoción, utilizan la propaganda en el cuerpo de la noticia [Koumouridis, 2020], el uso de títulos impactantes y alarmantes o en ocasiones sin títulos [Yang et al., 2018] y el propio uso del lenguaje empleado. Este proyecto de

tesis intentará responder la siguiente **pregunta de investigación:**

¿Pueden, las técnicas de enmascaramiento sobre el estilo o contenido de la información, discriminar entre noticias falsas y verdaderas?

1.4 Objetivos

Como **Objetivo General** nos planteamos: Detectar noticias falsas siguiendo un enfoque basado en estilo, mediante técnicas de enmascaramiento para el idioma inglés y español en notas periodísticas con resultados semejantes al estado del arte.

Para dar cumplimiento a nuestro objetivo general, presentamos los siguientes **objetivos específicos**:

1. Capturar características generales del estilo de las noticias que permitan distinguir entre noticias falsas y verdaderas.

2. Proponer representaciones que mejor discriminen el contenido falso del verdadero.

3. Analizar diferencias y puntos en común entre atributos específicos de estilo en las noticias, tanto para el idioma español como para el inglés.

1.5 Estructura de Tesis

La tesis está estructurada como sigue:

En el **Capítulo 2 Marco Teórico:** se presentan los conceptos para comprender el documento. Se exponen las técnicas de enmascaramiento utilizadas, la forma de representación de los textos, los algoritmos de aprendizaje utilizados y las medidas de evaluación. Todo el soporte teórico que se utilizará a lo largo de todo el documento.

Con el **Capítulo 3 Trabajos Relacionados:** se realiza un compendio de los trabajos pertenecientes al estado del arte en la detección de noticias falsas, haciendo mayor

referencia en aquellos que están más relacionados con el objetivo de la tesis.

Capítulo 4 Método Propuesto: este capítulo describe el método propuesto para obtener la implementación de la tesis y se describe cada una de las fases en dicho método: selección de los conjuntos de las palabras más frecuentes para definir los valores de k, selección del modelo (basado en estilo o contenido), los algoritmos de enmascaramiento definidos y por último antes de la clasificación la selección de características.

En el **Capítulo 5 Experimentos:** se describe en detalle todos los experimentos y los resultados obtenidos, se describen las colecciones utilizadas así como la configuración experimental y se realiza un pequeño resumen conclusivo de los resultados obtenidos y lo que podrían significar en el **Capítulo 6: Análisis de los Resultados**.

Por último el **Capítulo 7 Conclusiones y Trabajo Futuro:** concluye el trabajo de investigación realizado en este proyecto de tesis y sugiere trabajo futuro potencial que se puede explorar en el campo de la detección de noticias falsas.

Capítulo 2

Marco Teórico

En este capítulo vamos a describir los conceptos relacionados a la tarea de detección automática de noticias falsas mediante el enmascarado de la información textual; algoritmos de enmascaramiento utilizados, las distintas representaciones del texto, algoritmos de aprendizaje y medidas de evaluación.

2.1 Preprocesamiento

Los algoritmos de enmascaramiento son métodos para transformar los textos aplicando un proceso de distorsión que permita resaltar la información textual relevante a la tarea en cuestión. En nuestro caso, la idea principal es proporcionar una nueva versión del texto, conservando la mayor parte de la información relacionada con el estilo [Stamatatos, 2017] y distorsionando el resto. Estos métodos, seleccionan un conjunto de palabras, por ejemplo, las palabras más frecuentes del corpus, el conjunto W_k, y transforman el texto de entrada enmascarando aquellas palabras fuera de ese conjunto W_k. El enmascaramiento puede hacerse usando un asterisco por palabra o múltiples asteriscos conservando la longitud de la palabra original.

Los algoritmos tradicionales de enmascaramiento definidos por Stamatatos [2017] y que constituyen la base para los utilizados en nuestro trabajo los podemos ver representados a continuación.

El Algoritmo 2.1, Vista Distorsionada Múltiples Asteriscos, sustituye cada carácter del token por el símbolo de $(*)$ y cada dígito por el símbolo de $(\#)$, con esta técnica se mantiene la longitud de los tokens. A diferencia de este, con el Algoritmo 2.2, Vista

Algoritmo 2.1: DV-MA (Vista Distorsionada Múltiples Asteriscos)

Entrada: *Texto*, W_k
Salida : *Texto*

Tokenizar Texto
para cada *token t en Texto* **hacer**
  si *minúsculas*$(t) \notin W_k$ **entonces**
    remplazar cada dígito en t con #
    remplazar cada letra en t con ∗
  fin
fin

Distorsionada Simple Asterisco, se pierde totalmente este dato, dado que DV-SA sustituye todo el token por el símbolo de ($*$) y toda la secuencia de dígitos por un simple símbolo de (#).

Algoritmo 2.2: DV-SA (Vista Distorsionada Simple Asterisco)

Entrada: *Texto*, W_k
Salida : *Texto*

Tokenizar Texto
para cada *token t en Texto* **hacer**
  si *minúsculas*$(t) \notin W_k$ **entonces**
    remplazar cada secuencia de dígitos en t con un solo #
    remplazar cada secuencia de dígitos en t con un solo ∗
  fin
fin

2.2 Representaciones

En esta sección se describen las representaciones utilizadas en nuestra investigación y que han mostrado relevancia para la tarea de detección de noticias falsas, que son simples de obtener y procesar.

2.2.1 n-gramas

La definición más simple de un n-grama es la unión de uno o varios tokens, palabras o caracteres. La construcción de los n-gramas se lleva por medio de combinaciones entre

tokens o caracteres vecinos. Para llevar esto a cabo se crea una ventana del tamaño del n-grama deseado: si se quiere un unigrama el tamaño de la ventana será de 1, si son bigramas (o digramas) la ventana será de tamaño 2 y así sucesivamente hasta llegar al tamaño de n-grama final deseado. Esta ventana se mueve a través del texto y abarca la cantidad de tokens o caracteres que el tamaño de la ventana indica. Un ejemplo de formación de n-gramas de tokens es el siguiente (ver Tabla 2.1):

Frase: El Covid-19 no es un virus.

Tokens								
El	Covid	-	19	no	es	un	virus	.
Unigramas								
El	Covid	-	19	no	es	un	virus	.
Bigramas								
El Covid	Covid -	- 19	19 no	no es	es un	un virus	virus .	
Trigramas								
El covid -	-19 no	no es un	un virus .					

Tabla 2.1: Ejemplos de n-gramas de tokens.

Al igual que se extraen los n-gramas de tokens, estos pueden ser de caracteres, o sea, secuencias de n caracteres consecutivos. En este caso, dichas secuencias podrán estar conformadas con caracteres alfanuméricos, signos de puntuación, e incluso el espacio en blanco. La Tabla 2.2 muestra algunos ejemplos de este tipo de atributo. Aquí, para mayor claridad el espacio en blanco es sustituido por guión bajo.

Trigramas de: El covid-19 no es un virus.															
El_	l_c	_co	cov	ovi	vid	id-	d-1	-19	19_	9_n	_no	no_	o_e	_es	es_	....

Tabla 2.2: N-gramas de caracteres de la expresión: *El covid-19 no es un virus.*

2.2.2 *Word Embeddings* Pre-entrenados

Los modelos de *embeddings* de palabras previamente entrenados son la forma más sencilla de comenzar a trabajar con técnicas de *embeddings* de palabras [Kaliyar et al., 2020]. La principal ventaja de usar estos modelos es la capacidad de entrenar con conjuntos de datos masivos. Los *embeddings* generalmente representan codificaciones geométricas de palabras basados en la frecuencia con la que aparecen juntos en un corpus de texto. Al utilizar un modelo previamente entrenado, se puede reducir el consumo de tiempo para entrenar el modelo, limpiar, y procesar, grandes conjuntos de datos. Los modelos pre-entrenados se pueden clasificar en dos tipos de modelos, sin contexto y basados en el contexto. Los modelos libres de contexto, como *word2Vec* o *GloVe*, crean una representación solitaria de "*embeddings* de palabras"para cada palabra del vocabulario. Los modelos basados en el contexto crean una representación de cada palabra que depende de diferentes palabras en una oración.

GloVe es un algoritmo de aprendizaje no supervisado (Cerisara et al., 2018) que se utiliza para descubrir la cercanía de dos palabras, con su separación en un espacio vectorial. Estas representaciones vectoriales creadas se denominan vectores de *embeddings* de palabras. En *GloVe*, el entrenamiento se realiza en estadísticas globales de co-ocurrencia de palabras agregadas [Kaliyar et al., 2020] de un corpus. En el caso de *GloVe*, la matriz de recuento se procesa previamente normalizando los recuentos y suavizándolos logarítmicamente.

El objetivo de *GloVe* es sencillo, hacer cumplir los vectores de palabras para capturar relaciones sublineales en el espacio vectorial. *GloVe* da menor peso a los pares de palabras muy frecuentes para evitar las palabras vacías sin sentido como .[el]", ün", etc. Antes de entrenar el modelo real, se construye una matriz de co-ocurrencia X basada en palabras, donde una celda X_{ij} es un valor, que representa la frecuencia con la que la palabra i aparece en el contexto de la palabra j.

En este trabajo de investigación, hemos utilizado el paquete más pequeño de incrustaciones de palabras de 822 Mb, llamado "glove.6B.zip". Se entrenó sobre un conjunto de datos de mil millones de tokens (palabras) con un vocabulario de 400 mil palabras. Existen diferentes tamaños de vector de *embeddings*, con dimensiones de 50, 100, 200 y 300. Elegimos la versión de 300 dimensiones.

2.3 Algoritmos de Aprendizaje

En el campo de aprendizaje de máquina se establecen dos pilares principales llamados aprendizaje supervisado y aprendizaje no supervisado. Los algoritmos de aprendizaje supervisado basan su aprendizaje en un juego de datos de entrenamiento previamente etiquetados. Por etiquetado entendemos que para cada ocurrencia del juego de datos de entrenamiento conocemos el valor de su atributo objetivo. Esto le permitirá al algoritmo poder “aprender” una función capaz de predecir el atributo objetivo para un juego de datos nuevo. Las dos grandes familias de algoritmos supervisados son: los algoritmos de regresión cuando el resultado a predecir es un atributo numérico y los algoritmos de clasificación cuando el resultado a predecir es un atributo categórico. Algunos métodos y algoritmos que podemos implementar son los siguientes: K vecinos más cercanos (*K-nearest neighbors*), Redes neuronales artificiales (*Artificial neural networks*), Máquinas de Soporte Vectorial (*Support vector machines*), Clasificador Bayesiano ingenuo (*Naïve Bayes classifier*), Árboles de decisión (*Decision trees*) y Regresión logística (*Logistic regression*).

En nuestra investigación hacemos uso de las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) y redes neuronales.

2.3.1 Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (creadas por Vladimir Vapnik) constituyen un método basado en aprendizaje para la resolución de problemas de clasificación y regresión, incluidas aplicaciones médicas de procesamiento de señales, procesamiento del lenguaje natural y reconocimiento de imágenes y voz. El objetivo del algoritmo SVM es encontrar un hiperplano en un espacio N-dimensional (N-el número de características) que separe de la mejor forma posible dos clases diferentes de puntos de datos con un margen máximo.

Específicamente, SVM pertenecen a una clase de algoritmos de *Machine Learning* denominados métodos kernel, donde se puede utilizar una función de kernel para ajustar puntos de datos que no pueden separarse fácilmente o puntos de datos que son multidimensionales [Ahmad et al., 2020] . Existen modelos de SVM sigmoide, SVM polinomial, SVM gaussiano y SVM lineal básico.

Aunque los algoritmos SVM están formulados para la clasificación binaria, los algoritmos SVM multiclase se construyen combinando varios clasificadores binarios.

2.3.2 Redes Neuronales Convolucionales

Los modelos de aprendizaje profundo, específicamente en la tarea de detección de noticias falsas son bien conocidos por sus resultados de vanguardia en dicha clasificación. Las redes neuronales convolucionales (CNN) son una red neuronal artificial que se utiliza con frecuencia para diversas aplicaciones, como la clasificación de imágenes, el reconocimiento facial y el procesamiento del lenguaje natural [Krizhevsky et al., 2017]. En el campo del procesamiento del lenguaje natural, CNN exhibe un buen desempeño como red neuronal para la clasificación.

Las CNN constan de capas de entrada, ocultas y de salida. Las capas constan de

mapas de características y una capa completamente conectada con capas convolucionales y capas agrupadas. La capa convolucional y la capa de agrupación extraen las características de los valores de entrada y asignan los valores extraídos al mapa de características. En este proceso, las características de las oraciones se pueden extraer a través de la similitud semántica entre las palabras que constituyen la oración, y luego la capa completamente conectada tiene un valor de clasificación de las características extraídas para la clasificación [Jang et al., 2019].La capa completamente conectada finalmente envía el resultado a una capa de salida.

En la siguiente Figura 2.1, podemos ver el modelo general de una red convolucional. Nuestra fase inicial de experimentación con las redes utiliza una red convolucional sencilla con una sola convolución.

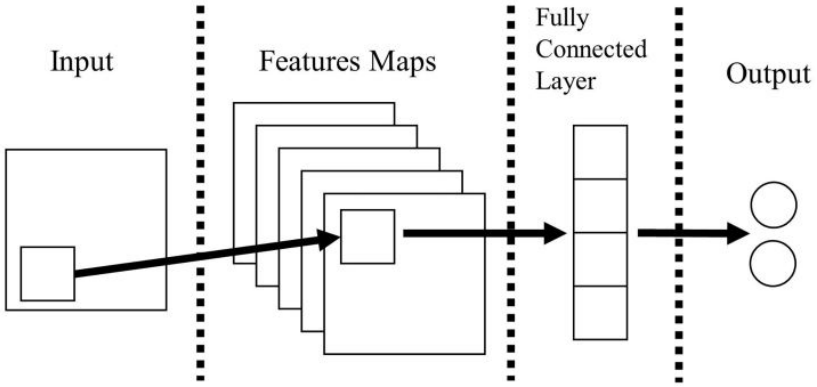

Figura 2.1: Modelo General CNN. [Jang et al., 2019]

2.4 Medidas de Evaluación

Para evaluar el desempeño de nuestro modelo propuesto hemos utilizado como métricas de evaluación: exactitud o *accuracy (Acc)*, precisión, recuerdo y la medida F1.

2.4.1 Exactitud

La Exactitud o *accuracy* (Acc) es una medida del número correcto de predicciones respecto al número total de predicciones en los datos [Sriram, 2020]. Es una manera de medir cuantas instancias desconocidas para el clasificador son etiquetadas por este de manera correcta. Puede calcularse como:

$$Acc = \frac{PrediccionesCorrectas}{PrediccionesTotales} \tag{2.1}$$

Mediante esta medida, es posible tener una idea de cuanto acierta el clasificador, independientemente de con que categoría acierta mas. Es decir, un clasificador entrenado con un corpus desbalanceado respecto a las clases, por ejemplo, muchos ejemplos de noticias verdaderas y muy pocas de noticias falsas podría predecir mas instancias como verdaderas y muy pocas como falsas, lo cual no se vería reflejado en el valor de Acc. Por tanto, dicha medida es adecuada principalmente cuando el corpus de entrenamiento esta balanceado.

2.4.2 Precisión

La precisión es otra manera de medir cuántas predicciones positivas tuvo el sistema, es la relación entre los verdaderos positivos y todos los positivos. Del número de instancias positivas que se predijeron correctamente, cuántas son realmente positivas:

$$Precisión = \frac{TP}{TP + FP} \tag{2.2}$$

Un sistema es preciso con respecto a una clase, siempre que las predicciones hechas hacia esa clase sean mayormente acertadas. Sin embargo, que la precisión del sistema sea alta respecto a una clase, no significa que reconozca bastantes instancias de la clase

correcta, sino que al predecir esa clase, se equivoca poco.

2.4.3 Recuerdo

También se conoce como tasa de verdaderos positivos. Es la medida de todas las instancias positivas que el modelo predijo correctamente [Sriram, 2020].

$$Recuerdo = \frac{TP}{TP + FN} \tag{2.3}$$

Con el valor del recuerdo y de la precisión, es posible determinar cuantas instancias de una clase es capaz de reconocer el sistema, y cuantas está reconociendo erróneamente. Por ejemplo, si el sistema predice que todas las instancias son de la clase falsa, cuando en realidad solo la mitad es de esta clase, el recuerdo sera máximo (i.e., 1), mientras que la precisión será de 0.5, es decir, el sistema podría predecir que todas las instancias son falsas, incluso aquellas que no lo sean.

2.4.4 Medida F1

La medida F1 es una media armónica entre recuerdo y precisión. Una puntuación F1 alta indica que el nivel de exactitud del modelo es muy bueno.

$$medida - F1 = \frac{2 * Precisión * Recuerdo}{Precisión + Recuerdo} \tag{2.4}$$

Cuando se desea obtener un solo valor de F1 aun teniendo dos clases, promediar la evaluación de cada clase puede ofrecer una alternativa. En este tesis se utiliza la medida F1-macro:

$$F1 - macro = \frac{F1(fake) + F1(True)}{2} \tag{2.5}$$

Con la cual se le da igual peso a ambas clases en caso de que el corpus este desbalanceado.

Capítulo 3

Trabajos Relacionados

En este trabajo de investigación ponemos especial énfasis en el paso del preprocesamiento del texto mediante el enmascarado de la información textual encaminado a la detección de noticias falsas, teniendo en cuenta el contenido y el estilo de los textos falsos y verdaderos para la clasificación.

En este capítulo pretendemos realizar un recorrido por los principales trabajos relacionados a la detección de noticias falsas prestando importancia a aquellos que han utilizado un enfoque estilístico en la selección de sus características, artículos que han utilizado SVM, n-gramas de palabras y caracteres, *embeddings* de palabras pre-entrenados y redes neuronales. También las tareas en las que se han utilizado los algoritmos de enmascaramiento y la posibilidad de investigación de esto en la detección de noticias falsas.

3.1 Detección de Noticias Falsas

Diversos son los autores que han abordado el problema de detección de noticias falsas, siguiendo varias líneas de investigación (estilo, contenido, propagación o híbridos). En la Tabla 3.1 se resumen aspectos que se consideraron importantes de cada uno de los trabajos relacionados en el estado del arte con respecto a la detección automática de noticias falsas utilizan algoritmos tradicionales del aprendizaje automático, los trabajos presentados se encuentran organizados por año, la mayoría de ellos abordan la detección de noticias falsas no solo utilizando características de estilo, sino también, incorporando otras características: de contenido, metadatos y el uso de recursos externos.

	Caracterización				Metadatos (imágenes, información de la red)	Método de Clasificación
	Estilo		Contenido	Recursos Externos		
	Superficiales	Profundas				
Ahmed et al., [2017]	n-gramas de palabras n=[1,2,3,4]	—	—	—	—	Stochastic Gradient Descent (SGD) Support Vector Machines (SVM) Linear Support Vector Machines (LSVM) K-Nearest Neighbour (KNN) Decision Trees (DT) Pesado TF y TF-IDF
Pérez-Rosas et al., [2017]	n-gramas de palabras n=[1,2,3,4] Métricas de legibilidad	Sintácticas	—	LIWC	—	Linear Support Vector Machines (LSVM) Pesado TF-IDF
Singh et al., [2017]	—	—	—	LIWC	—	Logistic Regression (LR) Support Vector Machines (SVM) Random Forest (RF) Decision Tree (DT) K-Nearest Neighbour (KNN)
Horne and Adali, [2017]	Características léxicas	Etiquetas POS. Complejidad a nivel de oración. Complejidad a nivel de palabra.	—	LIWC	—	Support Vector Machines (SVM)
Pisarevskaya [2017]	—	Etiquetas POS.	—	—	—	Support Vector Machines (SVM) Random Forest (RF)
Potthast et al., [2018]	BOW n-gramas de caracteres n=[1,2,3] n-gramas de palabras vacías, n=[1,2,3] Métricas de legibilidad.	n-gramas de etiquetas POS. n=[1,2,3]	—	General Inquirer Dictionaries	—	Random Forest (RF)
Zhou et al., [2019]	BOW	Etiquetas POS.	—	LIWC	—	Logistic Regression (LR) Naïve Bayes (NB) Support Vector Machine (SVM) Random Forests (RF) XGBoost
Reis et al., [2019]	BOW Métricas de legibilidad. Características Léxicas	Etiquetas POS	Puntuación de toxicidad obtenida Puntuaciones de subjetividad y sentimiento del texto.	LIWC	Imágenes Fuente de la noticia Entorno de la noticia: cantidad de likes, comentarios y vistas de los usuarios	K-Nearest Neighbour (KNN) Naïve Bayes (NB) Random Forest (RF) Logistic Regression (LR) Linear Support Vector Machines (LSVM) XGBoost
Castelo et al., [2019]	Morfológicas Métricas de legibilidad	—	—	LIWC	Web-markup	Support Vector Machine (SVM) K-Nearest Neighbors (KNN) Random Forest (RF)

Tabla 3.1: Tabla de comparación de los trabajos relacionados con la detección de noticias falsas.

Autores como Ahmed et al. [2017] se han centrado en la detección automática de contenido falso utilizando reseñas falsas en línea. Los autores también han explorado dos métodos diferentes de extracción de características para clasificar las noticias falsas. Utilizan técnicas de análisis de n-gramas de palabras con TF y con TF-IDF y aprendizaje automático. Investigaron y compararon dos técnicas de extracción de características diferentes y seis técnicas de clasificación de máquinas diferentes (*Stochastic Gradient Descent* (SGD), Máquinas de Soporte Vectorial (SVM), Máquinas de Soporte Vectorial Lineal (LSVM), K-vecinos más cercanos (KNN) y Árboles de Decisión (DT). La evaluación experimental produjo el mejor rendimiento utilizando el pesado TF-IDF y la máquina de soporte vectorial lineal (LSVM) como clasificador, con una exactitud del 92 % para unigramas de palabras. El conjunto de datos sobre el que se entrena y prueba el modelo propuesto, fue elaboración propia de los autores y consta de 12600 notas falsas y 12600 notas verdaderas (perfectamente balanceado). Un aspecto importante, es que cuando observamos los resultados obtenidos con el resto de los algoritmos de clasificación, observamos que la exactitud obtenida varía en un intervalo de 61 a 89 % respectivamente.

Al igual que Ahmed et al. [2017], los autores Pérez-Rosas et al. [2017] utilizan unigramas y bigramas de palabras para la identificación automática de contenido falso en artículos de noticias *online*, a través de un análisis exhaustivo para la identificación de rasgos lingüísticos en el contenido de las noticias falsas. Para su investigación, presentan dos nuevos conjuntos de datos que cubren siete dominios diferentes. Uno de los conjuntos de datos se recopila mediante una combinación de esfuerzos de anotación manual y colaborativa (40 noticias por cada uno de los seis dominios representados, para un total de 240 noticias verdaderas). A partir de estas notas verdaderas, fueron generadas versiones falsas a través de *Amazon Mechanical Turk*. El segundo conjunto de datos se recopila directamente de la web, dirigido principalmente a noticias sobre celebridades.

Para este caso, se coleccionaron 100 noticias falsas sobre celebridades y estas, de la misma forma que el anterior, fueron llevadas a notas verdaderas, para un total de 200 noticias. Para el primer conjunto de datos: *FakeNewsAMT*, se obtiene una exactitud de un 78 % a partir de las métricas de legibilidad usadas y para el segundo conjunto de datos: *Celebrity*, el modelo más exacto, con un 74 % de exactitud, se crea utilizando la combinación de todos los conjuntos de características lingüísticas extraídas.

El uso de otras características como los diccionarios psicolingüísticos como LIWC (Análisis lingüístico y recuento de palabras), también han sido utilizados para la tarea de detección de noticias falsas [Reis et al., 2019; Zhou et al., 2019; Castelo et al., 2019] de conjunto con otros tipos de características. Sin embargo, autores como por ejemplo, Singh et al. [2017] utilizan solo las características obtenidas con LIWC para cada uno de los artículos (el recuento de palabras, la autenticidad, la influencia, el tono y la analítica). En este trabajo, los autores mezclan dos conjuntos de datos: Kaggle Fake News, para las noticias falsas, y un nuevo conjunto de datos de elaboración propia con 345 artículos de noticias verdaderas, procedentes de agencias de noticias reconocidas (ej. New York Time). Utilizan métodos tradicionales de aprendizaje automático para clasificar noticias falsas y obtuvieron una exactitud del 87 % con SVM como clasificador.

Existen otros trabajos de investigación en la literatura consultada que han seguido un enfoque puramente lingüístico para la detección de noticias falsas. Los autores en [Horne and Adali, 2017] trabajaron el problema de detección de noticias falsas utilizando una extensa colección de señales lingüísticas, (63 en total) agrupadas en tres amplias categorías: complejidad, psicología (LIWC) y características léxicas. El algoritmo de aprendizaje automático que usaron fue un clasificador SVM con núcleo lineal en el conjunto de datos de Buzzfeed y la puntuación obtenida fue del 77 % de exactitud promedio con validación cruzada de 5 pliegues.

Siguiendo este enfoque puramente lingüístico, Pisarevskaya [2017] se propuso detec-

tar engaño en reportes de noticias en el idioma ruso. El conjunto de datos, de elaboración propia utilizado, recopila un total de 174 textos y 48 diferentes temas y revela diferencias entre informes de noticias falsas y verdaderas teniendo en cuenta la estructura de la noticia. Pisarevskaya realizó dos experimentos; el primero basado en marcadores a nivel léxico (Etiquetas POS) y el segundo a nivel del discurso mediante la teoría de estructuras retóricas (RST). Los mejores resultados se obtuvieron con las máquinas de soporte vectorial, una validación cruzada de 10 pliegues y como características, los marcadores léxicos, alcanzando una medida F1 de 0.65 y una exactitud de 0.64.

Potthast et al. [2018], en su investigación, entre otros aspectos, buscaban dar respuesta sobre si, el simple uso del estilo podía discriminar correctamente las notas falsas de las verdaderas. Para ello, los autores se centran en la detección de noticias falsas pero con una particularidad: el estilo de redacción de noticias falsas en relación con el estilo de noticias hiperpartidistas. Analizaron varios enfoques: primero, si las noticias hiperpartidistas se podían distinguir por su estilo de las noticias convencionales (se alcanzó un $F1 = 0{,}78$), segundo, si la sátira se puede distinguir de ambas (se alcanzó un $F1 = 0{,}81$) y **si las noticias falsas se pueden detectar a través de solo el estilo** (se alcanzó un $F1 = 0{,}46$ para la clase falsa y una exactitud de 55 % para todo). Para los autores, su modelo de clasificación de noticias falsas basado en el estilo de escritura no les funcionó bien, no obstante, proponen en trabajos futuros, volver a probar este modelo pero en conjuntos de datos más variables, sobre todo en dominios, dado que el conjunto de datos utilizado para entrenamiento y prueba (BuzzFeed) es puramente político.

Un modelo basado en estilo, ayuda sobre todo a la detección temprana de las noticias falsas, cuando aun no han sido extendidas en la red. Para predecir noticias falsas antes de que comiencen a propagarse en las redes sociales, este trabajo [Zhou et al., 2019], realiza un estudio interdisciplinario: representa el contenido de las noticias capturando

su estilo de escritura respectivamente a nivel léxico (BOW), nivel de sintaxis (POS), nivel semántico (LIWC) y nivel del discurso. Los resultados experimentales arrojaron que utilizando XGBoost con todas las características a nivel del discurso, podían obtener para el conjunto de datos PolitiFact una exactitud y medida F1 de un 89 % respectivamente y para BuzzFeed un 87.9 % respectivamente para ambas medidas, superando todas las líneas bases que incluyen contenido, propagación e híbrido (contenido + propagación). Un dato muy importante a expresar es que para este estudio los conjuntos de datos utilizados están equilibrados con un 50 % de noticias verdaderas y un 50 % de noticias falsas.

Un gran número de trabajos exploran no solo estas características estilométricas antes mencionadas en los trabajos previos, sino que investigan otros varios tipos de características también extraídas de las noticias, incluida la fuente y las publicaciones de las redes sociales. En el siguiente trabajo, [Reis et al., 2019], los autores, además de explorar las principales características propuestas en la literatura para la detección de noticias falsas (sintácticas, léxicas, psicolingüísticas y semánticas), profundizan en la búsqueda de una amplia variedad de características, publicaciones e historias que pueden ayudar a predecir noticias falsas con mayor precisión: el sesgo, la confiabilidad, el compromiso, la ubicación del dominio y los patrones temporales. Para su investigación, utilizaron un conjunto de datos que contenía 2282 artículos de noticias (buzzfeed) y una amplia variedad de algoritmos de clasificación: KNN, Naïve Bayes, Random Forest, Support Vector Machine y el algoritmo XGBoost. Descubrieron que XGBoost funciona mejor que todos con una puntuación macro F1 de 0.81 y el uso de todas las características antes mencionadas.

Por su parte, Castelo et al. [2019], proponen una clasificación independiente del tema *topic-agnostic*, (TAG), estrategia que utiliza características lingüísticas y de *web-markup* para identificar noticias falsas. Informan resultados experimentales utilizando

múltiples conjuntos de datos que muestran que su enfoque alcanza una alta exactitud en la identificación de noticias falsas, incluso cuando los temas de las noticias, según los autores evolucionan con el tiempo. Reportaron resultados que mostraron que su enfoque fue efectivo, obteniendo con máquinas de soporte vectorial y el uso de todas las características: para el conjunto de datos *Celebrity* una exactitud de un 78 %, para *US-Election2016* una exactitud del 86 % y para *PoliticalNews*, conjunto de datos de elaboración propia, una exactitud del 83 %. Proponen añadir al modelo características adicionales, por ejemplo, la participación del usuario y la estructura de la red para elevar la exactitud obtenida.

Dado que el NLP tiene que ver con datos textuales, es razonable explotar las características lingüísticas que pueden capturar los diferentes estilos de escritura de un escritor o un grupo de documentos. Estas características se derivan del contenido del texto de diferentes niveles de la organización de un documento, como palabras, oraciones, caracteres especiales e incluso todo el documento [Koumouridis, 2020].

Como se mencionó en el capítulo **1 Introducción**, existen cuatro perspectivas de estudio de las noticias falsas que están muy relacionadas con el ciclo de vida de la misma [Zhou et al., 2019]: basada en el conocimiento, basada en el estilo, basada en la propagación y basada en la credibilidad. El enfoque de nuestra tesis está encaminado a seguir el estudio de la detección de noticias falsas basado en el estilo.

La mayoría de los trabajos que abordan el estudio de detección de noticias falsas siguiendo un enfoque basado en estilo (estilo de escritura, características lingüísticas), incorporan al análisis de texto otras características (características de contenido, recursos externos, metadatos: análisis de imágenes, uso del contexto en redes sociales, etc.) para poder elevar la exactitud de los resultados y los pocos que utilizan solo el enfoque de una manera más limpia no han obtenido del todo los resultados esperados.

Nuestro trabajo busca determinar el alcance del simple uso de datos textuales,

enfocándose principalmente al estilo con que se escriben las noticias falsas.

3.2 Redes Neuronales para la Detección de Noticias Falsas

Los modelos basados en redes neuronales profundas se han vuelto tendencia en la detección de noticias falsas. Muchos estudios han utilizado algoritmos de aprendizaje automático y han creado clasificadores basados en características como el contenido, el nombre del autor y el título del trabajo, utilizando muchos modelos como la red neuronal convolucional (CNN), la red neuronal recurrente (RNN), y la red neuronal de memoria a largo-corto plazo (LSTM) para encontrar un modelo óptimo y reportar sus resultados. En esta sección, discutimos los métodos de aprendizaje profundo que se han utilizado en los trabajos existentes para la detección de noticias falsas. En particular, los métodos CNN, RNN y LSTM. La siguiente Tabla 3.2, resume los aspectos fundamentales descritos en la bibliografía consultada.

	Entrada			**Metadatos (imágenes, información de la red)**	**Arquitectura Red Neuronal**
	Característica	**Dimensión**	**Pre-entrenado**		
Wang, [2017]	embeddings de palabras	300	Word2Vec	Tema Perfil del difusor Contexto de la noticia...	CNN + Bi-LSTM
Long et al. [2017]	embeddings de palabras skip-gram	200	—	Perfil del difusor	LSTM + Attention
Girgis et. al. [2018]	embeddings de palabras	—	—	—	GRU LSTM Vanilla RNN
Roy et al. [2018]	embeddings de palabras	300	Word2Vec	Tema Perfil del difusor Contexto de la noticia...	Ensamble: Bi-LSTM CNN
O'Brien et al., [2018]	embeddings de palabras	1000	Word2Vec	—	CNN
Trueman et al. [2021]	embeddings de palabras	100	GloVe	—	CNN + BiLSTM + Attention

Tabla 3.2: Tabla de Comparación de los trabajos relacionados con la detección de noticias falsas utilizando redes neuronales profundas.

Wang [2017] ha presentado un nuevo conjunto de datos (LIAR) en el dominio político, para ayudar en la tarea de detección automática de noticias falsas. Han propuesto una arquitectura híbrida para resolver el problema de las noticias falsas; una es una red neuronal convolucional para el aprendizaje de representación de metadatos, seguida de una red neuronal de memoria a corto-largo plazo (LSTM). El modelo es complicado con muchos parámetros para optimizar, su modelo propuesto, dado la naturaleza del conjunto de datos multiclase (6 clases) y la corta extensión del texto de la noticia, no funciona bien, la mejor exactitud obtenida es de un 27.4 % utilizando las características textuales y todos los metadatos analizados.

Siguiendo los pasos de Wang [2017] al utilizar el conjunto de datos LIAR, pero esta vez, utilizando un modelo híbrido de LSTM con modelos de atención, Long et al. [2017] propuso integrar el perfil del difusor (afiliación partidista, ubicación del orador, puesto de trabajo del orador e historial crediticio) para detectar noticias falsas. Específicamente, los autores incluyeron los perfiles de difusor de dos maneras: en el modelo de atención y el otro como datos de entrada adicionales. Su estudio indicó que el modelo propuesto mejora la exactitud en un 14,5 % para el conjunto de datos de noticias falsas con una exactitud de 41.5 %, comparado con el estado del arte propuesto por Wang [2017] en su investigación.

Girgis et al. [2018] han presentado un trabajo para identificar noticias falsas utilizando, de igual forma que los trabajos previos, el conjunto de datos LIAR. Han obtenido una precisión del 21,7 %, 21,66 % y 21,5 % con los modelos GRU, LSTM y Vanilla, respectivamente. Este trabajo, se compara como baseline, con el modelo implementado por Wang [2017] sobre el mismo conjunto de datos utilizando solo datos textuales y no superan el 27 % obtenido con un modelo CNN. La entrada utilizada para alimentar las redes constituyeron *embeddings* de palabras entrenados sobre el propio conjunto de datos, pero desconocemos las dimensiones de los mismos ya que el dato es omitido por

los autores.

Otro enfoque aplicado de igual forma que los trabajos anteriores, sobre el conjunto de datos LIAR fue propuesto por Roy et al. [2018] que intentaron desarrollar una arquitectura basada en modelos de ensambles. Los modelos individuales, CNN y BiLSTM, sobre el conjunto de datos obtuvieron una exactitud de 42,89 % y 42,65 %, respectivamente utilizando los mismos metadatos que Wang [2017]. Los autores asignaron la representación obtenida de CNN y BiLSTM a un modelo MLP (perceptrón multicapa) para obtener la exactitud de clasificación final (44,87 %) utilizando la totalidad de los metadatos relacionados con el tema, el perfil del difusor, contexto de la noticia, etc.

O'Brien et al. [2018] utilizan en su investigación otro conjunto de datos diferente al utilizado hasta el momento, las notas falsas las extrae de Kaggle y las notas verdaderas constituyen artículos de noticias tomados del *New York Times*, este conjunto de datos a diferencia del LIAR es binario y no multiclase. Estos autores no utilizaron ningún tipo de metadatos en su investigación, sin embargo, demostraron que las redes neuronales convolucionales pueden ser una herramienta poderosa para detectar noticias falsas en temas novedosos, únicamente desde el lenguaje. En su estudio, lograron una exactitud del 87,7 % y abrieron la caja negra de los detectores de redes neuronales examinando las palabras de los artículos de entrada que fueron más relevantes para la clasificación.

En este artículo, los autores, Trueman et al. [2021], proponen una red convolucional de memoria a largo y corto plazo basada en la atención para detectar automáticamente noticias falsas utilizando el conjunto de datos LIAR. Primero, preprocesaron el conjunto de datos usando la conversión de mayúsculas y minúsculas, eliminando símbolos y puntuación, y utilizando técnicas de tokenización. En segundo lugar, generaron vectores de palabras para los datos de entrada utilizando los *embeddings* de palabras previamente entrenadas por GloVe. Finalmente, el modelo híbrido AC-BiLSTM propuesto se emplea en estos vectores de palabras para predecir noticias falsas en un entorno de clases

múltiples obteniendo una exactitud de 35.1 % y una medida micro F1 de 39 %.

Todos los trabajos estudiados que utilizan un enfoque de aprendizaje profundo alimentan la red neuronal con *embeddings* ya sean pre-entrenados o no y complementan su método con la incorporación de la información adicional que proporcionan los metadatos. En nuestro caso muy particular alimentamos la red con texto enmascarado y no utilizamos ningun tipo de información adicional fuera de la pura información textual.

3.3 Técnica de Enmascaramiento del Texto

Las técnicas de enmascaramiento (o distorsión) del texto, fueron introducidas primeramente por Granados et al. [2011] para el problema específico de la agrupación de texto basada en comprensión. El autor plantea en su documento que el eliminar las *stop-words* ayudaría en la clasificación y para ello realiza este preprocesamiento utilizando una lista genérica de palabras vacías. Estas palabras que deben ser eliminadas, son las que el autor enmascara mediante caracteres aleatorios y sustitución por asteriscos. Luego, el uso del concepto de enmascaramiento del texto ha sido abordado en otras áreas de investigación.

Fue Stamatatos [2017] en la tarea de Atribución de Autoría quien presenta un método ya elaborado que mejora la eficacia de la atribución de autoría al introducir un paso de distorsión del texto antes de extraer medidas estilométricas. El método propuesto intenta enmascarar información específica de un tema que no está relacionada con el estilo personal de los autores. Dadas las k palabras más frecuentes del lenguaje, el autor enmascara todos los *tokens* del texto que no pertenecen al conjunto de estas k palabras. Utilizan los n-gramas de tokens y caracteres y SVM como clasificador. Demostraron que el enfoque propuesto puede mejorar los métodos existentes en la tarea de Atribución de Autoría. Stamatatos introduce dos algoritmos de enmascaramiento:

DV-MA o *Distorted View Multiple Asterisks*, que no es más que sustituir cada caracter de la palabra o token por un asterisco y cada dígito de un número por el símbolo # y DV-SA o *Distorted View Simple Asterisks* donde se sustituye la ocurrencia total de la palabra o token por un simple asterisco y cualquier número por un solo símbolo #.

Cuando utilizamos los algoritmos de enmascaramiento, no solo nos referimos a la técnica utilizada DV-MA o DV-SA (métodos tradicionales) propuestos por Stamatatos [2017] sino que dependemos de una lista de k palabras que puede ser o un recurso externo o determinada a partir del vocabulario del propio conjunto de datos en uso. A partir de esta lista decidiremos si enmascaramos estilo (enmascaramos todas las ocurrencias en el texto de las palabras que pertenecen a la lista de las k palabras, dejando intacto el resto) o contenido (enmascaramos el resto de las palabras en el texto que no pertenecen a mi lista de k palabras). La Tabla 3.3 muestra un resumen de los diferentes trabajos relacionados con el uso de las técnicas de enmascaramiento utilizadas para las diferentes tareas y muestra en que posición nos encontramos con nuestro método y en que nos diferenciamos del resto, esto último será explicado al final de la sección.

Un año más tarde el propio Stamatatos [2018], reutiliza las técnicas de enmascaramiento, nuevamente en la tarea de atribución de autoría, esta vez examina un conjunto más rico de técnicas de distorsión de texto. DV-EX que se inspira en estudios psicológicos que indican que las letras exteriores son más importantes que las letras interiores en la lectura de oraciones y DV-L2 es un intento de mantener sufijos de palabras que generalmente indican información morfosintáctica (por ejemplo, tiempo, número, parte del habla, etc.). Los autores se enfocaron esta vez en la atribución de temas cruzados donde el corpus de entrenamiento comprende textos sobre varios temas (en lugar de un área temática general única).

Ghanem et al. [2018] proponen un enfoque que utiliza una técnica de distorsión de texto para detectar afirmaciones que son dignas de verificación en debates presidencia-

	Tarea	**Técnica de enmascaramiento**	**Lista de K palabras**		**Que se enmascara**	**Método de Clasificación**
			Recurso Externo	**Recurso Interno**		
Stamatatos [2017]	Atribución de Autoría	Método Tradicional (DV-MA, DV-SA)	Palabras más frecuentes del idioma inglés (BNC corpus)	—	Contenido	Support Vector Machine (SVM)
Stamatatos [2018]	Atribución de Autoría	Método Tradicional (DV-MA, DV-SA, DV-EX, DV-L2)	Palabras más frecuentes del idioma inglés (BNC corpus)	—	Contenido	Support Vector Machine (SVM)
Ghanem et al. [2018]	Afirmaciones verificables en debates políticos	Método Modificado (DV-MA)	—	Frecuencia de los términos en el documento (más frecuentes)	Estilo	K-Nearest Neighbors(KNN)
Sánchez-Junquera et al. [2019]	Detección de Engaño	Método Modificado (DV-MA, DV-SA)	—	Frequently Co-occurring Entropy (FCE)	Estilo y Contenido	Naive Bayes (NB)
Jimenez-Villar et al. [2019]	Perfilado de Autor	Método Tradicional (DV-MA)	—	* Frecuencia de los términos en el documento (DF) * Frequency Co-occurring Entropy (FCE) * Ganancia de Información (IG)	Contenido	Support Vector Machine (SVM)
Sánchez-Junquera et al. [2019]	Detección de Hiperpartidismo en noticias	Método Modificado (DV-SA)	Palabras más frecuentes del idioma inglés (BNC corpus)	—	Estilo y Contenido	Naive Bayes (NB) Support Vector Machine (SVM) Random Forest (RF)
Sánchez-Junquera et al. [2021]	Identificación de estereotipos sobre inmigrantes	Método Modificado (DV-MA)	—	* Lista de palabras con mayor Frecuencia Relativa * Palabras con mayor Frecuencia Absoluta	Contenido	Logistic Regression (LR)
Nuestro Modelo	**Detección Automática de Noticias Falsas**	**Método Modificado (DV-MA, DV-SA)**	**Palabras más frecuentes del idioma inglés (BNC corpus)**	*** Palabras con mayor Frecuencia Absoluta** *** Frequency Co-occurring Entropy (FCE)**	**Estilo y Contenido**	**Support Vector Machine (SVM)** **Convolutional Neural Network (CNN)**

Tabla 3.3: Tabla de Comparación de los trabajos relacionados con la técnica de enmascaramiento.

les. Consideraron que este tipo de tarea era más temática que estilística, donde el estilo de escritura no es tan importante como las palabras temáticas. En su enfoque, utilizaron la técnica de distorsión de texto para detectar afirmaciones valiosas, ocultaron palabras que tienen una alta frecuencia en los documentos y mantuvieron (resaltando) otras palabras claves que se usan más en afirmaciones fácticas (entidades nombradas y claves lingüísticas). Pudieron concluir que el método de distorsión de texto funcionó mejor que usar el texto completo en el proceso de clasificación. Mejoraron los resultados en comparación con la línea de base con el método BOW normal.

De igual forma Sánchez-Junquera et al. [2019], en su investigación, utiliza un enfoque de adaptación de dominio para la detección de engaño entre dominios en textos. Su propuesta consiste en modificar los textos originales de los dominios de origen y destino en una forma en la que se mantenga la información de estilo pero se enmascare la información específica del dominio. Sus experimentos demuestran que la técnica de enmascaramiento es una buena idea para detectar engaños en escenarios de dominios cruzados. Los autores realizan ciertas modificaciones a las técnicas tradicionales propuestas por Stamatatos [2017], si el token es independiente del dominio (por ejemplo, comas y puntos), se mantiene. Por otro lado, si se encuentra que es específico del dominio (por ejemplo, comillas, paréntesis o términos compuestos como y/o), se reemplaza por el símbolo @. Además, para considerar todos los detalles numéricos que suelen dar los comunicadores veraces, enmascararon los números (por ejemplo, uno, dos, tres, etc.) con un solo símbolo +.

Para la tarea de Perfilado de Autor (*author profiling*) Jimenez-Villar et al. [2019] aplicaron algunas técnicas de enmascaramiento que les permitieron enfatizar los términos relevantes ofuscando los irrelevantes pero manteniendo información sobre la longitud de los textos. Encontraron que el enmascaramiento utilizando múltiples asteriscos como lo propuso Stamatatos [2017] era el más adecuado para la tarea de perfilado de

autor. Una conclusión interesante arrojó que enmascarar todas las palabras produjo una exactitud muy competitiva, lo que sugiere la importancia de la longitud del texto para distinguir a los *bots* de los humanos.

Un poco más reciente, Sánchez-Junquera et al. [2021] vuelve a retomar el enmascaramiento pero esta vez en la identificación de estereotipos sobre inmigrantes utilizando dos enfoques diametralmente opuestos: BETO, un modelo de aprendizaje profundo basado en *Transformers*; y utilizando la técnica de enmascaramiento de texto que ha sido reconocida por su capacidad para ofrecer buenos resultados a la vez que comprensibles para los humanos. Abordaron dos tareas de clasificación, la detección de estereotipo frente a la detección de no estereotipo y la identificación de las dimensiones de víctimas frente a amenazas mediante un conjunto de datos anotado en Español. Demostraron que ambos enfoques son adecuados para la identificación de estereotipos de inmigrantes; y curiosamente, la técnica de enmascaramiento logra casi los mismos resultados que BETO, a pesar de su simplicidad.

En el ámbito de las noticias, las técnicas de enmascaramiento del texto no han sido muy utilizadas. Sánchez-Junquera et al. [2019] utiliza estas técnicas del enmascarado del texto en la Detección de Hiperpartidismo en noticias, que les permitió evaluar el papel del estilo frente al contenido para la tarea en cuestión, enmascarando las k palabras más frecuentes o enmascarando el resto del texto y dejando intactas estas k palabras. Aunque los autores utilizan un conjunto de datos de noticias falsas, la tarea en cuestión era la detección de el hiperpartidismo en las noticias.

Cuando realizamos el recorrido por los diferentes trabajos del estado del arte que utilizan la técnica de enmascaramiento del texto, podemos decir, que no ha sido abordada para la tarea de detección automática de de noticias falsas; por lo que constituye, en este caso una oportunidad de investigación. Por otra parte los diferentes trabajos utilizan a la hora de definir la lista de las k palabras indistintamente o recursos externos

o internos solamente, en nuestro caso utilizamos ambos para así poder diferenciar cual es más ventajoso y entender el por qué. Por último ninguno autor utiliza en sus métodos de clasificación las redes neuronales, nuestro trabajo alimenta una red convolucional con texto enmascarado y los resultados obtenidos como se verá en próximos capítulos son bastante alentadores.

3.4 Brechas de Investigación

Al realizar este recorrido por los trabajos relacionados con la tarea de detección de noticias falsas; pudimos apreciar que existe un amplio mundo de investigación para la detección de noticias falsas, tanto en notas periodísticas (nuestra investigación se enfoca solo en conjuntos de datos de notas periodísticas) como en redes sociales, sin embargo se han reportado pocos trabajos enfocados en el estilo de las noticias falsas solo trabajando con información textual y sobre todo utilizando las técnicas del texto enmascarado. Esta brecha en la investigación es la que utilizamos como principal motivación para el desarrollo de nuestra tesis de maestría, consideramos que son aspectos que podemos explotar y obtener resultados mejores o similares a los vistos durante este capítulo en el estado del arte para la tarea de detección de noticias falsas en notas periodísticas.

Capítulo 4

Método Propuesto

La idea principal con esta tesis es proponer un método para la detección de noticias falsas mediante el enmascarado de los textos de entrada, de tal manera que se mantenga la estructura textual, relacionada con el estilo de las noticias, mientras enmascaramos las ocurrencias de las palabras menos frecuentes, correspondientes a información temática. Para efectos de evaluación y comparación también se considera el escenario inverso, se enmascaran las ocurrencias de las palabras más frecuentes y se conservan las palabras con información temática, relacionadas al contenido de la noticia. En el presente capítulo se describe el método propuesto. De manera general, como se muestra en la Figura 4.1, tenemos cuatro etapas fundamentales, cada una de estas etapas se describen en cada una de las siguientes secciones de este capítulo:

(Etapa 1): Selección del tipo de enfoque que vamos a utilizar, de esta etapa se derivan las k palabras más frecuentes que posteriormente utilizaremos para el enmascaramiento.

(Etapa 2): Selección del modelo: si queremos enmascarar siguiendo un modelo basado en estilo o un modelo basado en contenido.

(Etapa 3): Algoritmos de enmascaramiento, utilizaremos una vista distorsionada con múltiples asteriscos o con simples asteriscos.

(Etapa 4): Selección de las características (n-gramas de palabras o caracteres), que pasaremos luego al clasificador.

En la Figura 4.1, antes referida, también se puede contemplar una serie de cuestiones que nos preguntamos en cada una de estas etapas, y que son la base fundamental del por qué de esta experimentación, que se pretende lograr con cada una de estas etapas y que nos permitirán llegar a conclusiones referentes al método utilizado para la detección

de noticias falsas mediante el enmascaramiento de la información textual aportando así a nuestra contribución.

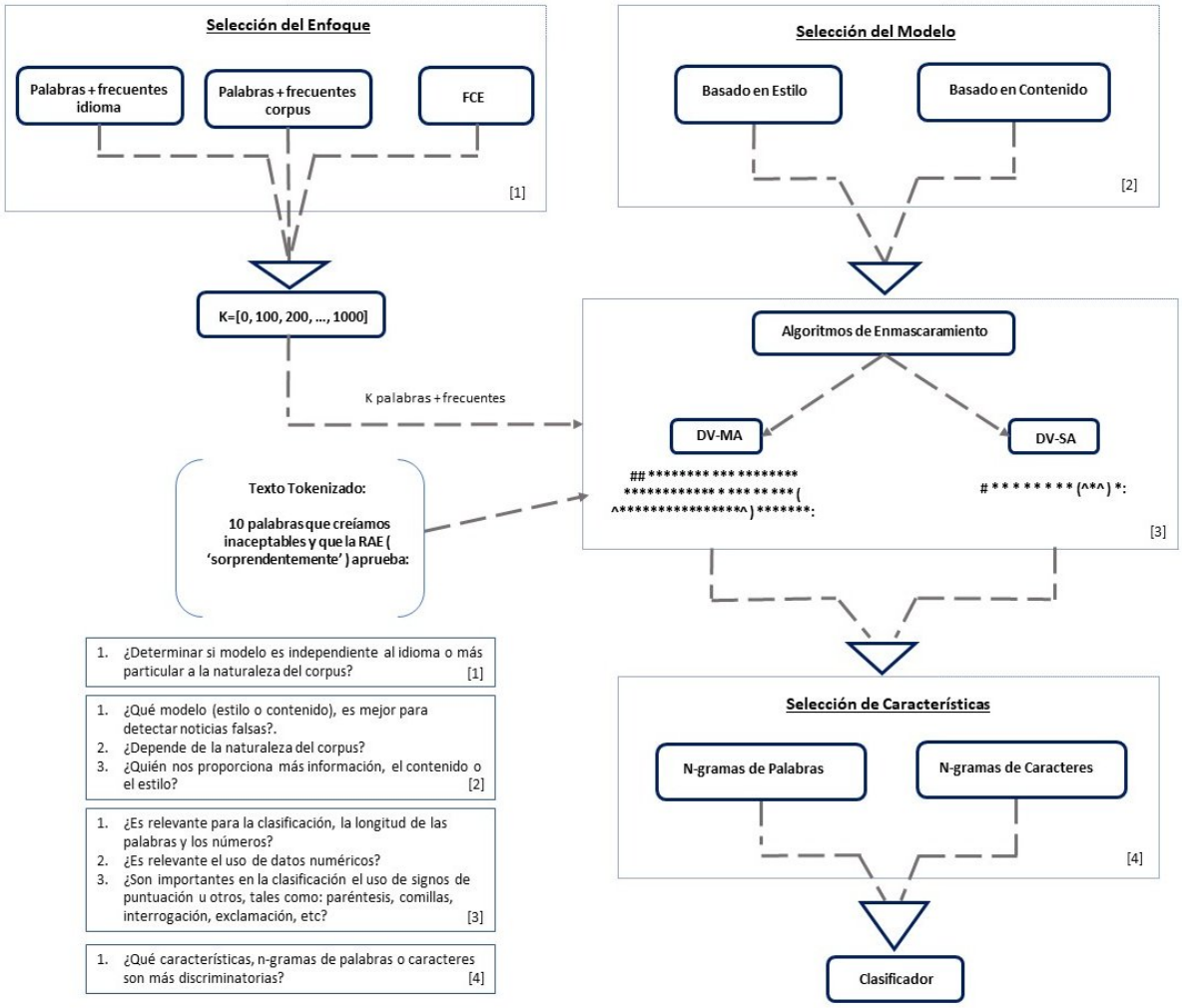

Figura 4.1: Esquema General del Método Propuesto. Se muestran cuatro etapas fundamentales y las principales interrogantes que queremos resolver en cada una de ellas.

4.1 Selección del Enfoque. Palabras más Frecuentes

En esta etapa se usan diferentes métodos para identificar las palabras asociadas al estilo de las notas periodísticas. El objetivo es la selección de un conjunto de k-palabras que utilizaremos en el proceso de enmascarado. Se exploraron tres diferentes

enfoques: primeramente mediante un recurso externo, segundo utilizando la frecuencia absoluta de las palabras del vocabulario del corpus, y tercero mediante el cálculo de la *Frequently Co-occurring Entropy* (FCE). Con estos diferentes enfoques queremos comprobar: primeramente, si nuestro modelo es independiente al idioma y segundo si es un poco más específico y depende de la naturaleza propia de cada uno de los conjuntos de datos.

Estos términos, los denominados generales (o más frecuentes), tienen probabilidades de ocurrencia relativamente más altas y similares tanto en las notas falsas como en las verdaderas y son menos independientes al contenido de cada dominio. Dentro de estos términos generales, podemos encontrar patrones asociados al estilo, a un nivel superficial de análisis. Por ejemplo: las *stopwords* (preposiciones, conjunciones, artículos, algunos adverbios) conjunciones verbales, etc. A continuación se detallan estos tres enfoques.

4.1.1 Recurso Externo: Palabras más Frecuentes del Idioma

Para este enfoque se utilizaron las listas de palabras más frecuentes del idioma. Para determinar las frecuencias de las palabras se utilizaron corpus representativos del idioma español e inglés. Idiomas en los cuales se experimentó nuestra propuesta.

Las palabras más frecuentes del idioma español fueron obtenidas del: "Corpus de Referencia del Español actual" (CREA)[1] y las palabras más frecuentes del idioma inglés fueron obtenidas del: *"British National Corpus"* (BNC)[2].

El Corpus de Referencia del Español Actual (CREA) es un conjunto de textos de diversa procedencia, del que es posible extraer información para estudiar las palabras, sus significados y sus contextos. Atendiendo a este criterio, el CREA cuenta, con algo más de ciento sesenta millones de formas léxicas. Se compone de una amplia variedad

[1]https://www.rae.es/banco-de-datos/crea
[2]https://www.english-corpora.org/bnc/

de textos escritos y orales, producidos en todos los países de habla hispana desde 1975 hasta 2004. Los textos escritos, seleccionados tanto de libros, como de periódicos y revistas, abarcan más de 100 materias distintas.

Además, desde su aparición, el CREA ha sido el punto de partida forzoso para investigaciones sobre el español actual, principalmente lingüísticas, pero también de campos tan dispares como el de la publicidad, la terminología o la sociología, así como para la elaboración de numerosos productos derivados: gramáticas, diccionarios, tesauros, correctores ortográficos, métodos de didáctica del español y aplicaciones de traducción automática, entre otros.

Por su parte, el *British National Corpus* (BNC) fue creado originalmente por la prensa de la Universidad de Oxford en la década de 1980 y principios de la de 1990, y contiene 100 millones de palabras de textos de una amplia gama de géneros (por ejemplo, ficción, revistas, periódicos y académico).

4.1.2 Palabras más Frecuentes del Corpus

Este segundo enfoque se limita al propio vocabulario del conjunto de datos utilizado y las frecuencias absolutas de cada uno de estos términos en el propio conjunto de datos. Se obtuvo la frecuencia de todas las palabras y se ordenaron de forma descendente por frecuencia.

4.1.3 *Frequently Co-occurring Entropy* (FCE)

A diferencia de los enfoques discutidos anteriormente que suelen ser un poco más descriptivos, el cálculo del FCE es un poco más discriminativo. Para nuestro método intentamos contrastar dos conjuntos de documentos, que podrían ser las dos clases (falsa y verdadera) o podrían ser dos conjuntos de datos de diferentes dominios.

FCE garantiza que las palabras más generales, por ejemplo, sean frecuentes en las dos clases (o dominios) y que las frecuencias de cada palabra en ambas clases (o dominios) sean similares entre sí. A continuación describiremos como se realiza el cálculo del FCE, tomando como referencia a [Tan et al., 2009] y [Sanchez Junquera, 2018].

El FCE se calcula según la fórmula 4.1

$$FCE_w = log(\frac{P_S(w) * P_T(w)}{|P_S(w) - P_T(w)| + \beta}) \tag{4.1}$$

Para el caso muy particular de que $|P_S(w) - P_T(w)| = 0$ es que se define en la ecuación 4.1 la constante β[3]

Sean S y T una clase u otra (o el dominio fuente u objetivo) respectivamente, entonces, $P_S(w)$ y $P_T(w)$ son las probabilidades de ocurrencia del término w en cada uno de ellos. El término w es general si $P_S(w)$ y $P_T(w)$ son ambas "altas y similares".

Las probabilidades $P_S(w)$ y $P_T(w)$ se calculan, teniendo en cuenta la cantidad de instancias en las que aparecen el término w en S y T respectivamente y la cantidad total de instancias por clases (o por dominio).

$$P_S(w) = \frac{N_w^S + \alpha}{N^S + 2\alpha} \tag{4.2}$$

$$P_T(w) = \frac{N_w^T + \alpha}{N^T + 2\alpha} \tag{4.3}$$

La constante α la utilizan los autores [Tan et al., 2009] y [Sanchez Junquera, 2018] para manejar el desbordamiento.

[3]En esta tesis se toman los valores dados en [Tan et al., 2009], es decir $\alpha = \beta = 0,0001$.

4.2 Selección del Modelo: Basado en Estilo o Contenido

Como se mencionó en párrafos anteriores, el enfoque propuesto busca identificar las noticias falsas observando el estilo con que éstas son redactadas. Para ello, se enmascaran todas aquellas palabras referentes al contenido de la nota y se conservan aquellas palabras referentes al estilo, así como la estructura de nota. No obstante, para efectos de evaluar el enfoque propuesto también se realizaron experimentos enmascarando el estilo y conservando el contenido.

La idea con esta etapa, es poder determinar si a través del estilo de las noticias podemos clasificar una nota como verdadera o falsa. Al realizar experimentos evaluando también un modelo basado en el contenido buscamos ver si estos modelos dependen o no de la naturaleza y características propias de cada conjunto de dato en particular.

De esta manera, siguiendo un **modelo basado en estilo**, enmascaramos información relacionada con el contenido para mantener el estilo de la escritura predominante, es decir, seleccionamos las k palabras más frecuentes y enmascaramos las ocurrencias del resto de las palabras, las menos frecuentes. Es decir, enmascaramos todos los tokens w que cumplen con:

$$w \notin W_k \tag{4.4}$$

Siguiendo un **modelo basado en contenido**, enmascaramos información relacionada con el estilo para permitir que el sistema se entrene solo en las diferencias relacionadas con el contenido; es decir, enmascaramos las k palabras más frecuentes y mantenemos intacto el resto. Es decir, enmascaramos todos los tokens w que cumplen con:

$$w \in W_k \tag{4.5}$$

w son los términos o tokens del documento y W_k es el conjunto de las k palabras más frecuentes.

4.3 Transformación de los textos: Algoritmos de enmascaramiento

Según el diccionario de la Real Academia Española, **enmascarar**, como bien lo dice la palabra, no es más que encubrir, disfrazar o modificar la apariencia de algo para ocultar su verdadera identidad o características.

Cuando enmascaramos los textos, ocultamos la palabra mediante la sustitución de la misma por otros caracteres o símbolos. Stamatatos [Stamatatos, 2017] propone dos alternativas para enmascarar los términos de acuerdo a si son palabras o números. Los métodos, transforman los textos según un tipo de distorsión, enmascarando (sustituyendo) los términos w tal que:

(i) (DV-MA) ***Distorted View with Multiple Asterisks***: Cada palabra w se enmascara remplazando cada uno de sus caracteres con un asterisco ($*$); y cada dígito en el texto se remplaza con un símbolo de (#).

(ii) (DV-SA) ***Distorted View with Single Asterisks***: Cada palabra w se enmascara remplazando cada ocurrencia de la palabra por un solo asterisco ($*$); y cada secuencia de dígitos en el texto se remplaza por un símbolo de (#).

Cuando enmascaramos siguiendo **DV-MA** los textos son modificados al remplazar ciertas palabras y dígitos, sin embargo, mantienen la longitud de la palabra y número original, algo que se pierde completamente cuando enmascaramos con **DV-SA**. Cuando analicemos el desempeño de estas transformaciones podremos comprobar si la longitud de los términos es importante para la clasificación. Cabe mencionar que, una medida estilística relevante es la longitud de las palabras. Ya que las palabras vacías (artículos,

preposiciones, pronombres, etc.) tienden a ser palabras de pocos caracteres.

4.3.1 Enmascaramiento usado

Para esta tarea en específico, donde uno de los aspectos que pretendemos es identificar qué modelo es más discriminatorio para la clasificación (estilo o contenido) o si esto depende de la naturaleza o no del corpus, se transformaron los algoritmos de enmascaramiento (los tradicionales) antes mencionados y propuestos por [Stamatatos, 2017].

Como se hacía mención en secciones anteriores cuando enmascaramos siguiendo el modelo de Stamatatos [2017] sustituimos con (∗) y los números con (#). Para poder definir correctamente nuestro modelo y dar respuesta a: qué términos debo enmascarar y cuáles no para poder discriminar entre las notas falsas y las veraces; se realizó toda una primera fase de experimentación utilizando exactamente y tal cual, estos algoritmos tradicionales. Estos experimentos arrojaron que otros patrones asociados al estilo (signos de puntuación, exclamación e interrogación, comillas, paréntesis, etc) pudieran aportar información relevante para la clasificación más allá de las palabras y los números. Uniendo todos estos detalles se transformaron estos algoritmos tradicionales a un enmascaramiento más fino y detallado.

La siguiente tabla resume como fueron enmascarados estos nuevos atributos en nuestros algoritmos de enmascaramiento.

4.4 Selección de características

Una vez aplicados los algoritmos de enmascaramiento y modificado los textos, se extraen los n-gramas de palabras y caracteres para la representación. De acuerdo a la Etapa 4 de la Figura: 4.1, se seleccionan las características y se entrena el clasificador

Tokens	Enmascarado
Signos de puntuación, independientes del tipo de noticia: punto (.), coma (,), punto y coma (;) y dos puntos (:)	Se mantienen tal cual (. , ; :)
'Comillas simples', "comillas dobles" y <<comillas tipográficas>>	$\frown$
Paréntesis (), llaves {} y corchetes []	()
Exclamación ¡! e Interrogación ¿?	μ
Guión (-), guión bajo (_)	α
Signos matemáticos como: peso ($), por ciento (%), suma (+) e igual (=)	π
Slach (/)	Se mantiene tal cual (/)
otros símbolos	$\sim$

Tabla 4.1: Tabla de cómo se enmascararon el resto de los tokens

sobre estos n-gramas y se evalúa su clasificación sobre las instancias del conjunto de prueba. En el próximo capítulo, se evalúa la efectividad de estas transformaciones a los textos, en el sentido de obtener una representación general que permita la detección automática de noticias falsas.

Capítulo 5

Experimentos

En la fase de experimentación se realizaron una gran cantidad de pruebas que nos permitieran llegar a la mejor configuración para cada una de las colecciones, que serán los resultados mostrados en este capítulo (si se desea ver más resultados con todas las variantes y configuraciones se pueden consultar los Anexos correspondientes). La Figura 5.1 nos ilustra el esquema de experimentación seguido teniendo en cuenta los diferentes parámetros y enfoques. Los valores de k varían entre 0, cuando enmascaramos todo y 1000, cada uno de estos valores de k son tomados de cada una de las listas (palabras más frecuentes del idioma, palabras más frecuentes del corpus y FCE), definimos entonces que modelo vamos a seguir (estilo o contenido) y los algoritmos de enmascaramiento a usar.

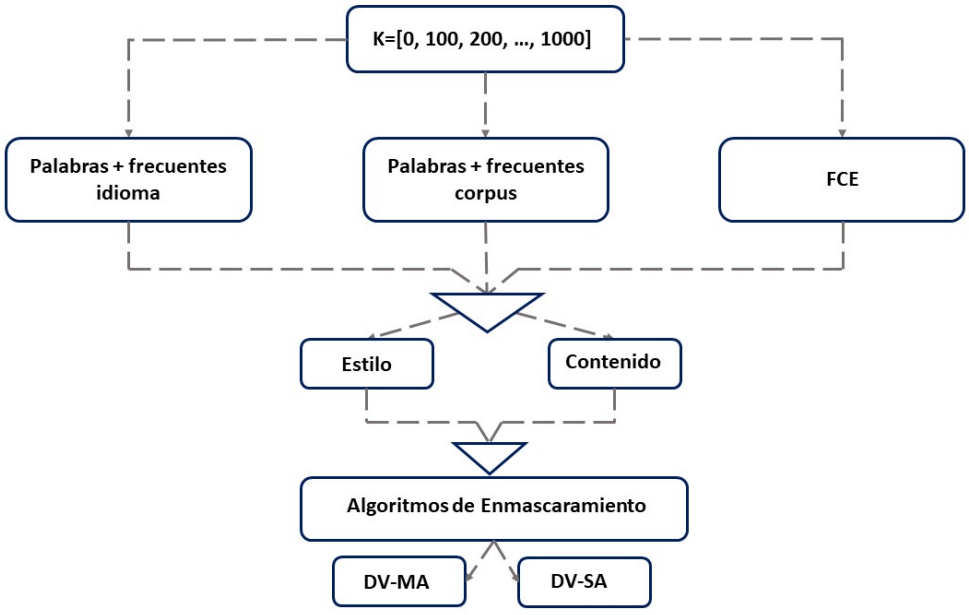

Figura 5.1: Esquema de Experimentación

El capítulo, inicia presentando y describiendo las diferentes colecciones de datos usadas en nuestra experimentación. A continuación, se resumen mediante tablas de comparación con el estado del arte, los principales resultados utilizando modelos tradicionales de aprendizaje y redes neuronales para cada uno de los conjuntos por individual. La principal razón al mostrarlo de esta forma, es que los conjuntos de datos utilizados fueron evaluados en condiciones muy diversas y no siempre se utilizaron para evaluarlos las mismas medidas de evaluación en el estado del arte. Para finalizar se resumen nuestros mejores resultados y se arriban a conclusiones previas.

5.1 Colecciones de Datos

En la detección automática de noticias falsas existen muy pocos corpus que trabajan este tema para el idioma español, el objetivo es considerar nuestro método no solo para el inglés sino también para el español y valorar la capacidad del método propuesto de ser independiente del idioma.

Finalmente, para la evaluación del método propuesto se trabajan cuatro corpus de noticias periodísticas empleados en diferentes trabajos del estado del arte: dos en idioma español y dos en inglés. Conjuntos de datos de naturaleza diferente, tanto en idioma como en dominios temáticas, los cuales serán descritos cada uno en las siguientes subsecciones.

5.1.1 MEX-A3T

El Corpus de noticias falsas en español MEX-A3T, tomado de [Posadas-Durán et al., 2019] contiene una colección de noticias compiladas a partir de varios recursos en la Web: sitios web de periódicos establecidos, sitios web de empresas, sitios web especiales dedicados a validar noticias falsas y sitios web designados por diferentes periodistas

como sitios que publican noticias falsas periódicamente.

El corpus presentado fue etiquetado considerando solo dos clases (verdadera con 491 notas y falsas con 480 notas). Cubre noticias de nueve temas diferentes: ciencia, deporte, economía, educación, entretenimiento, política, salud, seguridad y sociedad. Los autores dividieron el conjunto de datos en entrenamiento y conjunto de prueba, utilizando el 70 % del corpus para el entrenamiento y el resto para prueba. Según los autores, realizaron una distribución jerárquica del corpus, es decir, todas las categorías mantienen la relación 70 %-30 %.

Para llevar a cabo nuestra experimentación, tomamos la partición de entrenamiento y la dividimos en 80-20, para ello, utilizamos el 80 % para entrenamiento y el 20 % restante como una partición de validación. La siguiente Figura 5.2 muestra la distribución de los datos para el MEX-A3T utilizadas para el desarrollo de la experimentación.

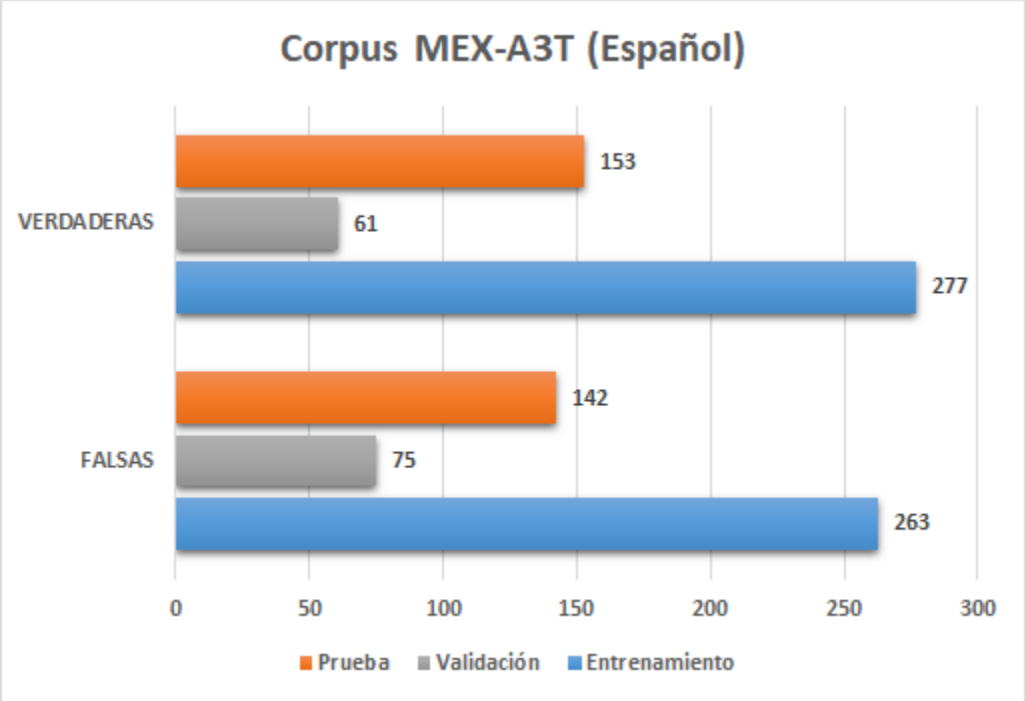

Figura 5.2: Distribución de los datos del MEX-A3T

5.1.2 RAW-CovidES

Para crear un conjunto de datos de *Fake News* en español con documentos de noticias pertenecientes al dominio de la salud (temas como COVID-19 que es un 50 % del conjunto de datos), los autores [Bonet-Jover et al., 2020], recopilaron automáticamente y de manera equilibrada, noticias falsas y verdaderas de varios periódicos en línea, blogs y sitios web de verificación de hechos. Se recopilaron un total de 200 documentos de noticias, de ellas 105 noticias verdaderas y 95 noticias falsas. Como podemos apreciar, es un conjunto de datos con muy pocas instancias.

Para el trabajo con el conjunto de datos, los autores presentaron una división aleatoria del 80 % del conjunto para entrenamiento y el 20 % restante para prueba, realizando un total de cinco corridas y reportando las medidas promedio.

Según como se muestra en la Figura 5.3, nuestra distribución de los datos consistió en la división de igual forma del 80-20 de cinco colecciones individuales para entrenamiento y prueba; y luego, de la partición de entrenamiento volvimos a tomar otro 20 % para validación.

5.1.3 LIAR

Liar es un conjunto de datos que fue puesto públicamente por [Wang, 2017]. Incluye 12.8K declaraciones cortas etiquetadas por humanos y comprende seis etiquetas de calificaciones de veracidad: *pants-fire, false, barely-true, half-true, mostly-true*, y *true*. En nuestro trabajo, utilizamos el conjunto de datos LIAR modificado y reportado por [Khan et al., 2019], donde los autores se enfocan principalmente en clasificar las noticias como reales y falsas. Para la clasificación binaria de noticias, transformaron las seis etiquetas antes mencionadas en dos etiquetas: *pants-fire, false, barely-true* se contemplan como falso y *half-true, mostly-true, true* son igualmente verdaderos, para un total de

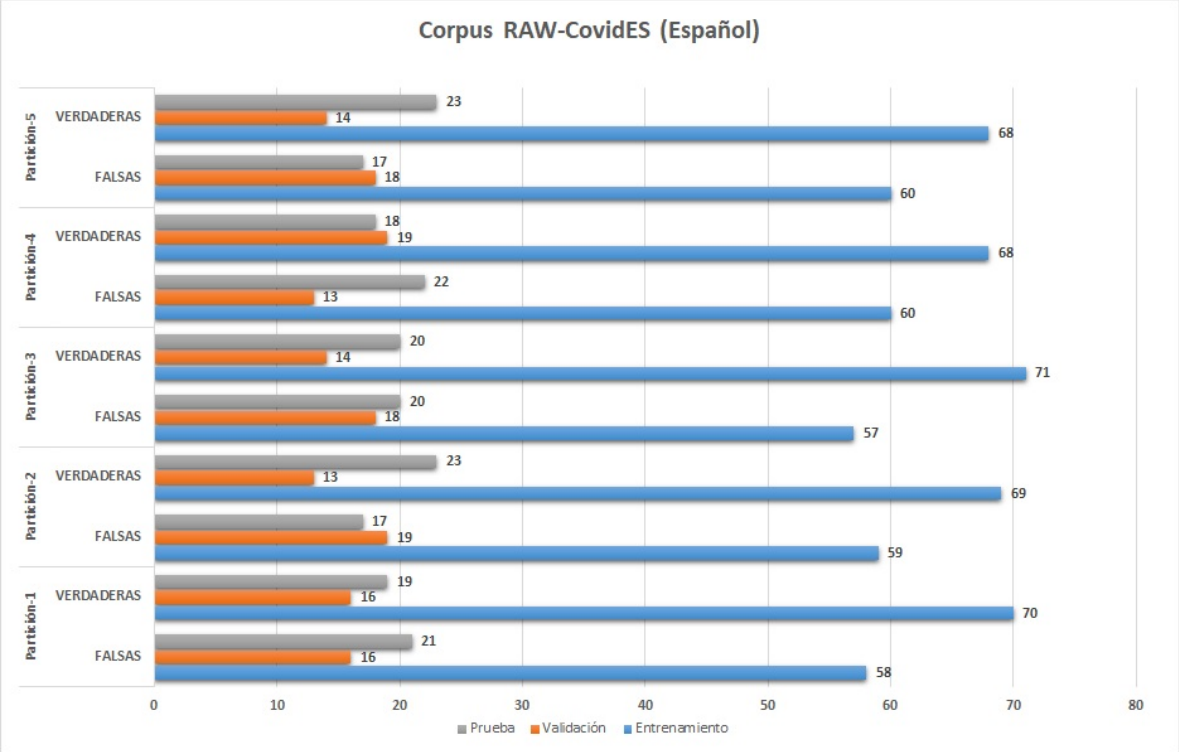

Figura 5.3: Distribución de los datos del RAW-CovidES

un 56 % de noticias verdaderas y un 44 % falsas.

Este conjunto de datos se ocupa principalmente de cuestiones políticas y cuenta con una partición para entrenamiento, una para validación y una partición de prueba. Las estadísticas de este conjunto de datos se muestran en la Figura 5.4.

5.1.4 COAID

CoAID (*Covid-19 Healthcare Misinformation Dataset*) es un conjunto de datos que como bien lo indica su nombre, incluye artículos de noticias falsas y verdaderas relacionadas con temas de salud y principalmente sobre el Covid-19. Según los autores [Cui and Lee, 2020], las fechas de publicación de la información recopilada va desde el 1 de diciembre de 2019 hasta el 1 de septiembre de 2020. En total, según el artículo lograron recopilar 204 artículos de noticias falsas y 3565 noticias verdaderas. Usaron en su

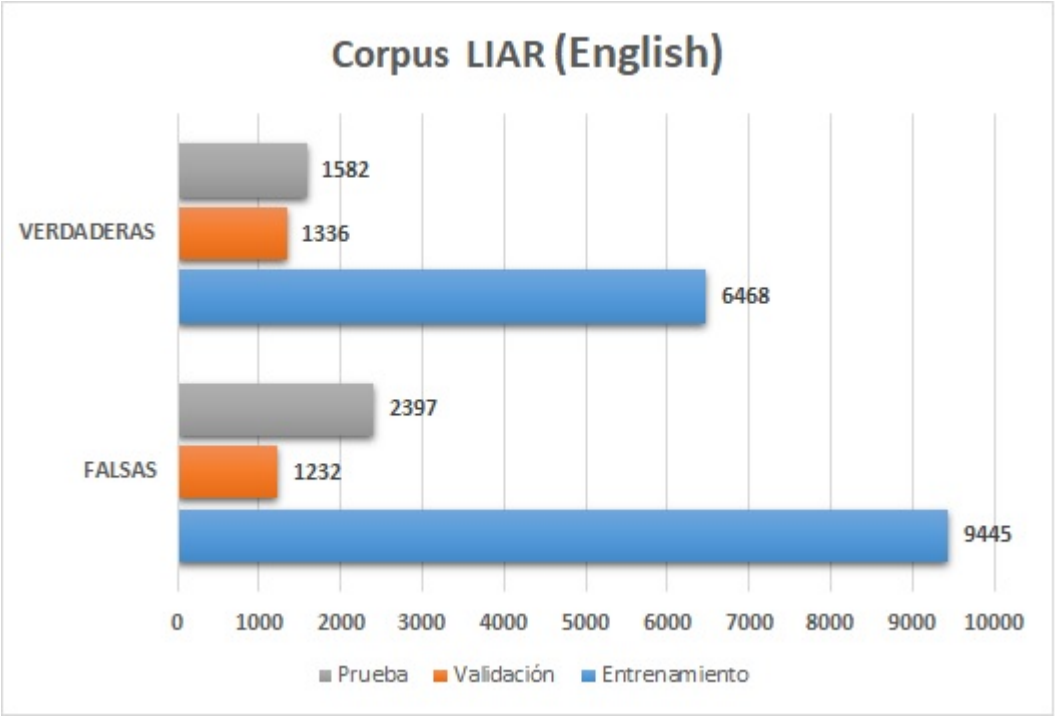

Figura 5.4: Distribución de los datos de LIAR

experimentación aleatoriamente las etiquetas del 75 % de los artículos de noticias para el entrenamiento y el 25 % restante para predecir. Finalmente, ejecutan cada método cinco veces e informan el puntaje promedio.

Para nuestros experimentos y siguiendo las pautas planteadas por los autores del conjunto de datos, y el enfoque seguido hasta ahora en nuestros experimentos, tomamos de igual forma la división aleatoria del 75-25 para obtener 4 particiones fijas del conjunto de datos y luego del 75 % utilizado para entrenamiento extrajimos un 20 % para definir nuestra partición de validación. La Figura 5.5 muestra la distribución realizada. Como bien se puede observar, según las tablas presentadas por los autores [Cui and Lee, 2020], contaban con 204 artículos de noticias falsas y 3565 de noticias verdaderas, en revisión realizada al conjunto de datos detectamos un total de 19 artículos de noticias falsas con contenido vacío, al igual que 398 de noticias verdaderas, por tanto, para nuestro trabajo contamos con 185 artículos de noticias falsas y 3167 artículos de noticias verdaderas.

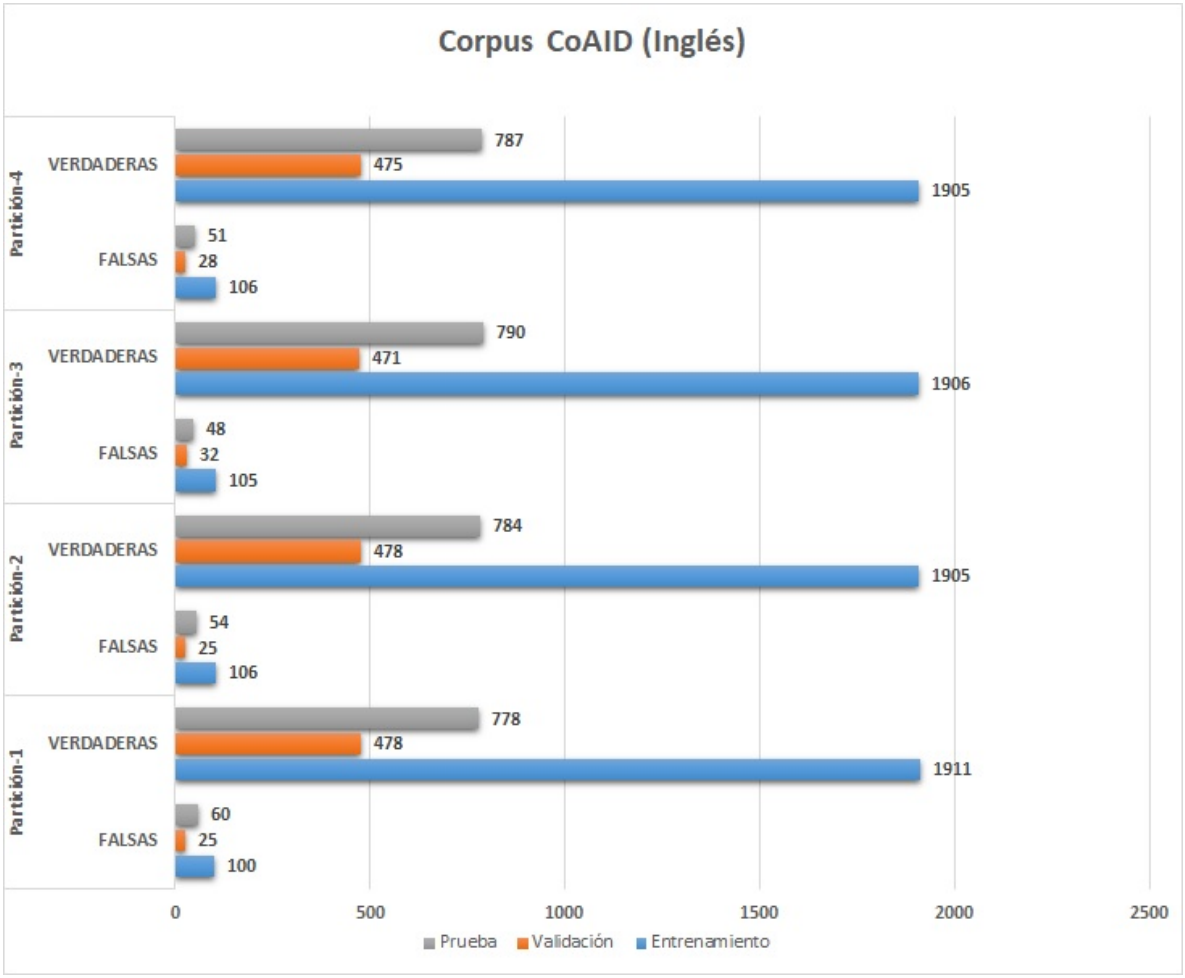

Figura 5.5: Distribución de los datos de CoAID

Otro aspecto importante sobre este conjunto de datos es el alto grado de desbalance entre sus clases.

5.2 Resultados con Modelos Tradicionales de Aprendizaje

En esta sección presentamos los principales resultados obtenidos para cada uno de los conjuntos de datos utilizando el enmascaramiento de la información textual, evaluados en el conjunto de prueba y comparado con el estado del arte existente, así

como algunas otras experimentaciones sobre las mejores configuraciones presentadas. Utilizando modelos tradicionales de aprendizaje.

5.2.1 Resultados: MEX-A3T

Nuestro mejor modelo para el MEX-A3T se obtuvo utilizando la siguiente configuración: (i) 900 palabras más frecuentes del corpus (K=900), (ii) un modelo basado en estilo y (iii) enmascarando con simples asteriscos (DV-SA).

En base a esta configuración y sobre el conjunto de prueba, obtuvimos un F1 de 0.7619 y un F1-macro de 0.7830. A efectos de tener una comparación usando el mismo clasificador que el reportado por Aragón et al. [2020], evaluamos esta configuración utilizando *Random Forest* como clasificador. Este es el clasificador usado en el trabajo antes mencionado usando *bolsa de palabras* y con la combinación de los tamaños de n-gramas de palabras [1,2,3].

Como podemos observar en la Tabla A.14, al realizar el experimento con este nuevo clasificador, el método propuesto funciona mejor, alcanzando un puntaje similar al equipo en el cuarto lugar de la competencia del 2020.

Modelo_MEX-A3T	Falsa	Verdadera	F1-macro	Exactitud
Idiap-UAM-1	**0.8444**	**0.8688**	**0.8566**	**0.8576**
Idiap-UAM-2	0.8406	0.8599	0.8502	0.8508
Ares	0.8188	0.8151	0.8169	0.8169
CIMAT-1	0.7943	0.8117	0.8030	0.8034
Nuestro modelo: Random Forest, DV-SA, Modelo basado en Estilo, K=900, palabras más frecuentes del corpus Unigramas de palabras	0.7931	0.8000	0.7966	0.7966
Baseline (BoW-RF)	0.7850	0.7879	0.7864	0.7864
Intensos-2	0.7703	0.7883	0.7793	0.7797
Nuestro modelo: SVM, DV-SA, Modelo basado en Estilo, K=900, palabras más frecuentes del corpus Unigramas de palabras	0.7619	0.7635	0.7627	0.7627
Intensos-1	0.7597	0.7376	0.7487	0.7492
Baseline (INGEOTEC)	0.7596	0.7723	0.7659	0.7661
ITCG-SD	0.7464	0.7771	0.7617	0.7627

Tabla 5.1: Comparación MEX-A3T con el estado del arte

De igual forma comparamos nuestro modelo, con los resultados reportados por

Posadas-Durán et al. [2019] en el artículo que presenta al conjunto de datos como un nuevo recurso en idioma Español para analizar y detectar información engañosa presente en las noticias.

La Tabla A.15, nos muestra dos de las configuraciones que utilizaron los autores, SVM y Random Forest con una simple bolsa de palabras, para ambos casos, podemos ver que nuestro método, con estos mismos clasificadores y con bolsa de palabras, supera al reportado por ellos en ambos casos, lo que nos indica que trabajar con la información enmascarada del texto de las noticias, hace que ambos clasificadores encuentren patrones relevantes que le permiten distinguir entre la noticia falsa y la verdadera.

Modelo	**Exactitud**
SVM (BoW)	0.7152
Random Forest (BoW)	0.7627
Nuestro Modelo: SVM, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (Unigramas de Palabras)	0.7627
Nuestro Modelo: Random Forest, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (Unigramas de Palabras)	**0.7966**

Tabla 5.2: Comparación nuestro modelo con el MEX-A3T del artículo original. Resultados en el conjunto de prueba en términos de la Exactitud

Posadas-Durán et al. [2019], experimenta también utilizando la combinación de los tamaños [3,4,5] de los n-gramas de palabras. Para este esquema, también utilizamos nuestro modelo ganador y lo llevamos a la característica propuesta por ellos. No obstante, como en nuestro proyecto el tamaño de n-gramas que utilizamos cuando nos referimos a palabras es de [1,2,3], pues mezclamos también estos tamaños y nos comparamos con ellos y con nosotros mismos cuando utilizamos la variación de [3,4,5] propuesta en el artículo.

La Tabla 5.3, hace referencia a los resultados obtenidos, en la cual podemos ver como nuestro método supera el propuesto por ellos. Llama la atención como nuestro propio

método pero utilizando la variación de n-gramas de [1,2,3] supera de igual forma la variación de [3,4,5]. Lo que nos permite entrever que para este modelo, basado en estilo, el uso de n-gramas de palabras tan grandes como de cuatro y cinco son características indicadores del contenido más que del estilo.

Modelo	**Exactitud**
SVM (n-gramas palabras [3,4,5])	0.5118
Random Forest (n-gramas palabras [3,4,5])	0.5084
Nuestro Modelo: SVM, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (n-gramas palabras [3,4,5])	0.7322
Nuestro Modelo: Random Forest, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (n-gramas palabras [3,4,5])	0.5831
Nuestro Modelo: SVM, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (n-gramas palabras [1,2,3])	0.7729
Nuestro Modelo: Random Forest, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (n-gramas palabras [1,2,3])	**0.7831**

Tabla 5.3: Comparación nuestro modelo con el MEX-A3T del artículo original utilizando combinaciones de los tamaños de los n-gramas de palabras en términos de la Exactitud

5.2.2 Resultados: RAW-CovidES

Cuando explicábamos las características y configuraciones de este conjunto de datos, mencionábamos que los autores usaron un esquema de validación cruzada en 5 pliegues. Nuestra experimentación retoma este esquema, no obstante, se tuvo el cuidado de conservar un porcentaje de los datos de cada pliegue de entrenamiento para validación.

El mejor modelo obtenido para estos datos, estuvo dado por la siguiente configuración: (i) palabras más frecuentes del idioma inglés (específicamente con una K=1000), (ii) un modelo basado en estilo, enmascarando con simples asteriscos (DV-SA) y penta-

gramas de caracteres. Este modelo obtuvo un F1 de 0.665 y como F1-macro un 0.789. La Tabla 5.4, muestra estos resultados y los compara con los resultados reportados por Bonet-Jover et al. [2020]. Es importante observar, que a pesar de que el modelo propuesto tiene un menor F1 para la clase de interés (la clase falsa), se tiene una alta precisión, lo que confirma la importancia del estilo en esta tarea.

Modelo_RAW-CovidES	Verdadera			Falsa			Exactitud	F1-macro
	Precisión	Recuerdo	F1	Precisión	Recuerdo	F1		
Baseline (Random)	0.551	0.549	0.548	0.498	0.500	0.497	0.526	0.522
Baseline (TF-IDF)	0.609	0.868	0.715	0.726	0.381	0.494	0.637	0.605
Full pipeline	**0.920**	0.550	0.790	0.680	**0.950**	**0.690**	0.750	0.740
Nuestro Modelo: SVM, DV-SA, Modelo basado en Estilo, K=1000, palabras frecuentes del idioma, pentagramas de caracteres	0.776	**0.896**	**0.914**	**0.959**	0.517	0.665	**0.825**	**0.789**

Tabla 5.4: Comparación RAW-CovidES con el estado del arte

5.2.3 Resultados: LIAR

LIAR es uno de los conjuntos de datos clásicos en la tarea de detección automática de noticias falsas, el hecho de ser frases cortas los resultados en ninguno de los experimentos son muy alentadores.

Con nuestro modelo de FCE para un K=1000, teniendo en cuenta un modelo basado en estilo y enmascarando con múltiples asteriscos, obtuvimos un F1-macro de 0.56. Sin embargo, todos los resultados reportados para el conjunto LIAR por Khan et al. [2019] son similares y ninguno supera el 0.60 (ver Tabla 5.5).

Modelo-LIAR	Representación	Exactitud	Precisión	Recuerdo	Valor-F1
SVM	Lexical	0.56	0.56	0.56	0.48
SVM	Lexical+Sentimiento	0.56	0.57	0.56	0.48
LR	Lexical+Sentimiento	0.56	0.56	0.56	0.51
Decision Tree	Lexical+Sentimiento	0.51	0.51	0.51	0.51
Adaboost	Lexical+Sentimiento	0.56	0.56	0.56	0.54
Naive Bayes	**Bigram (TF-IDF)**	**0.60**	**0.59**	**0.60**	**0.59**
k-NN	Empath Features	0.53	0.53	0.53	0.53
Nuestro Modelo: SVM, DV-MA, Estilo, K=1000, FCE	Trigramas de Palabras	0.59	0.59	0.57	0.56

Tabla 5.5: Comparación LIAR con el estado del arte

Nuestro modelo, para este conjunto de datos, aunque la diferencia no es significativa, no supera al Naive Bayes con Bigramas, no obstante si nos comparamos con el SVM utilizado por los autores, nuestro modelo es superior en comparación con el 0.48 de medida F1 obtenido por ellos.

5.2.4 Resultados: CoAID

Como habíamos referido anteriormente, el conjunto de datos CoAID, es un conjunto de datos desbalanceado con respecto a la clase falsa. En la experimentación realizada, explicábamos que en el artículo, sus autores [Cui and Lee, 2020] reportan las medidas promedio después de realizar aleatoriamente cinco corridas sobre el 75 % de entrenamiento y prediciendo sobre el otro 25 % restante. A diferencia de ellos, realizamos la división aleatoria de igual forma 75-25, tomando de ese 75 % de entrenamiento una partición para validación. Ejecutamos nuestro modelo un total de cuatro veces sobre particiones fijas y al igual que ellos, reportamos las medidas promedio obtenidas.

La Tabla 5.6, muestra nuestro mejor resultado, comparado con el estado del arte reportado en [Cui and Lee, 2020], es importante destacar aquí, que nuestro método supera todos los resultados reportados, sobre todo, podemos marcar la diferencia si nos comparamos solamente el SVM que ellos reporta.

Teniendo en cuenta, que utilizamos ambos el mismo clasificador, quisimos realizar sobre nuestra mejor configuración algunos otros experimentos. A pesar de conocer el desbalance que presenta el conjunto de datos, los autores, no utilizan la versión para datos desbalanceado que ofrece *sklearn* para SVM. Por nuestra parte, evaluamos este modelo pero utilizando esta versión antes mencionada para datos desbalanceados. También evaluamos nuestra configuración con una bolsa de palabras como lo reportan los autores.

La Tabla 5.7 muestra los resultados a estos experimentos, donde podemos comprobar

Modelo-CoAID	Precisión	Recuerdo	Valor-F1
SVM	0.4036	0.1322	0.1986
LR	0.4287	0.0690	0.1143
RF	0.6056	0.0581	0.045
CNN	0.9653	0.1238	0.1983
BiGRU	0.7476	0.0524	0.0930
CSI	0.6814	0.2109	0.2283
SAME\v	0.8922	0.2991	0.3400
HAN	0.6965	0.4659	0.5471
dEFEND	**0.8965**	0.4847	0.5814
Nuestro Modelo: SVM, DV-SA, Estilo, K=700, palabras frecuentes del corpus, cuatrigramas de caracteres	0.8903	**0.6315**	**0.6882**

Tabla 5.6: Comparación CoAID con el estado del arte

no solo la efectividad de nuestro método sino de igual forma las ventajas de utilizar la opción para datos desbalanceados que se nos ofrece cuando utilizamos las máquinas de soporte vectorial.

Modelo-CoAID	Precisión	Recuerdo	Valor-F1
SVM	0.4036	0.1322	0.1986
dEFEND	**0.8965**	0.4847	0.5814
Nuestro Modelo: SVM, DV-SA, Estilo, K=700, palabras frecuentes del corpus, cuatrigramas de caracteres	0.8903	0.6315	0.6882
Nuestro Modelo: SVM (balanceado), DV-SA, Estilo, K=700, palabras frecuentes del corpus, cuatrigramas de caracteres	**0.8661**	**0.8019**	**0.8299**

Tabla 5.7: Comparación CoAID con el estado del arte. Utilizando un SVM para datos desbalanceados

5.3 Resultados utilizando Modelos basados en Redes Neuronales

De igual forma que la fase de experimentación anterior con modelos tradicionales, en la experimentación con redes neuronales (CNN), evaluamos siguiendo un modelo

basado en estilo y para efectos de evaluar el enfoque propuesto también se realizaron experimentos enmascarando el estilo y conservando el contenido (modelo basado en contenido).

Para este caso específico con las redes neuronales, alimentamos nuestra red con el texto ya enmascarado y *embeddings* pre-entrenados de Glove. Pasamos a la red solamente texto enmascarado con simples asteriscos (algoritmo DV-SA).

Se realizaron los experimentos sobre los mismos cuatro conjuntos de datos descritos anteriormente, y se utilizó la misma distribución de los datos para cada uno de ellos como se expuso en la Sección 5.1.

5.3.1 Resultados: MEX-A3T

Como podemos apreciar en la Tabla 5.8 podemos ver nuevamente el estado del arte para el conjunto de datos de noticias falsas en español, MEX-A3T, pero esta vez comparamos nuestro modelo utilizando CNN con el resto de los resultados reportados.

Podemos observa como, utilizando el texto enmascarado como entrada a la red, obtenemos un resultado bastante apropiado y que nos sitúa en la tercera posición de la tabla para el caso muy particular del MEX-A3T.

Modelo_MEX-A3T	Falsa	Verdadera	F1-macro	Exactitud
Idiap-UAM-1	**0.8444**	**0.8688**	**0.8566**	**0.8576**
Idiap-UAM-2	0.8406	0.8599	0.8502	0.8508
Nuestro Modelo: CNN, DV-SA, Modelo Basado en Estilo Kernel=4 y K= 1000 y las palabras más frecuentes del corpus	0.8251	0.8153	0.8202	0.8203
Ares	0.8188	0.8151	0.8169	0.8169
CIMAT-1	0.7943	0.8117	0.8030	0.8034
Baseline (BoW-RF)	0.7850	0.7879	0.7864	0.7864
Intensos-2	0.7703	0.7883	0.7793	0.7797
Intensos-1	0.7597	0.7376	0.7487	0.7492
Baseline (INGEOTEC)	0.7596	0.7723	0.7659	0.7661
ITCG-SD	0.7464	0.7771	0.7617	0.7627

Tabla 5.8: Comparación de nuestro modelo para el MEX-A3T con Redes Neuronales Convolucionales con el estado del arte reportado

5.3.2 Resultados: RAW-CovidES

Con nuestro modelo, utilizando redes neuronales convolucionales, para el conjunto de datos RAW-CovidES no se obtienen los resultados esperados. En la Tabla 5.9 se muestran los resultados referidos anteriormente.

Modelo RAW-CovidES	Falsa	Verdadera	F1-macro	Exactitud
Baseline (Random)	0.497	0.548	0.522	0.526
Baseline (TF-IDF)	0.494	0.715	0.605	0.637
Full pipeline	**0.690**	0.790	0.740	0.750
Nuestro Modelo: CNN, DV-SA, Modelo Basado en Contenido Kernel = 2 y FCE, K=300	0.579	0.438	0.509	0.519

Tabla 5.9: Comparación de nuestro modelo para el RAW-CovidES con Redes Neuronales Convolucionales y el estado del arte reportado

A diferencia de lo que pudimos observar con el MEX-A3T, donde la CNN obtenía su mejor resultado siguiendo un modelo basado en estilo, este conjunto de datos, reporta su mejor resultado, con un modelo basado en contenido. La diferencia radica en que el MEX-A3T es de múltiples dominios y variadas temáticas, mientras que el RAW-CovidES es temático para el dominio de salud.

5.3.3 Resultados: LIAR

Como se había planteado con anterioridad, las características propias del conjunto de datos LIAR (*statement* u oraciones muy cortas) lo hacía aun más difícil a la hora de la clasificación, por ello, los resultados obtenidos no siempre fueron buenos. La Tabla 5.10 muestra nuestro mejor resultado para este conjunto utilizando CNN y comparado con el estado del arte reportado, nos referimos al mejor resultado que reportan los autores en [Khan et al., 2019] utilizando modelos tradicionales de *Machine Learning* y utilizando modelos de redes neuronales.

Al igual que para el conjunto de datos RAW-CovidES, que son conjuntos temáticos. El dominio de las noticias de LIAR es político, por tanto, aquí vemos nuevamente como

Modelo-LIAR	Características	Exactitud	Precisión	Recuerdo	Valor-F1
Naive Bayes	**Bigram (TF-IDF)**	**0.60**	**0.59**	**0.60**	**0.59**
CNN	Glove Embedding	0.58	0.58	0.58	0.58
Conv-HAN	Glove Embedding	0.59	0.59	0.59	0.59
Nuestro modelo: CNN, DV-SA, Modelo basado en Contenido Kernel = 4 y FCE, K=100	**Glove Embedding**	**0.60**	**0.59**	**0.58**	**0.58**

Tabla 5.10: Comparación de nuestro modelo para LIAR con Redes Neuronales Convolucionales y el estado del arte reportado

la red muestra para estos casos su mejor resultado en el modelo basado en contenido. Aunque no se supera el estado del arte para este conjunto los resultados obtenidos son bastante similares a su mejor modelo con Naive Bayes.

5.3.4 Resultados: CoAID

Los resultados mostrados en la Tabla 5.11 resumen la comparación de nuestro método de enmascarado utilizando las redes neuronales convolucionales, donde apreciamos como el valor F1 de 0.8371 supera el estado del arte reportado por los autores Cui and Lee [2020].

Modelo-CoAID	Precisión	Recuerdo	Valor-F1
SVM	0.4036	0.1322	0.1986
dEFEND	0.8965	0.4847	0.5814
Nuestro modelo: CNN, DV-SA, Modelo Basado en Contenido Kernel = 5, palabras_frecuentes_corpus K=100	**0.9134**	**0.7868**	**0.8371**

Tabla 5.11: Comparación de nuestro modelo para el conjunto de datos CoAID, ahora con Redes Neuronales Convoluciones con el estado del arte reportado y con nuestro mejor modelo utilizando SVM

Volvemos a encontrar aquí, que dada la naturaleza temática del corpus CoAID (dominio salud), la red neuronal devuelve el mejor resultado cuando enmascara el estilo, dejando libre las palabras de contenido, es decir para un modelo basado en contenido.

5.4 Conclusiones Previas

Con el fin de comprender que hemos podido ir concluyendo de todas nuestras experimentaciones, en esta sección daremos respuestas tentativas a cada una de las preguntas que nos realizamos en la Figura 4.1, del Capítulo anterior: Método Propuesto.

En primera instancia, como datos generales, tenemos que, cuando utilizamos las técnicas de enmascaramiento con los algoritmos de aprendizaje tradicional siempre nos fue mejor siguiendo un modelo basado en estilo, en cambio con las redes neuronales, empieza a funcionar mejor con un modelo basado en contenido.

El método propuesto, depende de la naturaleza del corpus con el que se esté tratando. Para la mayoría de los conjuntos de datos cuando enmascarábamos utilizando tanto SVM como la CNN, los mejores resultados se obtenían con las k palabras más frecuentes del corpus o a partir del cálculo del FCE. El FCE es más discriminatorio cuando trabajamos con conjuntos que son temáticos.

La naturaleza o el tipo de corpus también influye en el tipo de modelo que vamos a escoger. Para este aspecto, la red neuronal fue más consistente: si el conjunto de datos es temático utilizar un modelo basado en contenido es la mejor opción, en cambio, cuando clasificamos un conjunto con varios dominios o temas de noticias, el clasificador se centra más en las características estilísticas que de contenido, por ello, un modelo basado en estilo es quien resuelve el problema para este caso.

Si enmascaramos la totalidad de nuestros textos ($k = 0$), siguiendo la estrategia de múltiples asteriscos o simples asteriscos, pudimos apreciar que es relevante para la clasificación la longitud original del token. Fueron significativas las diferencias entre ambos algoritmos, obteniendo con DV-MA las más altas puntuaciones. No obstante, en el enmascarado con un $k > 0$ en tres de los cuatros conjuntos de datos (excepto LIAR), el enmascarado con simple asteriscos (DV-SA) resultó mejor, lo que nos pudiera indicar

que el clasificador está buscando patrones más allá de las palabras o los números que enmascaramos, contempla más bien el uso de signos y otras características de estilo, que serán mejor analizadas en el próximo capítulo.

Otro aspecto importante a analizar es el parámetro K y sus variaciones. Para un modelo basado en estilo, el valor de k se hace más relevante a medida que sobrepasa el $k >= 600$ y para un modelo basado en contenido la relevancia de k, no excede la $k = 400$.

En cuanto a los n-gramas, para algunos conjuntos nos fue mejor con palabras y para otros con caracteres. Para la CNN, solo experimentamos con n-gramas de palabras y sus variaciones estuvieron dadas desde [1-5], aquí, mientras más grande el tamaño del n-grama [4 o 5] se capturaba mejor el contenido de la noticia y mejor era la puntuación obtenida.

Capítulo 6

Análisis de Resultados

En el presente capítulo, se presentan aspectos relevantes que afectan el rendimiento del método propuesto.

A partir de los experimentos realizados, mostrados en el capítulo anterior, se tienen respuestas preliminares a las interrogantes planteadas en la Figura 4.1. A continuación se listas algunas de estas respuestas:

- En primera instancia, y principal conclusión es que, el método propuesto depende de la naturaleza de los conjuntos de datos evaluados.
- Funciona mucho mejor y es más discriminativo cuando enmascaramos con las k palabras más frecuentes del corpus y las extraídas a partir del cálculo del FCE entre las clases.
- Si el conjunto de datos es puramente temático utilizar un modelo basado en contenido es la mejor opción.
- Cuando clasificamos un conjunto con varios dominios o temas de noticias, el clasificador se centra más en las características estilísticas que de contenido, por ello, un modelo basado en estilo es el indicado.

Para continuar respondiendo a la lista de preguntas planteadas, las siguientes secciones presentan el análisis de los resultados enfocados en dos aspectos: (i) observar el comportamiento del método al modificar el parámetro k. Diferentes variaciones del parámetro k son analizadas, se observaron condiciones como cuándo $k = 0$, un caso extremo donde todo el texto es enmascarado; así como situaciones cuando $k >= 600$ o cuando $k <= 400$, límites que se observaron en los experimentos previos; (ii) analizar

a través de un análisis basado en ganancia de información el tipo de palabras que son relevantes para la discriminación en esta tarea.

6.1 Efecto del parámetro k

Como se puede notar (ver Tablas de resultados en el Apéndice A), para cualquier valor de k, los resultados de $F1$ son muy variables, obtener un valor fijo para este parámetro es casi imposible, un valor adecuado para este parámetro depende de las características específicas de cada una de las clases, en cada una de los corpus trabajados.

Sin embargo, se puede observar la importancia de seleccionar un conjunto de términos ($k > 0$), ya que en general, se obtienen mayores valores de $F1$ con valores de k mayores a cero.

Cuando $k <= 400$, quiere decir que nuestro conjunto de palabras más frecuentes está prácticamente constituido de palabras vacías o *stop-words*, algunos adverbios, algunos verbos y formas verbales, en fin, palabras tan generales que son utilizadas de igual manera en cualquiera de los dos tipos de noticias. Entonces, cuando evaluamos siguiendo un modelo basado en contenido, estamos enmascarando la información relacionada con el estilo, es decir el clasificador recibe esas k palabras enmascaradas y deja al descubierto el resto, o términos más específicos a cada una de las clases es por ello que se obtiene los mejores resultados para estas $k <= 400$.

En cambio, cuando evaluamos para un modelo basado en estilo, obtenemos nuestros mejores resultados cuando $k >= 600$, quiere decir que estamos seleccionando un rango de palabras frecuentes que ya, no solo incluye aquellas que son extremadamente generales sino, que también, incorpora términos específicos a cada una de las clases, y que a pesar de seguir siendo de las más frecuentes ya nos brindan cierta información de contenido.

Para el caso más específico y que mencionábamos con anterioridad, cuando la $k = 0$, quiere decir que enmascaramos todo el documento, pudimos apreciar que siempre el mejor puntaje de F1 macro se obtenía cuando enmascarábamos utilizando $DV - MA$, es decir manteniendo la longitud de las palabras, en cambio con $DV - SA$, los valores de F1, siempre resultaban ser muy bajos. Esto nos alertaba acerca de la importancia de la longitud de los términos para la clasificación. La siguiente Figura 6.1 muestra la longitud de las palabras por clases para cada uno de los siguientes conjuntos de datos.

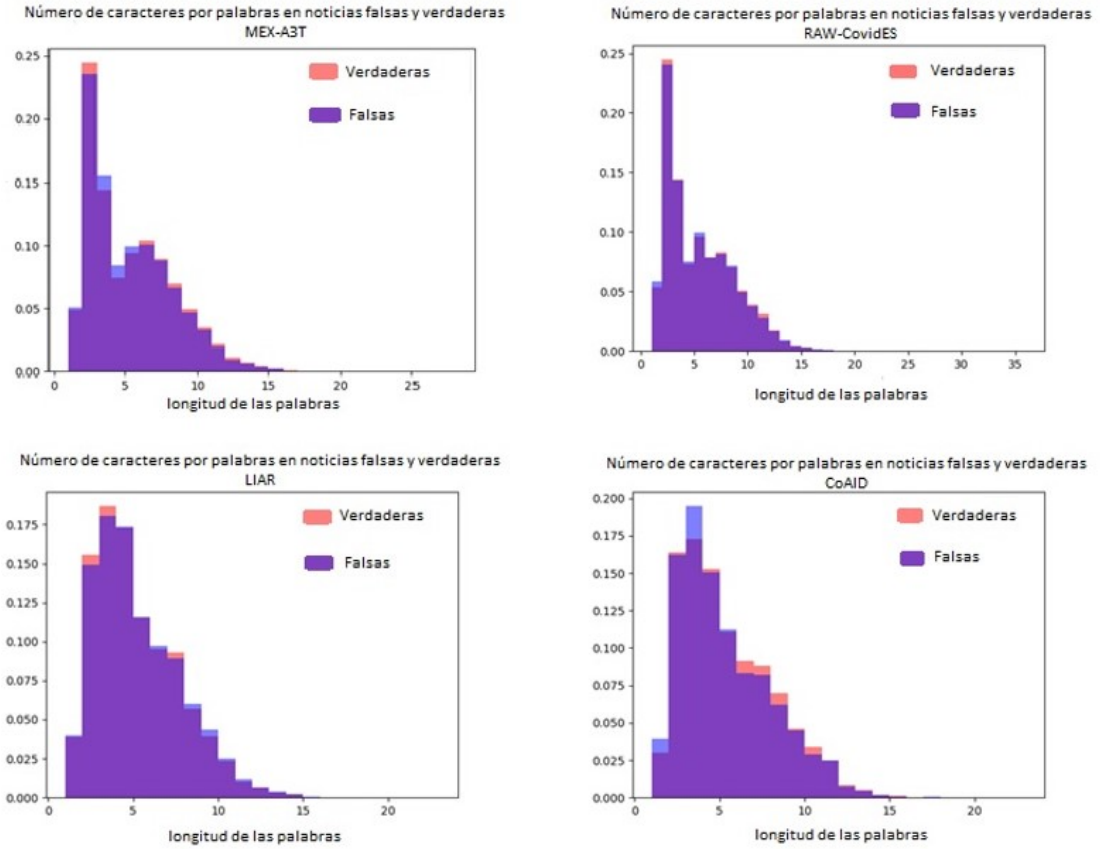

Figura 6.1: Longitud de las palabras por clases para cada uno de los conjuntos de datos.

Se puede ver, como, no hay tanta diferencia de tamaños y proporción de las palabras entre las clases. En la siguiente sección veremos algunos ejemplos de palabras con mayor ganancia de información para cada tipo de conjuntos de datos (español o inglés).

6.2 Análisis de la ganancia de Información

Mediante el análisis de la ganancia de información (IG por sus siglas en inglés), pudimos determinar patrones que nos ayudaron a entender, los aspectos más relevantes en la clasificación de las noticias falsas mediante el enmascaramiento del texto.

Uno de estos aspectos fue la longitud de las palabras con ganancia de información para cada conjunto de datos. Para los conjuntos de datos en español, si observamos la Figura 6.1, podemos ver que no existe mucha diferencia entre las longitudes entre clases, no obstante; palabras de tamaño 1, 2, 6, 8, 10 y 11 se muestran con ciertas diferencias y se obtienen con ganancia de información entre 0.093 al 0.054.

Por ejemplo, para las palabras con mayor ganancia de información, pudimos encontrar, sobre todo en las notas falsas, que tienen mayor frecuencia adverbios con la terminación 'mente' (finalmente, totalmente, detenidamente, terriblemente, precariamente, retóricamente). De igual forma, en el corpus del español mexicano otras muchas como: faltas de ortografía, palabras para el uso de expresiones idiomáticas o expresiones coloquiales comunes que solo se presentan en las notas falsas y sobre las que el clasificador puede estar enfocándose para clasificar estas notas falsas: (malcogidos, malheridos, sansseacabó, totalquien, híjole, jajaja, jijiji, pollas, Wuuuuu, huevones, malumización, jijos, güey).

Para los conjuntos de datos en inglés palabras de tamaño 2, 3, 4, 5 y 7 se muestran con ciertas diferencias entre clases, y se obtienen con alta ganancia de información para estos conjuntos. Para este caso muy particular del inglés, es muy difícil diferenciar los usos de cada una de estas palabras por clases ya que la mayoría se utilizan en ambas noticias y con la misma finalidad, solo que su uso es más repetido en una clase u en otra.

Para el caso del CoAID donde existe un desbalance abrupto de los datos sobre la

clase falsa, ésta puede ser una de las razones por las cuales el clasificador se centra en los patrones que encuentra más a menudo en las verdaderas y clasifica con mejor puntaje F1 esta clase. Por su parte, para LIAR, cuyos datos no están desbalanceados, algunas diferencias significativas las podemos apreciar sobre todo en el uso de abreviaturas de tamaño 3 en la clase falsa: ('gov' en vez de 'government', 'rep' en vez de 'republican', así como los meses del año en abreviatura, por ejemplo 'dec' en vez de 'december').

Otros aspectos importantes que fueron conclusivos a partir de la ganancia de información:

- El uso de datos numéricos.
- El uso del punto (.) y la coma (,).
- El uso de paréntesis abierto (y cerrado), que engloba el uso de: (paréntesis) , [corchetes] y {llaves}.
- El uso de las comillas.
- El uso de los dos puntos (:).
- El uso del punto y coma (;)

6.2.1 Uso de datos numéricos

Para todos los conjuntos de datos, pudimos comprobar que el uso de datos numéricos es un aspecto importante en la clasificación, el mayor uso de valores numéricos los vemos en la clase verdadera, pero en realidad este hecho está dado porque la longitud promedio por textos es mayor para esta clase.

En las noticias falsas tienen mayor participación datos numéricos cuando nos referimos a: fechas y edades principalmente, y algunos pocos a valores estadísticos. Al

parecer, lo que sucede es que el escritor de engaño pretende crear la sensación de una noticia real al estar correctamente ubicada en tiempo y espacio, sin embargo evade dar datos estadísticos, ya sea por ignorancia o porque estos pueden ser verificables y/o comparables.

En las noticias verdaderas al igual que en las falsas: fechas y edades están presentes, pero se mencionan mucho más datos porcentuales, de conteo o dinero y de uso estadístico. Al parecer, el escritor de una nota verdadera sustenta su veracidad sobre la ciencia y no teme dar todo detalle matemático que permita validar su información como correcta.

6.2.2 Uso del punto (.), la coma (,) y el punto y coma (;)

En nuestro proceso del enmascarado es bueno recordar que estos signos de puntuación: punto, coma y el punto y coma, no se enmascararon, se mantuvieron tal cual y la principal razón para ello, fue que son signos de puntuación de uso general.

Consideramos que su relevancia en estos casos tuvo mucho que ver con la longitud de las noticias por clases. Para hacerlo más ilustrativo la Tabla 6.1 muestra la longitud promedio de la nota por palabras y por oraciones.

	MEX-A3T		**RAW-CovidES**		**LIAR**		**CoAID**	
	Fake	**True**	**Fake**	**True**	**Fake**	**True**	**Fake**	**True**
Longitud promedio por palabras	304	473	532	623	17	18	72	63
Longitud promedio por oraciones	9.0	17.0	22	24	1	1	5	4

Tabla 6.1: Longitud promedio de los textos por oraciones y por palabras

6.2.3 El uso de las comillas

Normalmente en nuestra escritura utilizamos las comillas para: enmarcar citas textuales en un párrafo, delimitar los títulos de partes internas de obras o episodios de series, en un texto, delimitar los títulos de leyes, programas, planes, proyectos o asignaturas que son muy extensos, delimitar los títulos de ponencias, discursos y exposiciones y para resaltar las palabras que se emplean en tono irónico.

Para ambas clases el uso de las comillas es relativamente similar, la principal diferencia radica en el uso de estas para enmarcar la ironía y sobre todo su uso es mucho más común por no decir que del todo común en las notas falsas.

Capítulo 7

Conclusiones y trabajo futuro

7.1 Conclusiones

En este trabajo se define ***Fake News*** como *noticias publicadas por un medio de comunicación, que incluyen: afirmaciones, declaraciones, discursos, publicaciones, entre otros tipos de información y su autenticidad no es verificable (falsa)*, [Zhou et al., 2019; Abonizio et al., 2020].La mayoría de las perspectivas de estudio de la noticia identificaban las noticias falsas después de su propagación y un modelo basado en estilo podría apoyar la tarea de detección temprana. Por tanto, nuestro enfoque estuvo encaminado hacia la detección de noticias falsas siguiendo una perspectiva basada en estilo: donde se analiza la información del contenido de la noticia e incluye el estilo de escritura [Koumouridis, 2020].

Con el objetivo de seleccionar el conjunto de las palabras más frecuentes, pudimos concluir que el método propuesto, depende en su mayoría de la naturaleza del corpus con el que se esté tratando, para 3 de 4 de los conjuntos de datos, los mejores resultados se obtuvieron con las palabras más frecuentes del corpus o a partir del cálculo del FCE.

La naturaleza o el tipo de corpus también influyó en el tipo de modelo, ya fuera estilo o contenido: si el conjunto de datos es temático utilizar un modelo basado en contenido es la mejor opción, en cambio, para un conjunto con varios dominios o temas de noticias, un modelo basado en estilo es el que resuelve mejor el problema de clasificación.

Para tres de los cuatros conjuntos de datos, el enmascarado con simple asteriscos ($DV - SA$) siempre resultó mejor, siguiendo un modelo basado en estilo, el clasificador está buscando patrones más allá de las palabras o los números que enmascaramos,

contempla más bien el uso de signos y otras características de estilo, como fueron, el punto, la coma, el punto y coma, dos puntos, el uso del paréntesis y las comillas.

Se comprobó que los valores del parámetro K jugaron un papel crucial en el desempeño del método. Para un modelo basado en estilo el valor de k se hizo más relevante para una $k >= 600$, y para un modelo basado en contenido la relevancia de k ocurría cuando no excedía la $k = 400$.

En cuanto a los n-gramas, para algunos conjuntos nos fue mejor con palabras y para otros con caracteres. Para el caso particular de la CNN, solo experimentamos con palabras y sus variaciones estuvieron dadas desde [1-5]-gramas, aquí, mientras más grande el tamaño del n-grama [4 ó 5] se capturaba mejor el contenido de la noticia y mejor era la puntuación obtenida.

Pudimos concluir, además, que el método propuesto es de fácil implementación, y no requiere de recursos externos.

Las principales limitaciones del método propuesto están relacionadas primeramente a la dependencia de cada tipo de corpus según su naturaleza, lo que lo hace dependiente a la selección de los parámetros y modelos. Los valores de sus parámetros k y n.

En este sentido, nuestra contribución está dada por:

- Se utiliza por primera vez el método del enmascaramiento en el área de detección de noticias falsas.

- En comparación con resto de los trabajos del estado del arte, el modelo propuesto fue evaluado en diferentes conjuntos de datos en condiciones muy diversas (diferentes idiomas, un solo dominio o tema y múltiples dominios). Concluyendo que el método puede ser fácilmente aplicado a diferentes idiomas, ya que es independiente a estos.

- Comparado con lo algoritmos tradicionales de enmascaramiento, nuestro método

utiliza un enmascaramiento más detallado de patrones asociados al estilo (signos de puntuación y otros símbolos como los paréntesis, las comillas, etc) más allá de las palabras y los números.

- Se ubica en el estado del arte de la detección automática de noticias falsas, con nuevos resultados, y satisfactorios para las versiones empleadas en esta tesis de los conjuntos de datos MEX-A3T, RAW-CovidES, LIAR y CoAID.

7.2 Trabajo Futuro

Atendiendo a la literatura consultada, el método propuesto es la primera vez que se utiliza para la detección de noticias falsas. A continuación se indican algunas consideraciones a seguir para trabajos futuros.

- Poder caracterizar el comportamiento de los modelos ante las diferentes bases de datos.
- Buscar alternativas para pasar a la red neuronal características a nivel carácter; dado el desempeño alcanzado con estas caracterizaciones usando modelos tradicionales.
- Evaluar nuestro método en dominios cruzados. Para ello se tendría que acompañar el método propuesto con algún método de selección de instancias en el dominio fuente para poder realizar con éxito la adaptación entre dominios.
- Dada la independencia del método respecto al idioma, también sería interesante evaluar el alcance del método en escenarios translingües, donde se desearía aprovechar la existencia de grandes colecciones ya existentes en otros idiomas.

Bibliografía

Abonizio, H. Q., de Morais, J. I., Tavares, G. M., and Junior, S. B. (2020). Language-independent fake news detection: English, Portuguese, and Spanish mutual features. *Future Internet*, 12(5):87.

Ahmad, I., Yousaf, M., Yousaf, S., and Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, 2020.

Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Aragón, M. E., Jarquín-Vasqueza, H., Montes-Y-Gómez, M., Escalante, H. J., Villasenõr-Pineda, L., Gómez-Adorno, H., Posadas-Durán, J. P., and Bel-Enguix, G. (2020). Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican Spanish. In *CEUR Workshop Proceedings*, volume 2664, pages 222–235.

Bacciu, A., Morgia, M. L., Mei, A., Nemmi, E. N., Neri, V., and Stefa, J. (2019). Bot and gender detection of twitter accounts using distortion and LSA notebook for PAN at CLEF 2019. *CEUR Workshop Proceedings*, 2380.

Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P., and García-Cumbreras,

M. Á. (2020). Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Systems with Applications*, page 114340.

Buntain, C. and Golbeck, J. (2017). Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215. IEEE.

Castelo, S., Santos, A., Almeida, T., Pham, K., Freire, J., Elghafari, A., and Nakamura, E. (2019). A topic-agnostic approach for identifying fake news pages. In *2019 World Wide Web Conference, WWW 2019*, pages 975–980. Association for Computing Machinery, Inc.

Cui, L. and Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Ghanem, B. (2018). Fact checking: Detecting and verifying facts. *CEUR Workshop Proceedings*, 2251:19–23.

Ghanem, B., Montes-Y-Gómez, M., Rangel, F., and Rosso, P. (2018). UPV-INAOE-autoritascheck that: Preliminary approach for checking worthiness of claims. *CEUR Workshop Proceedings*, 2125.

Girgis, S., Amer, E., and Gadallah, M. (2018). Deep learning algorithms for detecting fake news in online text. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pages 93–97. IEEE.

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., et al. (2018). Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, pages 17–21.

Granados, A., Cebrián, M., Camacho, D., and De Borja Rodríguez, F. (2011). Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1090–1102.

Hardalov, M., Koychev, I., and Nakov, P. (2016). In search of credible news. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 172–180. Springer.

Horne, B. D. and Adali, S. (2017). This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. In *Eleventh International AAAI Conference on Web and Social Media*.

Jang, B., Kim, I., and Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, 14(8):e0220976.

Jimenez-Villar, V., Sánchez-Junquera, J., Montes-Y-Gómez, M., Villaseñor-Pineda, L., and Ponzetto, S. P. (2019). Bots and Gender Profiling using Masking Techniques Notebook for PAN at CLEF 2019. *CEUR Workshop Proceedings*, 2380.

Kaliyar, R. K., Goswami, A., Narang, P., and Sinha, S. (2020). Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44.

Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., and Afroz, S. (2019). A benchmark study on machine learning methods for fake news detection. *arXiv preprint arXiv:1905.04749*.

Koumouridis, G. (2020). Improving fake news detection with linguistic cues. Accessed: December 7, 2021.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kumar, S., Kumar, S., Yadav, P., and Bagri, M. (2021). A survey on analysis of fake news detection techniques. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 894–899. IEEE.

Kumar, S. and Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. (2017). Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256.

Ma, J., Gao, W., and Wong, K. F. (2018). Rumor detection on twitter with tree-structured recursive neural networks. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, pages 1980–1989. Association for Computational Linguistics.

O'Brien, N., Latessa, S., Evangelopoulos, G., and Boix, X. (2018). The language of fake news: Opening the black-box of deep learning based detectors.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Pisarevskaya, D. (2017). Deception detection in news reports in the russian language: Lexics and discourse. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 74–79.

Posadas-Durán, J.-P., Gomez-Adorno, H., Sidorov, G., and Escobar, J. J. M. (2019). Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240.

Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

Roy, A., Basak, K., Ekbal, A., and Bhattacharyya, P. (2018). A deep ensemble framework for fake news detection and classification. *arXiv preprint arXiv:1811.04670*.

Ruchansky, N., Seo, S., and Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *International Conference on Information and Knowledge Management, Proceedings*, volume Part F1318 of *CIKM '17*, pages 797–806. Association for Computing Machinery.

Sanchez Junquera, J. (2018). *Adaptación de dominio para la detección automática de textos enganosos*. PhD thesis, Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica.

Sánchez-Junquera, J., Rosso, P., Montes, M., Chulvi, B., et al. (2021). Masking and bert-based models for stereotype identication. *Procesamiento del Lenguaje Natural*, 67:83–94.

Sánchez-Junquera, J., Rosso, P., Montes-y Gómez, M., and Ponzetto, S. P. (2019). Unmasking Bias in News. *arXiv preprint arXiv:1906.04836*.

Sánchez-Junquera, J., Villaseñor-Pineda, L., Montes-y Gómez, M., Rosso, P., and Stamatatos, E. (2020). Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, 135:122–130.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3):1–42.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Singh, V., Dasgupta, R., Sonagra, D., Raman, K., and Ghosh, I. (2017). Automated fake news detection using linguistic analysis and machine learning. In *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation (SBP-BRiMS)*, pages 1–3.

Sriram, S. (2020). An evaluation of text representation techniques for fake news detection using: Tf-idf, word embeddings, sentence embeddings with linear support vector machine.

Stamatatos, E. (2017). Authorship attribution using text distortion. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, volume 1, pages 1138–1149.

Stamatatos, E. (2018). Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and technology*, 69(3):461–473.

Tan, S., Cheng, X., Wang, Y., and Xu, H. (2009). Adapting naive bayes to domain adaptation for sentiment analysis. In *European Conference on Information Retrieval*, pages 337–349. Springer.

Trueman, T. E., Kumar, A., Narayanasamy, P., and Vidya, J. (2021). Attention-based c-bilstm for fake news detection. *Applied Soft Computing*, page 107600.

Vishwakarma, D. K. and Jain, C. (2020). Recent state-of-the-art of fake news detection: A review. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2:422–426.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., and Yu, P. S. (2018). TI-CNN: Convolutional Neural Networks for Fake News Detection. *arXiv preprint arXiv:1806.00749*.

Zhang, J., Dong, B., and Yu, P. S. (2020). FakeDetector: Effective fake news detection with deep diffusive neural network. In *Proceedings - International Conference on Data Engineering*, volume 2020-April, pages 1826–1829. IEEE.

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. (2019). Fake news early detection: An interdisciplinary study. *arXiv preprint arXiv:1904.11679*.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Apéndice A

Resultados de los experimentos realizados. Utilizando SVM

A continuación, se muestran los resultados de todos los experimentos realizados sobre cada uno de los conjuntos de datos para validar el método propuesto de el enmascarado de la información textual de la noticia para la detección automática de noticias falsas; donde $DV-MA$ es la vista distorsionada con múltiples asteriscos y $DV-SA$ es la vista distorsionada con simples asteriscos. Utilizando una variación del parámetro K que va desde K= [0, 100, 200, ..., 1000] teniendo en cuenta para su extracción: las palabras más frecuentes del idioma, las palabras más frecuentes extraídas del vocabulario del propio corpus y las palabras más frecuentes teniendo en cuenta el FCE calculado del propio vocabulario del corpus.

A.1 Conjunto de Datos: MEX-A3T (Español)

Para el conjunto de datos (MEX-A3T); presentamos a continuación todos los experimentos realizados entrenando y probando sobre nuestro conjunto de validación en la siguiente subsección: Etapa de Validación.

A partir del mejor resultado obtenido en esta fase de experimentación; nos referimos al resultado o configuración mostrado en la Tabla: B.3, entonces evaluamos nuestro modelo para el conjunto de prueba. Los resultados sobre esta evaluación los mostramos en la subsección: Etapa de Evaluación y comparación con el estado del arte; donde como bien lo expresa el título del acápite, con nuestro mejor modelo comparamos nuestro resultado con el estado del arte para este conjunto de datos.

A.1.1 Etapa de Validación

La validación de los experimentos se llevó a cabo sobre una división del 80-20, para las particiones de entrenamiento y validación respectivamente.

MEX-A3T K=0	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
DV-MA	0.7862	0.7711	0.8030	0.8087	0.0263	0.3244	**0.8108**	**0.7925**	0.7947	0.7693	0.8026	0.7763
DV-SA	0.0000	0.3096	0.1000	0.3625	0.7000	0.6909	0.0263	0.3244	0.1000	0.3625	0.6000	0.6421

Tabla A.1: Utilizando el enmascaramiento total del texto (K=0).

MEX-A3T DV-MA_Estilo	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	0.8212	0.7990	0.8176	0.7805	0.8199	0.7793
K=100	0.8000	0.7858	0.7467	0.7176	0.7518	0.7423	0.0000	0.3096	0.5045	0.5814	0.6718	0.6834
K=200	0.8056	0.7936	0.7500	0.7167	0.7536	0.7499	0.0000	0.3096	0.4860	0.5763	0.6560	0.6817
K=300	0.7943	0.7865	0.7632	0.7316	0.7536	0.7499	0.0000	0.3096	0.4860	0.5763	0.6452	0.6739
K=400	0.8028	0.7937	0.7582	0.7236	0.7391	0.7352	0.0000	0.3096	0.5273	0.6031	0.6116	0.6502
K=500	0.8028	0.7937	.7517	0.7540	0.7500	0.7500	0.0000	0.3096	0.5273	0.6031	0.6230	0.6933
K=600	0.8252	0.8157	0.7651	0.7403	0.7576	0.7645	0.0000	0.3096	0.5273	0.6031	0.6400	0.6669
K=700	0.8252	0.8157	0.7891	0.7706	0.7692	0.7790	0.0000	0.3096	0.5273	0.6031	0.6560	0.6817
K=800	0.8169	0.8085	0.7808	0.7634	0.7597	0.7715	0.0000	0.3096	0.5225	0.5967	0.6341	0.6661
K=900	**0.8333**	**0.8229**	0.7919	0.7700	0.7778	0.7639	0.0000	0.3096	0.4954	0.5790	0.6290	0.6591
K=1000	0.8194	0.8082	0.7651	0.7403	0.7302	0.7486	0.0000	0.3096	0.5133	0.5837	0.6667	0.6830

Tabla A.2: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes del idioma y DV-MA.

MEX-A3T DV-MA_Estilo	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	**0.8212**	**0.7990**	0.8000	0.7675	0.7943	0.7865	0.3838	0.5156	0.6822	0.6977	0.7413	0.7272
K=200	0.7919	0.7700	**0.8158**	**0.7912**	0.7576	0.7645	0.3636	0.4997	0.6825	0.7043	0.7552	0.7420
K=300	0.8082	0.7930	**0.8267**	**0.8068**	0.7258	0.7480	0.5739	0.6309	0.7500	0.7500	0.7692	0.7567
K=400	0.8163	0.8002	**0.8344**	**0.8139**	0.7167	0.7465	0.6825	0.7043	0.7391	0.7352	0.7832	0.7715
K=500	0.8056	0.7934	**0.8533**	**0.8365**	0.6607	0.7116	0.7519	0.7572	0.7972	0.7862	0.7945	0.7782
K=600	0.8276	0.8154	**0.8414**	**0.8301**	0.5849	0.6599	0.7862	0.7711	0.7947	0.7693	0.7973	0.7777
K=700	0.8276	0.8154	**0.8414**	**0.8301**	0.5577	0.6419	0.8054	0.7848	0.8212	0.7990	0.8188	0.7996
K=800	**0.8435**	**0.8298**	**0.8438**	**0.8298**	0.5294	0.6235	0.7974	0.7684	0.8105	0.7834	0.8054	0.7848
K=900	**0.8414**	**0.8301**	0.8356	0.8226	0.4694	0.5853	0.8228	0.7886	0.7947	0.7693	0.8133	0.7919
K=1000	0.8194	0.8082	0.8276	0.8154	0.4694	0.5853	**0.8280**	**0.7966**	0.8026	0.7763	0.8133	0.7919

Tabla A.3: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del propio corpus y DV-MA.

MEX-A3T	Palabras						Caracteres					
DV-MA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	**0.8112**	**0.8009**	0.7763	0.7465	0.7714	0.7645	0.0000	0.3096	0.2727	0.4625	0.5932	0.6408
K=200	**0.8143**	**0.8087**	0.7763	0.7465	0.7612	0.7647	0.0000	0.3096	0.2353	0.4439	0.5321	0.6096
K=300	0.7887	0.7790	**0.7943**	**0.7693**	0.7143	0.7339	0.0000	0.3096	0.2955	0.4792	0.5505	0.6249
K=400	0.7943	0.7865	**0.8212**	**0.7990**	0.6891	0.7236	0.1250	0.3802	0.3617	0.5123	0.6034	0.6543
K=500	**0.8112**	**0.8009**	0.8079	0.7841	0.7009	0.7375	0.1707	0.4064	0.6782	0.5534	0.7226	0.6775
K=600	0.8000	0.7858	**0.8212**	**0.7990**	0.6055	0.6708	0.1928	0.4191	0.4600	0.5730	0.6325	0.6775
K=700	0.7973	0.7777	**0.8054**	**0.7848**	0.5243	0.6172	0.2326	0.4389	0.5333	0.6200	0.6441	0.6857
K=800	**0.8027**	**0.7854**	0.8000	0.7770	0.4375	0.5653	0.2697	0.4572	0.5556	0.6314	0.7179	0.6693
K=900	0.8000	0.7770	**0.8054**	**0.7848**	0.4211	0.5552	0.2889	0.4686	0.5818	0.6489	0.6325	0.6775
K=1000	0.7919	0.7700	**0.8027**	**0.7854**	0.3871	0.5343	0.3918	0.5273	0.6071	0.6661	0.6325	0.6775

Tabla A.4: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del FCE del propio corpus y DV-MA.

MEX-A3T	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.7516	0.7062	0.7746	0.7642	0.7626	0.7572	0.7862	0.7711	0.7808	0.7634	**0.7917**	**0.7786**
K=200	0.7532	0.7156	0.7887	0.7790	0.8028	0.7937	0.7862	0.7711	0.7945	0.7782	**0.8188**	**0.7996**
K=300	0.7613	0.7225	0.8219	0.8078	0.8112	0.8009	0.8027	0.7854	0.8219	0.8078	**0.8267**	**0.8068**
K=400	0.7692	0.7294	**0.8252**	**0.8157**	0.8056	0.7934	0.7887	0.7790	0.8056	0.7934	0.8108	0.7925
K=500	0.7763	0.7465	0.8169	0.8085	0.7972	0.7862	0.8000	0.7858	0.8163	0.8002	**0.8400**	**0.8216**
K=600	0.7662	0.7306	0.8112	0.8009	0.8056	0.7934	0.8194	0.8082	0.8138	0.8006	**0.8299**	**0.8150**
K=700	0.7815	0.7544	0.8194	0.8082	0.8219	0.8078	0.7801	0.7718	**0.8310**	**0.8232**	0.8219	0.8078
K=800	0.7763	0.7465	0.8194	0.8082	0.8028	0.7937	0.7887	0.7790	0.8252	0.8157	**0.8276**	**0.8154**
K=900	0.7712	0.7386	0.8356	0.8226	0.8194	0.8082	0.8138	0.8006	0.8392	0.8304	**0.8435**	**0.8298**
K=1000	0.7815	0.7544	0.8299	0.8150	**0.8414**	**0.8301**	0.8056	0.7934	0.8082	0.7930	0.8218	0.8078

Tabla A.5: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas idioma español y DV-SA.

MEX-A3T	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.7771	0.7364	0.8027	0.7854	**0.8112**	**0.8009**	0.7763	0.7465	0.8054	0.7848	0.8082	0.7930
K=200	0.7867	0.7622	0.7755	0.7558	0.7861	0.7647	0.7838	0.7629	**0.7891**	**0.7706**	0.7692	0.7567
K=300	0.7919	0.7700	0.7801	0.7718	0.7500	0.7500	0.7891	0.7706	**0.7945**	**0.7782**	0.7887	0.7790
K=400	**0.8158**	**0.7912**	0.8055	0.8012	0.7681	0.7647	0.8180	0.7925	0.8028	0.7937	0.8000	0.7939
K=500	**0.8366**	**0.8133**	0.8000	0.7939	0.7660	0.7570	0.8299	0.8150	0.8219	0.8078	0.8169	0.8055
K=600	0.8212	0.7990	0.8227	0.8159	0.7606	0.7495	0.8299	0.8150	**0.8356**	**0.8226**	0.8276	0.8154
K=700	**0.8497**	**0.8282**	0.8058	0.8014	0.7591	0.7498	0.8378	0.8221	0.8299	0.8150	0.8276	0.8154
K=800	**0.8571**	**0.8354**	0.7943	0.7865	0.7660	0.7570	0.8163	0.8002	0.8369	0.8307	0.8143	0.8087
K=900	**0.8571**	**0.8354**	0.8028	0.7937	0.7606	0.7495	0.8456	0.8293	0.8310	0.8232	0.8286	0.8234
K=1000	0.8477	0.8288	0.8085	0.8012	0.7746	0.7642	**0.8514**	**0.8370**	0.81160	0.8088	0.7941	0.7941

Tabla A.6: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del vocabulario del propio corpus y DV-SA.

MEX-A3T	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.7632	0.7316	0.8219	0.8078	0.7919	0.7786	0.7891	0.7706	0.8163	0.8002	0.8163	0.8002
K=200	0.7552	0.7420	0.8276	0.8154	0.7801	0.7718	0.7832	0.7715	0.8082	0.7930	0.8082	0.7930
K=300	0.7917	0.7786	0.7972	0.7862	0.7972	0.7862	0.7945	0.7782	0.8138	0.8006	0.8163	0.8002
K=400	0.7917	0.7786	0.8276	0.8154	0.7626	0.7572	0.8219	0.8078	0.8356	0.8226	0.8188	0.7996
K=500	0.7832	0.7715	0.8219	0.8078	0.7714	0.7645	0.8252	0.8157	0.8276	0.8154	0.8243	0.8073
K=600	0.8028	0.7937	0.8000	0.7858	0.7660	0.7570	0.8163	0.8002	0.8243	0.8073	0.8133	0.7919
K=700	0.7972	0.7862	0.7945	0.7782	0.7626	0.7572	0.7945	0.7782	0.8243	0.8073	0.8212	0.7990
K=800	0.7917	0.7786	0.7832	0.7715	0.7591	0.7573	0.8082	0.7930	0.8212	0.7990	0.8158	0.7912
K=900	0.8000	0.7858	0.8000	0.7858	0.7368	0.7425	0.8163	0.8002	0.8289	0.8061	0.8182	0.7904
K=1000	0.8082	0.7930	0.8000	0.7858	0.7287	0.7420	0.8344	0.8139	0.8344	0.8139	0.8258	0.7975

Tabla A.7: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del FCE del propio vocabulario del corpus y DV-SA.

MEX-A3T	Palabras								Caracteres			
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.7153	0.7132	0.6984	0.7191	0.1250	0.3802	0.7832	0.7715	0.7534	0.7339	**0.7922**	**0.7605**
K=200	0.7143	0.7056	0.6720	0.6865	0.1928	0.4191	**0.7917**	**0.7786**	0.7755	0.7558	0.7815	0.7544
K=300	0.7143	0.7056	0.6825	0.7043	0.1707	0.4064	**0.8267**	**0.8068**	0.7919	0.7700	0.7792	0.7455
K=400	0.7143	0.7056	0.7132	0.7272	0.1928	0.4191	**0.8299**	**0.8150**	0.8054	0.7848	0.8052	0.7755
K=500	0.7183	0.7053	0.7031	0.7196	0.1928	0.4191	**0.8027**	**0.7854**	0.7867	0.7622	0.7974	0.7684
K=600	0.7138	0.7053	0.6829	0.7106	0.1707	0.4064	**0.7945**	**0.7782**	0.7838	0.7629	0.7843	0.7535
K=700	0.7183	0.7053	0.6774	0.7036	0.1481	0.3934	0.7887	0.7790	**0.7891**	**0.7706**	0.7763	0.7465
K=800	0.7183	0.7053	0.6774	0.7036	0.1707	0.4064	0.7660	0.7570	0.7891	0.7706	**0.7895**	**0.7414**
K=900	0.7143	0.7056	0.6557	0.6879	0.1250	0.3802	0.7681	0.7647	**0.7919**	**0.7700**	0.7619	0.7410
K=1000	0.7194	0.7131	0.6333	0.6719	0.1481	0.3934	0.7591	0.7573	**0.7808**	**0.7634**	0.7483	0.7261

Tabla A.8: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes extraídas idioma Español y DV-MA.

MEX-A3T	Palabras								Caracteres			
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.7059	0.7059	0.7132	0.7272	0.0519	0.3388	0.7361	0.7196	0.7550	0.7246	**0.7763**	**0.7465**
K=200	0.6912	0.6912	0.7187	0.7344	0.1250	0.3802	**0.8000**	**0.7858**	0.7682	0.7395	0.7532	0.7156
K=300	0.6818	0.6909	0.7328	0.7423	0.2558	0.4559	**0.7857**	**0.7792**	0.7671	0.7486	0.7733	0.7473
K=400	0.7023	0.7128	0.7231	0.7348	0.3871	0.5343	0.7746	0.7642	0.7586	0.7415	**0.7891**	**0.7706**
K=500	0.7164	0.7205	0.7187	0.7344	0.4694	0.5853	0.7714	0.7645	0.7671	0.7486	**0.7838**	**0.7629**
K=600	0.7068	0.7131	0.7442	0.7567	0.5243	0.6172	**0.7755**	**0.7558**	0.7703	0.7480	0.7703	0.7480
K=700	0.7368	0.7425	0.7500	0.7639	0.5385	0.6264	**0.7943**	**0.7865**	0.7639	0.7491	0.7746	0.7642
K=800	0.7368	0.7425	0.7692	0.7790	0.5524	0.6355	**0.8029**	**0.8015**	0.7801	0.7718	0.7943	0.7865
K=900	0.74394	0.7498	0.7692	0.7790	0.5660	0.6045	**0.8112**	**0.8009**	0.7972	0.7862	0.8085	0.8012
K=1000	0.7368	0.7425	0.7692	0.7790	0.5794	0.6534	**0.8085**	**0.8012**	0.7972	0.7872	0.8058	0.8014

Tabla A.9: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes extraídas del vocabulario del propio corpus y DV-MA.

MEX-A3T	Palabras								Caracteres			
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.6912	0.6912	0.6719	0.6901	0.1250	0.3802	**0.7972**	**0.7868**	0.7483	0.7261	0.7922	0.7605
K=200	0.6917	0.6984	0.5882	0.6340	0.1928	0.4191	**0.8514**	**0.8370**	0.7778	0.7639	0.7600	0.7325
K=300	0.7121	0.7203	0.5833	0.6272	0.1707	0.4064	0.6935	0.7184	**0.7808**	**0.7634**	0.7785	0.7551
K=400	0.7015	0.7058	0.5641	0.6175	0.1250	0.3802	0.5688	0.6402	**0.7808**	**0.7634**	0.7703	0.7480
K=500	0.6866	0.6911	0.5641	0.6175	0.0519	0.3388	0.6595	0.4677	**0.7586**	**0.7415**	0.7568	0.7332
K=600	0.6963	0.6985	0.5641	0.6175	0.0263	0.3244	0.1928	0.4191	**0.7671**	**0.7486**	0.7517	0.7254
K=700	0.7286	0.7203	0.5000	0.5750	0.1013	0.3667	0.1481	0.3934	0.7338	0.7278	**0.7534**	**0.7339**
K=800	0.7101	0.7058	0.5091	0.5879	0.1707	0.4064	0.1250	0.3802	0.7246	0.7205	**0.7500**	**0.7344**
K=900	0.6963	0.6985	0.4860	0.5763	0.1013	0.3667	0.1250	0.3808	0.6866	0.6911	**0.7273**	**0.7125**
K=1000	0.6917	0.6984	0.4571	0.5579	0.1250	0.3808	0.1250	0.3802	0.6615	0.6758	**0.7133**	**0.6977**

Tabla A.10: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes extraídas del FCE del vocabulario del propio corpus y DV-MA.

MEX-A3T	Palabras								Caracteres			
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.6917	0.6984	0.7176	0.7276	0.3871	0.5343	**0.8138**	**0.8006**	0.7949	0.7595	0.7949	0.7595
K=200	0.6290	0.6591	0.6718	0.6834	0.5437	0.6328	**0.8322**	**0.8145**	0.7974	0.7684	0.7949	0.7595
K=300	0.5882	0.6340	0.6667	0.6762	0.5472	0.6290	**0.8400**	**0.8216**	0.8000	0.7675	0.7949	0.7595
K=400	0.5517	0.6092	0.7164	0.7205	0.6491	0.6980	**0.8322**	**0.8145**	0.8182	0.7904	0.8129	0.7825
K=500	0.5439	0.6074	0.7031	0.7196	0.5049	0.6015	**0.8212**	**0.7990**	0.7922	0.7605	0.8052	0.7755
K=600	0.4954	0.5790	0.6560	0.6817	0.4950	0.5984	**0.8182**	**0.7904**	0.7974	0.7684	0.8129	0.7825
K=700	0.4771	0.5637	0.6050	0.6489	0.4082	0.5374	**0.8000**	**0.7770**	0.7815	0.7544	0.8000	0.7675
K=800	0.4909	0.5726	0.5882	0.6340	0.3878	0.5215	**0.8000**	**0.7770**	0.7843	0.7535	0.7949	0.7595
K=900	0.4771	0.5637	0.5763	0.6258	0.3838	0.5156	0.7785	0.7551	**0.7821**	**0.7445**	0.7742	0.7375
K=1000	0.4360	0.5547	0.5641	0.6175	0.3469	0.4896	0.7534	0.7339	0.7600	0.7325	**0.7692**	**0.7294**

Tabla A.11: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del idioma y DV-SA.

MEX-A3T	Palabras						Caracteres					
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.5968	0.6295	0.6349	0.6599	0.5225	0.5967	**0.8133**	**0.7919**	0.7550	0.7246	0.7712	0.7386
K=200	0.5263	0.5923	0.6080	0.6373	0.5586	0.6271	**0.8378**	**0.8221**	0.7532	0.7156	0.7484	0.7075
K=300	0.5179	0.5902	0.6016	0.6364	0.5370	0.6161	**0.7862**	**0.7711**	0.7600	0.7325	0.7550	0.7246
K=400	0.5045	0.5814	0.5902	0.6284	0.5045	0.5814	**0.7917**	**0.7786**	0.7517	0.7254	0.7785	0.7551
K=500	0.4630	0.5547	0.6000	0.6421	0.5047	0.5917	**0.8000**	**0.7858**	0.7733	0.7473	0.7682	0.7395
K=600	0.4630	0.5547	0.6050	0.6489	0.5047	0.5919	**0.8054**	**0.7848**	0.7792	0.7545	0.7843	0.7535
K=700	0.4771	0.5637	0.5983	0.6475	0.5283	0.6135	0.7891	0.7706	0.7867	0.7622	**0.8000**	**0.7770**
K=800	0.4771	0.5637	0.6034	0.6543	0.5818	0.6489	**0.8243**	**0.8073**	0.8133	0.7919	0.7945	0.7782
K=900	0.4860	0.5763	0.6446	0.6799	0.5714	0.6357	**0.8378**	**0.8221**	0.8052	0.7755	0.7947	0.7693
K=1000	0.5000	0.5854	0.6154	0.6625	0.5455	0.6184	**0.8322**	**0.8145**	0.8312	0.8054	0.8133	0.7919

Tabla A.12: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes extraídas del vocabulario del propio corpus y DV-SA.

MEX-A3T	Palabras						Caracteres					
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.7536	0.7499	0.7438	0.7693	0.3263	0.3244	**0.8212**	**0.7990**	0.8176	0.7805	0.8199	0.7793
K=100	0.4486	0.5455	0.6190	0.6451	0.4000	0.5390	**0.8456**	**0.8293**	0.8387	0.8125	0.8129	0.7825
K=200	0.4118	0.5294	0.5345	0.5942	0.4078	0.5234	**0.8533**	**0.8365**	0.8333	0.8046	0.8153	0.7816
K=300	0.3673	0.5055	0.5128	0.5725	0.4423	0.5485	**0.8591**	**0.8442**	0.8077	0.7745	0.8205	0.7896
K=400	0.3043	0.4744	0.5133	0.5837	0.4571	0.5579	**0.8421**	**0.8211**	0.8333	0.8046	0.8153	0.7816
K=500	0.3043	0.4744	0.4528	0.5517	0.3838	0.5156	**0.8356**	**0.8226**	0.8153	0.7816	0.7974	0.7684
K=600	0.2667	0.4520	0.4444	0.5393	0.3673	0.5055	0.8000	0.7858	**0.8077**	**0.7745**	0.7785	0.7551
K=700	0.2472	0.4405	0.3922	0.5137	0.3299	0.4792	0.7801	0.7718	**0.7974**	**0.7684**	0.7600	0.7325
K=800	0.2472	0.4405	0.3800	0.5098	0.3333	0.4848	0.7606	0.7495	**0.7815**	**0.7544**	0.7671	0.7486
K=900	0.2069	0.4170	0.3960	0.5197	0.3333	0.4848	0.7518	0.7423	**0.7550**	**0.7246**	0.7397	0.7191
K=1000	0.2069	0.4170	0.4000	0.5256	0.3505	0.4953	0.7429	0.7351	**0.7483**	**0.7261**	0.7273	0.7125

Tabla A.13: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes extraídas del FCE del vocabulario del propio corpus y DV-SA.

A.1.2 Etapa de Evaluación y comparación con el Estado del Arte

El mejor enfoque obtenido a partir de la etapa de validación, es el que se evalúa en el conjunto de prueba, este último resultado es el que se reporta para este conjunto de datos y el cual comparamos con el estado del arte.

Model_MEX-A3T	Fake	True	F1-macro	Accuracy
Idiap-UAM-1	**0.8444**	**0.8688**	**0.8566**	**0.8576**
Idiap-UAM-2	0.8406	0.8599	0.8502	0.8508
Ares	0.8188	0.8151	0.8169	0.8169
CIMAT-1	0.7943	0.8117	0.8030	0.8034
Nuestro modelo: Random Forest, DV-SA, Modelo basado en Estilo, K=900, palabras más frecuentes del corpus Unigramas de palabras	0.7931	0.8000	0.7966	0.7966
Baseline (BoW-RF)	0.7850	0.7879	0.7864	0.7864
Intensos-2	0.7703	0.7883	0.7793	0.7797
Nuestro modelo: SVM, DV-SA, Modelo basado en Estilo, K=900, palabras más frecuentes del corpus Unigramas de palabras	0.7619	0.7635	0.7627	0.7627
Intensos-1	0.7597	0.7376	0.7487	0.7492
Baseline (INGEOTEC)	0.7596	0.7723	0.7659	0.7661
ITCG-SD	0.7464	0.7771	0.7617	0.7627

Tabla A.14: Comparación MEX-A3T con el estado del arte

Modelo	Accuracy
SVM (BoW)	0.7152
Random Forest (BoW)	0.7627
Nuestro Modelo: SVM, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (Unigramas de Palabras)	0.7627
Nuestro Modelo: Random Forest, DV-SA_Estilo, K=900, palabras_frecuentes_corpus, (Unigramas de Palabras)	**0.7966**

Tabla A.15: Comparación nuestro modelo con el MEX-A3T del artículo original. Resultados en el conjunto de prueba en términos del Accuracy

A.2 Conjunto de Datos: RAW-CovidES (Español)

Según el artículo de referencia realizan un 5-cross-validation con una partición 80-20 para entrenamiento y prueba respectivamente y reportan la medida promedio.

Para nuestro enfoque, Realizamos la siguiente distribución de los datos: 60-20-20, para entrenamiento, validación y prueba respectivamente, realizando el proceso de división de forma aleatoria cinco veces, obteniendo así cinco colecciones independientes: De igual forma realizamos las cinco corridas y reportamos la medida promedio, solo que validamos primero nuestro modelo y la mejor representación obtenida es la que evaluamos en el conjunto de prueba para reportar y compararnos con el estado del arte.

A.2.1 Etapa de validación

En esta fase de validación, la mejor configuración obtenida, la podemos ver en la Tabla: B.10.

RAW-CovidES K=0	Palabras								Caracteres			
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6803**	0.5040	0.6535	0.2630	0.5305
DV-MA	0.1083	0.4550	**0.4857**	**0.6669**	0.0891	0.4476	0.0000	0.4009	0.0000	0.4009	0.0000	0.4009
DV-SA	0.0000	0.4009	0.0000	0.4009	0.0000	0.4009	0.0000	0.4009	0.0000	0.4009	0.0000	0.4009

Tabla A.16: Utilizando el enmascaramiento total del Texto (K=0).

RAW-CovidES DV-MA_Estilo palabras_frecuentes_idioma	Palabras								Caracteres			
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.4252	0.6127	**0.5382**	**0.6957**	0.0263	0.4157	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=200	**0.5244**	**0.6809**	0.4551	0.6477	0.0263	0.4117	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=300	**0.4699**	**0.6474**	0.4266	0.6331	0.0263	0.4117	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=400	**0.5003**	**0.6617**	0.3821	0.6070	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=500	**0.4950**	**0.6589**	0.3874	0.6131	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=600	**0.5502**	**0.6927**	0.3745	0.6082	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=700	**0.5357**	**0.6831**	0.3096	0.5700	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=800	**0.6165**	**0.7341**	0.3539	0.5961	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=900	**0.6064**	**0.7270**	0.3333	0.5840	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=1000	**0.6063**	**0.7272**	0.2664	0.5449	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009

Tabla A.17: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes del idioma y DV-MA.

RAW-CovidES	Palabras						Caracteres					
DV-MA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	**0.6925**	**0.7786**	0.3386	0.5809	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=200	**0.7198**	**0.7908**	0.2113	0.5165	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=300	**0.7057**	**0.7738**	0.1937	0.5101	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=400	**0.7084**	**0.7756**	0.1637	0.4860	0.0263	0.4085	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=500	**0.7021**	**0.7630**	0.1637	0.4860	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=600	**0.7477**	**0.7965**	0.1637	0.4860	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=700	**0.7299**	**0.7931**	0.1637	0.4860	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=800	**0.7263**	**0.7982**	0.1420	0.4735	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=900	**0.7544**	**0.8104**	0.1420	0.4735	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=1000	**0.7757**	**0.8263**	0.1420	0.4735	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009

Tabla A.18: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del vocabulario del propio corpus y DV-MA.

RAW-CovidES	Palabras						Caracteres					
DV-MA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.4415	0.6374	**0.4550**	**0.6466**	0.0000	0.4009	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=200	**0.5157**	**0.6773**	0.3001	0.5617	0.0263	0.4126	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=300	**0.5892**	**0.7123**	0.2619	0.5430	0.0000	0.3937	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=400	**0.5860**	**0.7062**	0.1669	0.4839	0.0000	0.3937	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=500	**0.5399**	**0.6761**	0.1694	0.4931	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=600	**0.5312**	**0.6624**	0.1420	0.4735	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=700	**0.5997**	**0.7040**	0.1444	0.4749	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=800	**0.6055**	**0.7110**	0.1444	0.4749	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=900	**0.5465**	**0.6727**	0.0891	0.4435	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009
K=1000	**0.5465**	**0.6727**	0.0891	0.4435	0.0557	0.4250	0.0000	0.4009	0.0000	.4009	0.0000	0.4009

Tabla A.19: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del FCE del vocabulario del propio corpus y DV-MA.

RAW-CovidES	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6803**	0.5040	0.6535	0.2630	0.5305
K=100	0.0000	0.4009	0.2125	0.5181	**0.6547**	**0.7623**	0.0000	0.4009	0.2417	0.5346	0.4807	0.6622
K=200	0.0000	0.4009	0.3529	0.5965	**0.6580**	**0.7592**	0.0000	0.4009	0.2591	0.5420	0.4427	0.6293
K=3000.	0.0000	0.4009	0.3307	0.5865	**0.6275**	**0.7395**	0.0000	0.4009	0.3101	0.5744	0.4727	0.6400
K=400	0.0000	0.4009	0.3521	0.5990	**0.5295**	**0.6731**	0.0000	0.4009	0.3906	0.6224	0.5233	0.6712
K=500	0.0000	0.4009	0.3712	0.6107	0.4777	0.6500	0.0000	0.4009	0.3915	0.6228	**0.5117**	**0.6676**
K=600	0.0000	0.4009	0.3682	0.6020	0.5024	0.6640	0.0000	0.4009	0.4499	0.5446	**0.6250**	**0.7322**
K=700	0.0000	0.4009	0.4051	0.6202	0.4842	0.6531	0.0000	0.4009	0.4933	0.6803	**0.6593**	**0.7633**
K=800	0.0000	0.4009	0.4561	0.6438	0.4861	0.6586	0.0000	0.4009	0.5815	0.7138	**0.7465**	**0.8145**
K=900	0.0000	0.4009	0.4341	0.6307	0.3876	0.5997	0.0000	0.4009	0.5506	0.6951	**0.7368**	**0.8120**
K=1000	0.0000	0.4009	0.4361	0.6314	0.3270	0.5634	0.0000	0.4009	0.5854	0.7164	**0.7874**	**0.8480**

Tabla A.20: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes del idioma Inglés y DV-SA.

RAW-CovidES	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.0000	0.4009	0.4470	0.6445	0.2498	0.5363	0.0000	0.4009	**0.7472**	**0.8259**	0.7401	0.8127
K=200	0.0000	0.4009	0.4169	0.6322	0.1033	0.4594	0.2096	0.5161	**0.7191**	**0.7901**	0.6492	0.7370
K=300	0.0000	0.4009	0.4339	0.6425	0.1012	0.4510	0.4686	0.6636	**0.7476**	**0.8094**	0.6671	0.7711
K=400	0.0000	0.4009	0.4142	0.6328	0.1012	0.4510	0.5734	0.7111	**0.7186**	**0.7894**	0.6712	0.7730
K=500	0.0000	0.4009	0.3567	0.6065	0.0795	0.4385	0.5815	0.7136	**0.6926**	**0.7725**	0.6488	0.7553
K=600	0.0000	0.4009	0.4089	0.6258	0.795	0.4385	0.6596	0.7325	**0.7095**	**0.7706**	0.6455	0.7437
K=700	0.1143	0.4652	0.3501	0.5961	0.0795	0.4385	**0.6813**	**0.7510**	0.6740	0.7507	0.6490	0.7548
K=800	0.2710	0.5536	0.3448	0.5937	0.0557	0.4250	**0.6509**	**0.7237**	0.6253	0.7148	0.5408	0.6928
K=900	0.42500	.6382	0.3782	0.6137	0.0557	0.4250	**0.7075**	**0.7504**	0.6436	0.7305	0.5569	0.6987
K=1000	0.5552	0.7118	0.3349	0.5853	0.5557	0.4250	**0.7075**	**0.7504**	0.6069	0.6994	0.4956	0.6635

Tabla A.21: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del vocabulario del propio corpus y DV-SA.

RAW-CovidES	Palabras								Caracteres			
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.0000	0.4009	0.5167	0.6833	**0.6640**	**0.7594**	0.0000	0.4009	0.4308	0.6277	0.5950	0.7180
K=200	0.0000	0.4009	**0.5835**	**0.7140**	0.3097	0.5741	0.0000	0.4009	0.5704	0.7063	0.5675	0.6891
K=300	0.0000	0.4009	0.5525	0.6981	0.0807	0.4463	0.0557	0.4258	0.7025	0.7761	**0.7227**	**0.7863**
K=400	0.0000	0.4009	0.5902	0.7269	0.0807	0.4463	0.0557	0.4258	**0.6585**	**0.7427**	0.6338	0.7183
K=500	0.0294	0.4173	0.5295	0.6922	0.0557	0.4291	0.0557	0.4258	0.5436	0.6654	**0.5665**	**0.6824**
K=600	0.0294	0.4173	0.4547	0.6505	0.0557	0.4291	0.0557	0.4258	0.5346	0.6638	**0.4582**	**0.6323**
K=700	0.0294	0.4173	0.4253	0.6319	0.0557	0.4291	0.1896	0.5015	**0.5577**	**0.6865**	0.4154	0.5995
K=800	0.0294	0.4173	0.3638	0.6027	0.0557	0.4291	0.1319	0.4692	**0.4554**	**0.6271**	0.3050	0.5329
K=900	0.0294	0.4173	0.3260	0.5800	0.0557	0.4250	0.1607	0.4815	**0.4108**	**0.5987**	0.2702	0.5117
K=1000	0.0294	0.4173	0.3269	0.5846	0.0557	0.4250	0.1857	0.4956	**0.4012**	**0.5892**	0.2702	0.5117

Tabla A.22: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del FCE del vocabulario del propio corpus y DV-SA.

RAW-CovidES	Palabras								Caracteres			
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.2390	0.5199	0.0557	0.4250	0.0557	0.4250	0.0294	0.4071	**0.4464**	**0.6201**	0.2874	0.5363
K=200	0.2478	0.5243	0.0557	0.4291	0.0557	0.4250	0.0294	0.4142	**0.4156**	**0.6051**	0.2874	0.5363
K=300	0.2244	0.5107	0.0557	0.4291	0.0557	0.4250	0.0000	0.3977	**0.3081**	**0.5525**	0.2583	0.5159
K=400	0.1947	0.4942	0.0557	0.4291	0.0557	0.4250	0.0000	0.3977	0.2641	0.5268	**0.2780**	**0.5275**
K=500	0.1749	0.5482	0.0557	0.4291	0.0557	0.4250	0.0000	0.3977	0.2428	0.5183	**0.2528**	**0.5045**
K=600	0.1817	0.4860	0.0557	0.4291	0.0557	0.4250	0.0000	0.3977	0.1870	0.4786	**0.2755**	**0.5218**
K=700	0.1523	0.4695	0.0891	0.4476	0.0557	0.4250	0.0000	0.4009	0.1607	0.4598	**0.2755**	**0.5218**
K=800	0.1817	0.4860	0.0914	0.4519	0.0557	0.4250	0.0000	0.4009	0.1473	0.4597	**0.2747**	**0.5214**
K=900	0.1817	0.4860	0.0914	0.4519	0.0557	0.4250	0.0000	0.4009	0.1654	0.4706	**0.2777**	**0.5270**
K=1000	0.1817	0.4860	0.0914	0.4519	0.0557	0.4250	0.0000	0.4009	0.1473	0.4597	**0.2617**	**0.5299**

Tabla A.23: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del idioma y DV-MA.

RAW-CovidES	Palabras								Caracteres			
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.1895	0.4904	0.0557	0.4322	0.0557	0.4250	0.1083	0.4509	0.3798	0.5666	**0.5792**	**0.6668**
K=200	0.1895	0.4904	0.0557	0.4322	0.0557	0.4250	0.0250	0.4041	0.3620	0.5682	**0.5207**	**0.6419**
K=300	0.2371	0.5180	0.0807	0.4463	0.0557	0.4250	0.0888	0.4476	0.3292	0.5503	**0.5064**	**0.6445**
K=400	0.1589	0.4735	0.1033	0.4594	0.0557	0.4250	0.0000	0.4009	0.1497	0.4637	**0.3973**	**0.5848**
K=500	0.1185	0.4531	0.1239	0.4714	0.0557	0.4250	0.0000	0.4009	0.1080	0.4434	**0.2384**	**0.5029**
K=600	0.1464	0.4693	0.1239	0.4714	0.0557	0.4250	0.0000	0.4009	0.1330	0.4498	**0.2694**	**0.5203**
K=700	0.1497	0.4750	0.1500	0.4832	0.0557	0.4250	0.0000	0.4009	0.0875	0.4430	**0.2186**	**0.4913**
K=800	0.1282	0.4591	**0.1295**	**0.4711**	0.0557	0.4291	0.0000	0.4009	0.0357	0.4206	0.1038	0.4333
K=900	0.1282	0.4624	**0.1558**	**0.4860**	0.0557	0.4250	0.0000	0.4009	0.0620	0.4354	0.1554	0.4661
K=1000	0.0750	0.4324	**0.1771**	**0.4944**	0.0557	0.4250	0.0000	0.4009	0.0357	0.4206	0.1038	0.4370

Tabla A.24: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del corpus y DV-MA.

RAW-CovidES	Palabras								Caracteres			
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.2390	0.5233	0.0557	0.4250	0.0557	0.4250	0.0294	0.4173	**0.3246**	**0.5626**	0.3037	0.5426
K=200	0.1727	0.4846	0.0557	0.4322	0.0557	0.4250	0.0000	0.4009	0.0738	0.4211	**0.2058**	**0.4898**
K=300	**0.1510**	**0.4827**	0.0557	0.4322	0.0557	0.4250	0.0000	0.4009	0.0000	0.3939	0.1255	0.4497
K=400	0.1239	0.4713	0.0557	0.4322	0.0557	0.4250	0.0000	0.4009	0.0000	0.4009	**0.7887**	**0.4528**
K=500	**0.1003**	**0.4577**	0.0807	0.4463	0.0557	0.4250	0.0000	0.4009	0.0263	0.4157	0.0726	0.4230
K=600	0.0776	0.4447	**0.1033**	**0.4594**	0.0557	0.4250	0.0000	0.4009	0.0000	0.4009	0.0726	0.4230
K=700	**0.1003**	**0.4577**	**0.1033**	**0.4594**	0.0557	0.4250	0.0000	0.4009	0.0000	0.4009	0.0500	0.4183
K=800	**0.0976**	**0.4533**	0.0807	0.4463	0.0557	0.4250	0.0000	0.4009	0.0000	0.4009	0.0250	0.4080
K=900	0.0750	0.4403	**0.1033**	**0.4594**	0.0557	0.4250	0.0000	0.4009	0.0000	0.4009	0.0000	0.3977
K=1000	0.0750	0.4403	**0.1033**	**0.4594**	0.0557	0.4250	0.0000	0.4009	0.0000	0.4009	0.0000	0.3977

Tabla A.25: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del FCE del vocabulario del propio corpus y DV-MA.

RAW-CovidES	Palabras								Caracteres			
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.0000	0.4009	0.0807	0.4463	0.0557	0.4250	0.1827	0.4865	**0.4699**	**0.6361**	0.3185	0.5554
K=200	0.0000	0.4009	0.0557	0.4322	0.0557	0.4250	0.1827	0.4865	**0.4521**	**0.6210**	0.3019	0.5452
K=300	0.0000	0.4009	0.0557	0.4291	0.0557	0.4250	0.0866	0.4865	**0.4632**	**0.6311**	0.3001	0.5654
K=400	0.0000	0.4009	0.0557	0.4291	0.0557	0.4250	0.0532	0.4198	**0.4114**	**0.5991**	0.2827	0.5380
K=500	0.0000	0.4009	0.0557	0.4291	0.0557	0.4250	0.0294	0.4063	**0.4114**	**0.5991**	0.2827	0.5380
K=600	0.0000	0.4009	0.0557	0.4291	0.0557	0.4250	0.0544	0.4245	**0.3986**	**0.5949**	0.2604	0.5249
K=700	0.0000	0.4009	0.0544	0.4245	0.0557	0.4250	0.0544	0.4244	**0.3575**	**0.5707**	0.2577	0.5198
K=800	0.0000	0.4009	0.0542	0.4212	0.0557	0.4250	0.0294	0.4103	**0.3386**	**0.5674**	0.2604	0.5249
K=900	0.0000	0.4009	0.0557	0.4258	0.0557	0.4250	0.0294	0.4103	**0.3386**	**0.5674**	0.2370	0.5112
K=1000	0.0000	0.4009	0.0557	0.4258	0.0557	0.4250	0.0294	0.4103	**0.3409**	**0.5689**	0.2372	0.5038

Tabla A.26: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del idioma y DV-SA.

RAW-CovidES	Palabras								Caracteres			
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.0000	0.4009	0.2951	0.5775	0.0795	0.4385	0.1017	0.4402	0.4799	0.6229	0.4353	0.6024
K=200	0.0000	0.4009	0.3262	0.5660	0.2539	0.5037	0.1536	0.4769	**0.5473**	**0.6682**	0.5200	0.6500
K=300	0.0000	0.4009	0.3512	0.5909	0.3325	0.5570	0.1536	0.4769	**0.4940**	**0.6417**	0.5423	0.6648
K=400	0.0000	0.4009	0.2606	0.5435	0.3180	0.5632	0.0750	0.4324	0.3932	0.5821	**0.5063**	**0.6454**
K=500	0.0000	0.4009	0.1930	0.4999	0.3284	0.5702	0.0000	0.3939	0.4425	0.6148	**0.5670**	**0.6849**
K=600	0.0000	0.4009	0.0976	0.4494	0.3104	0.5553	0.0250	0.4080	0.4050	0.5869	**0.5662**	**0.6879**
K=700	0.0000	0.4009	0.0750	0.4365	0.0332	0.5813	0.0000	0.3900	0.4094	0.5937	**0.6073**	**0.7149**
K=800	0.0000	0.4009	0.5130	0.4229	0.2846	0.5538	0.0000	0.3977	0.3464	0.5658	**0.5667**	**0.6881**
K=900	0.0000	0.4009	0.0513	0.4229	0.2513	0.5354	0.0000	0.3977	0.3679	0.5824	**0.5599**	**0.6890**
K=1000	0.0000	0.4009	0.0513	0.4229	0.1905	0.4993	0.0000	0.3977	0.2547	0.5146	**0.5266**	**0.6659**

Tabla A.27: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del corpus y DV-SA.

RAW-CovidES	Palabras								Caracteres			
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	0.3626	0.5811	0.0557	0.4250	0.0557	0.4250	**0.5551**	**0.6003**	0.5040	0.6535	0.2630	0.5305
K=100	0.0000	0.4009	0.1795	0.4996	0.0557	0.4250	0.0794	0.4315	**0.4904**	**0.6487**	0.2934	0.5472
K=200	0.0000	0.4009	0.1533	0.4878	0.0557	0.4250	0.0000	0.3939	**0.4499**	**0.6301**	0.3211	0.5529
K=300	0.0000	0.4009	0.1044	0.4599	0.0557	0.4312	0.0000	0.3977	**0.3119**	**0.5500**	0.3002	0.5521
K=400	0.0000	0.4009	0.0750	0.4435	0.0807	0.4463	0.0000	0.4008	0.2516	0.5192	**0.2621**	**0.5291**
K=500	0.0000	0.4009	0.1013	0.4583	0.0807	0.4425	0.0000	0.4008	0.2527	0.5240	**0.2548**	**0.5251**
K=600	0.0000	0.4009	0.0776	0.4447	0.0807	0.4463	0.0000	0.4008	0.1885	0.4864	**0.2548**	**0.5251**
K=700	0.0000	0.4009	0.0513	0.4299	0.0807	0.4463	0.0000	0.4008	0.1709	0.4800	**0.2396**	**0.5199**
K=800	0.0000	0.4009	0.0250	0.4150	0.1270	0.4698	0.0000	0.4008	0.1482	0.4670	**0.2658**	**0.5349**
K=900	0.0000	0.4009	0.0250	0.4150	0.0750	0.4403	0.0000	0.4008	0.1244	0.4535	**0.2419**	**0.5213**
K=1000	0.0000	0.4009	0.0000	0.4008	0.0513	0.4267	0.0000	0.4008	0.1039	0.4414	**0.2419**	**0.5213**

Tabla A.28: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del FCE del propio corpus y DV-SA.

A.2.2 Etapa de Evaluación y comparación con el Estado del Arte.

La mejor configuración obtenida en la etapa de validación, es evaluada entonces en el conjunto de prueba y son los resultados que reportamos en este acápite y comparamos con el estado del arte.

Model_RAW-CovidES	True			Fake			Accuracy	F1-macro
	Precision	Recall	F1	Precision	Recall	F1		
Baseline (Random)	0.551	0.549	0.548	0.498	0.500	0.497	0.526	0.522
Baseline (TF-IDF)	0.609	0.868	0.715	0.726	0.381	0.494	0.637	0.605
Full pipeline	**0.920**	**0.550**	**0.790**	**0.680**	**0.950**	**0.690**	**0.750**	**0.740**
Nuestro Modelo: SVM, DV-SA, Modelo basado en Estilo, K=1000, palabras frecuentes del idioma, pentagramas de caracteres	**0.776**	**0.896**	**0.914**	**0.959**	**0.517**	**0.665**	**0.825**	**0.789**

Tabla A.29: Comparación RAW-CovidES con el estado del arte

A.3 Conjunto de Datos: LIAR (Inglés)

Para este conjunto de datos, contamos con tres particiones: entrenamiento, validación y prueba. De igual forma que con los conjuntos de datos anteriores, entrenamos y probamos en el conjunto de validación, para de ahí obtener nuestra mejor configuración y esta última es la que evaluamos finalmente en el de prueba y el resultado obtenido es el que comparamos con el estado del arte.

A.3.1 Etapa de Validación

Dentro de todos los experimentos realizados el mejor modelo lo obtuvimos con la siguiente configuración, Tabla: A.33.

LIAR K=0	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
DV-MA	0.0405	0.3798	0.4565	0.5854	**0.6826**	**0.7406**	0.0022	0.3609	0.0075	0.3635	0.0119	0.3656
DV-SA	0.0171	0.3680	0.0305	0.3751	0.1030	0.4107	0.0253	0.3724	0.0322	0.3759	0.0759	0.3978

Tabla A.30: Utilizando el enmascarado total del texto (K=0).

LIAR	Palabras						Caracteres					
DV-MA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.4356	0.5655	0.6824	0.7365	**0.9256**	**0.9359**	0.2600	0.4908	0.4238	0.5767	0.5276	0.6329
K=200	0.4750	0.5870	0.7376	0.7804	**0.9474**	**0.9543**	0.3806	0.5527	0.5213	0.6296	0.6006	0.6779
K=300	0.4980	0.6005	0.7716	0.8080	**0.9628**	**0.9675**	0.4438	0.5842	0.5698	0.6568	0.6376	0.7021
K=400	0.5108	0.6080	0.7873	0.8206	**0.9664**	**0.9706**	0.4689	0.5979	0.5925	0.6699	0.6568	0.7163
K=500	0.5204	0.6140	0.8072	0.8368	**0.9700**	**0.9737**	0.4932	0.6105	0.6005	0.6773	0.6741	0.7300
K=600	0.5376	0.6261	0.8196	0.8470	**0.9738**	**0.9770**	0.5067	0.6183	0.6151	0.6862	0.6852	0.7375
K=700	0.5435	0.6302	0.8304	0.8557	**0.9749**	**0.9779**	0.5218	0.6251	0.6256	0.6935	0.6997	0.7496
K=800	0.5479	0.6331	0.8386	0.8625	**0.9763**	**0.9792**	0.5306	0.6292	0.6366	0.7006	0.7090	0.7564
K=900	0.5568	0.6388	0.8455	0.8687	**0.9784**	**0.9810**	0.5354	0.6347	0.6442	0.7081	0.7144	0.7620
K=1000	0.5621	0.6429	0.8522	0.8743	**0.9800**	**0.9824**	0.5392	0.6372	0.6492	0.7131	0.7185	0.7656

Tabla A.31: Utilizando un modelo basado en estilo, las K palabras más frecuentes de idioma y DV-MA.

LIAR	Palabras						Caracteres					
DV-MA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.4373	0.5703	0.7159	0.7626	**0.9418**	**0.9572**	0.3647	0.5437	0.4872	0.6130	0.5666	0.6587
K=200	0.5117	0.6098	0.7896	0.8222	**0.9691**	**0.9729**	0.4775	0.6034	0.5886	0.6709	0.6574	0.7199
K=300	0.5444	0.6305	0.8266	0.8537	**0.9815**	**0.9837**	0.5262	0.6283	0.6304	0.6990	0.7010	0.7513
K=400	0.5574	0.6384	0.8583	0.8792	**0.9843**	**0.9862**	0.5649	0.6500	0.6542	0.7139	0.7245	0.7690
K=500	0.5657	0.6434	0.8711	0.8900	**0.9872**	**0.9887**	0.5779	0.6570	0.6694	0.7552	0.7356	0.7786
K=600	0.5754	0.6492	0.8847	0.9014	**0.9892**	**0.9905**	0.5926	0.6661	0.6782	0.7320	0.7561	0.7898
K=700	0.5911	0.6613	0.8955	0.9103	**0.9894**	**0.9907**	0.6049	0.6744	0.6879	0.7400	0.7638	0.8003
K=800	0.5952	0.6635	0.9052	0.9186	**0.9899**	**0.9911**	0.6072	0.6753	0.6972	0.7460	0.7736	0.8083
K=900	0.5986	0.6682	0.9102	0.9228	**0.9906**	**0.9917**	0.6152	0.6818	0.7021	0.7503	0.7852	0.8177
K=1000	0.6074	0.6742	0.9168	0.9283	**0.9910**	**0.9921**	0.6200	0.6852	0.7106	0.7572	0.7885	0.8204

Tabla A.32: Utilizando un modelo basado en estilo, las K palabras más frecuentes de corpus y DV-MA.

LIAR	Palabras						Caracteres					
DV-MA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.3809	0.5437	0.6901	0.7440	**0.9281**	**0.9381**	0.2312	0.4768	0.4224	0.5778	0.5241	0.6344
K=200	0.4692	0.5865	0.7718	0.8088	**0.9645**	**0.9690**	0.4067	0.5669	0.5478	0.6492	0.6295	0.7022
K=300	0.4997	0.6015	0.8152	0.8439	**0.9754**	**0.9784**	0.4655	0.5962	0.5929	0.6744	0.6704	0.7294
K=400	0.5316	0.6207	0.8413	0.8655	**0.9801**	**0.9825**	0.5106	0.6214	0.6227	0.6944	0.7007	0.7522
K=500	0.5469	0.6309	0.8583	0.8796	**0.9841**	**0.9860**	0.5401	0.6373	0.6449	0.7101	0.7252	0.7710
K=600	0.5675	0.6461	0.8767	0.8950	**0.9874**	**0.9889**	0.5712	0.6537	0.6660	0.7242	0.7429	0.7847
K=700	0.5771	0.6527	0.8821	0.8992	**0.9881**	**0.9895**	0.5814	0.6598	0.6770	0.7316	0.7559	0.7950
K=800	0.5827	0.6582	0.8938	0.9091	**0.9894**	**0.9907**	0.5941	0.6676	0.6804	0.7338	0.7665	0.8034
K=900	0.5906	0.6634	0.9013	0.9153	**0.9896**	**0.9908**	0.5957	0.6690	0.6832	0.7372	0.7697	0.8066
K=1000	0.6001	0.6699	0.9141	0.9261	**0.9902**	**0.9914**	0.6020	0.6732	0.6903	0.7439	0.7744	0.8106

Tabla A.33: Utilizando un modelo basado en estilo, las K palabras más frecuentes de FCE del vocabulario del corpus y DV-MA.

LIAR	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.3552	0.5325	0.5811	0.6632	**0.7213**	**0.7686**	0.4867	0.5961	0.5440	0.6366	0.6044	0.6776
K=200	0.4382	0.5707	0.6623	0.7196	**0.7934**	**0.8252**	0.5165	0.6168	0.5917	0.6679	0.6640	0.7206
K=300	0.4843	0.5950	0.6990	0.7483	**0.8400**	**0.8634**	0.5349	0.6268	0.6159	0.6844	0.6911	0.7418
K=400	0.4995	0.6029	0.7206	0.7654	**0.8569**	**0.8773**	0.5471	0.6337	0.6831	0.6970	0.7047	0.7532
K=500	0.5088	0.6096	0.7404	0.7823	**0.8742**	**0.8921**	0.5564	0.6410	0.6410	0.7036	0.7184	0.7640
K=600	0.5236	0.6181	0.7567	0.7949	**0.8880**	**0.9036**	0.5671	0.6484	0.6566	0.7143	0.7289	0.7718
K=700	0.5335	0.6240	0.7665	0.8035	**0.9001**	**0.9138**	0.5847	0.6593	0.6614	0.7192	0.7351	0.7770
K=800	0.5428	0.6305	0.7792	0.8136	**0.9063**	**0.9191**	0.5865	0.6617	0.6681	0.7244	0.7414	0.7824
K=900	0.5488	0.6337	0.7858	0.8191	**0.9160**	**0.9274**	0.5862	0.6617	0.6702	0.7263	0.7446	0.7857
K=1000	0.5540	0.6372	0.7942	0.8259	**0.7225**	**0.9330**	0.5929	0.6647	0.6737	0.7298	0.7535	0.7928

Tabla A.34: Utilizando un modelo basado en estilo, las K palabras más frecuentes del idioma y DV-SA.

LIAR	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.3928	0.5519	0.6167	0.6916	**0.7665**	**0.8066**	0.4977	0.6052	0.5688	0.6524	0.6278	0.6965
K=200	0.5056	0.6078	0.7230	0.7680	**0.8696**	**0.8886**	0.5573	0.6419	0.6347	0.6974	0.7016	0.7511
K=300	0.5364	0.6238	0.7688	0.8052	**0.9145**	**0.9259**	0.5795	0.6548	0.6578	0.7167	0.7322	0.7760
K=400	0.5532	0.6345	0.8021	0.8324	**0.9335**	**0.9421**	0.5987	0.6685	0.6727	0.7271	0.7467	0.7876
K=500	0.5651	0.6425	0.8255	0.8515	**0.9489**	**0.9553**	0.6030	0.6701	0.6836	0.7353	0.7585	0.7969
K=600	0.5733	0.6465	0.8428	0.8657	**0.9562**	**0.9616**	0.6111	0.6757	0.6923	0.7421	0.7737	0.8094
K=700	0.5897	0.6589	0.8536	0.8747	**0.9631**	**0.9676**	0.6217	0.6838	0.7001	0.7481	0.7823	0.8160
K=800	0.5954	0.6631	0.8687	0.8872	**0.9694**	**0.9730**	0.6256	0.6875	0.7102	0.7559	0.7904	0.8227
K=900	0.5998	0.6674	0.8758	0.8932	**0.9745**	**0.9775**	0.6266	0.6892	0.7153	0.7602	0.7982	0.8287
K=1000	0.6049	0.6717	0.9154	0.8990	**0.9765**	**0.9793**	0.6293	0.6914	0.7203	0.7647	0.8029	0.8328

Tabla A.35: Utilizando un modelo basado en estilo, las K palabras más frecuentes del corpus y DV-SA.

LIAR	Palabras						Caracteres					
DV-SA_Estilo	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.2907	0.5018	0.5843	0.6690	**0.7241**	**0.7729**	0.4498	0.5836	0.5249	0.6298	0.5962	0.6788
K=200	0.4476	0.5775	0.6962	0.7487	**0.8402**	**0.8651**	0.5217	0.6225	0.6102	0.6837	0.6800	0.7370
K=300	0.4842	0.5973	0.7945	0.7901	**0.8904**	**0.9061**	0.5414	0.6344	0.6351	0.7008	0.7122	0.7613
K=400	0.5269	0.6196	0.7831	0.8172	**0.9202**	**0.9310**	0.5685	0.6501	0.6545	0.7155	0.7321	0.7766
K=500	0.5421	0.6293	0.8098	0.8390	**0.9385**	**0.9465**	0.5841	0.6590	0.6695	0.7263	0.7523	0.7922
K=600	0.5654	0.6442	0.8323	0.8570	**0.9503**	**0.9566**	0.5994	0.6694	0.6821	0.7345	0.7601	0.7989
K=700	0.5752	0.6513	0.8488	0.8707	**0.9568**	**0.9621**	0.6073	0.6755	0.6932	0.7431	0.7688	0.8058
K=800	0.5842	0.6574	0.8567	0.8773	**0.9631**	**0.9676**	0.6145	0.6813	0.6972	0.7467	0.7802	0.8146
K=900	0.5915	0.6628	0.8658	0.8853	**0.9683**	**0.9721**	0.6169	0.6839	0.7000	0.7496	0.7855	0.8197
K=1000	0.5962	0.6666	0.8763	0.8942	**0.9710**	**0.9745**	0.6185	0.6856	0.7028	0.7526	0.7896	0.8225

Tabla A.36: Utilizando un modelo basado en estilo, las K palabras más frecuentes del FCE del vocabulario del propio corpus y DV-SA.

LIAR	Palabras						Caracteres					
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.7712	0.8083	0.9474	0.9544	**0.9890**	**0.9903**	0.6732	0.7296	0.7744	0.8102	0.8457	0.8689
K=200	0.7678	0.8054	0.9392	0.9475	**0.9865**	**0.9881**	0.6614	0.7227	0.7608	0.8004	0.8290	0.8561
K=300	0.7656	0.8043	0.9362	0.9449	**0.9861**	**0.9878**	0.6495	0.7157	0.7552	0.7972	0.8233	0.8517
K=400	0.7625	0.8016	0.9319	0.9414	**0.9842**	**0.9861**	0.6462	0.7142	0.7472	0.7913	0.8129	0.8435
K=500	0.7560	0.7971	0.9331	0.9424	**0.9831**	**0.9851**	0.6367	0.7089	0.7393	0.7855	0.8089	0.8405
K=600	0.7570	0.7984	0.9290	0.9391	**0.9803**	**0.9827**	0.6282	0.7038	0.7311	0.7795	0.8075	0.8399
K=700	0.7556	0.7976	0.9261	0.9366	**0.9795**	**0.9820**	0.6225	0.7000	0.7320	0.7807	0.8026	0.8357
K=800	0.7556	0.7976	0.9234	0.9344	**0.9781**	**0.9808**	0.6170	0.6964	0.7332	0.7820	0.8018	0.8351
K=900	0.7541	0.7963	0.9202	0.9323	**0.9768**	**0.9797**	0.6158	0.6961	0.7270	0.7772	0.7975	0.8322
K=1000	0.7498	0.7936	0.9207	0.9322	**0.9752**	**0.9783**	0.6154	0.6968	0.7226	0.7750	0.7912	0.8277

Tabla A.37: Utilizando un modelo basado en contenido, las K palabras más frecuentes del idioma y DV-MA.

LIAR	Palabras						Caracteres					
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.7708	0.8083	0.9456	0.9529	**0.9882**	**0.9896**	0.6566	0.7214	0.7655	0.8046	0.8400	0.8647
K=200	0.7648	0.8049	0.9329	0.7423	**0.9825**	**0.9846**	0.6196	0.7002	0.7384	0.7864	0.813	0.8438
K=300	0.7608	0.8025	0.9286	0.9389	**0.9744**	**0.9776**	0.5880	0.6805	0.7250	0.7773	0.7976	0.8323
K=400	0.7606	0.8030	0.9224	0.9336	**0.9700**	**0.9738**	0.5653	0.6674	0.7058	0.7647	0.7826	0.8218
K=500	0.7576	0.8009	0.9169	0.9293	**0.9633**	**0.9680**	0.5492	0.6592	0.6894	0.7538	0.7721	0.8143
K=600	0.7596	0.8030	0.9091	0.9229	**0.9582**	**0.9638**	0.5403	0.6543	0.6799	0.7476	0.7614	0.8067
K=700	0.7585	0.830	0.9032	0.9182	**0.9526**	**0.9590**	0.5277	0.6467	0.6682	0.7404	0.7542	0.8020
K=800	0.7533	0.7995	0.8951	0.9116	**0.9472**	**0.9545**	0.5148	0.6413	0.6582	0.7344	0.7403	0.7925
K=900	0.7490	0.7970	0.8899	0.9074	**0.9409**	**0.9492**	0.5010	0.6330	0.6429	0.7250	0.7282	0.7842
K=1000	0.7449	0.7951	0.8832	0.9022	**0.9371**	**0.9461**	0.4890	0.6270	0.6315	0.7185	0.7207	0.7795

Tabla A.38: Utilizando un modelo basado en contenido, las K palabras más frecuentes del corpus y DV-MA.

LIAR	Palabras						Caracteres					
DV-MA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.7697	0.8064	0.9495	0.9561	**0.9903**	**0.9915**	0.6762	0.7314	0.7762	0.8114	0.8461	0.8689
K=200	0.7675	0.8060	0.9416	0.6465	**0.9859**	**0.9876**	0.6512	0.7177	0.7563	0.7969	0.8215	0.8502
K=300	0.7648	0.8047	0.9303	0.9401	**0.9824**	**0.9845**	0.6332	0.7077	0.7483	0.7919	0.8102	0.8410
K=400	0.7663	0.8067	0.9234	0.9343	**0.9759**	**0.9789**	0.6122	0.6957	0.7341	0.7820	0.8022	0.8353
K=500	0.7682	0.8085	0.9207	0.9321	**0.9730**	**0.9764**	0.5914	0.6838	0.7222	0.7746	0.7958	0.8306
K=600	0.7649	0.8065	0.9126	0.9255	**0.9649**	**0.9694**	0.5694	0.6713	0.7011	0.7614	0.7810	0.8206
K=700	0.7594	0.8022	0.9093	0.9228	**0.9606**	**0.9658**	0.5573	0.6649	0.6918	0.7555	0.7678	0.8110
K=800	0.7554	0.7996	0.9018	0.9167	**0.9539**	**0.9602**	0.5465	0.6595	0.6772	0.7460	0.7608	0.8058
K=900	0.7548	0.7956	0.8964	0.9124	**0.9494**	**0.9563**	0.5300	0.6493	0.6701	0.7419	0.7548	0.8019
K=1000	0.7543	0.8002	0.8908	0.9081	**0.9441**	**0.9519**	0.5158	0.6414	0.6612	0.7367	0.7449	0.7956

Tabla A.39: Utilizando un modelo basado en contenido, las K palabras más frecuentes del FCE del vocabulario del propio corpus y DV-MA.

LIAR	Palabras						Caracteres					
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.7631	0.8028	0.9272	0.9373	**0.9795**	**0.9820**	0.6771	0.7317	0.7755	0.8108	0.8504	0.8727
K=200	0.7579	0.7987	0.9152	0.9273	**0.9717**	**0.9752**	0.6671	0.7256	0.7681	0.8051	0.8430	0.8669
K=300	0.7569	0.7985	0.9065	0.9200	**0.9633**	**0.9681**	0.6644	0.7245	0.7650	0.8038	0.8366	0.8619
K=400	0.7507	0.7940	0.9015	0.9159	**0.9584**	**0.9639**	0.6601	0.7214	0.7598	0.7997	0.8289	0.8558
K=500	0.7440	0.7896	0.8929	0.9089	**0.9536**	**0.9598**	0.6610	0.7230	0.7520	0.7943	0.8235	0.8515
K=600	0.7356	0.7840	0.8880	0.9048	**0.9499**	**0.9567**	0.6536	0.7183	0.7511	0.7940	0.8223	0.8507
K=700	0.7338	0.7828	0.8853	0.9027	**0.9448**	**0.9524**	0.6556	0.7199	0.7484	0.7919	0.8213	0.8499
K=800	0.7304	0.7805	0.8791	0.8976	**0.9427**	**0.9506**	0.6547	0.7194	0.7510	0.7942	0.8198	0.8488
K=900	0.7295	0.7801	0.8761	0.8952	**0.9380**	**0.9466**	0.6475	0.7150	0.7492	0.7927	0.8194	0.8488
K=1000	0.7230	0.7760	0.8717	0.8918	**0.9346**	**0.9438**	0.6494	0.7156	0.7417	0.7873	0.8138	0.8444

Tabla A.40: Utilizando un modelo basado en contenido, las K palabras más frecuentes idioma y DV-SA.

LIAR	Palabras						Caracteres					
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.7572	0.7988	0.9248	0.9352	**0.9749**	**0.9780**	0.6722	0.7311	0.7786	0.8143	0.8514	0.8741
K=200	0.7454	0.7921	0.9011	0.9157	**0.9602**	**0.9654**	0.6462	0.7143	0.7590	0.8006	0.8324	0.8590
K=300	0.7308	0.7826	0.8881	0.9052	**0.9454**	**0.9530**	0.6419	0.7125	0.7506	0.7946	0.8247	0.8524
K=400	0.7239	0.7791	0.8736	0.8939	**0.9341**	**0.9435**	0.6392	0.7109	0.7417	0.7885	0.8145	0.8451
K=500	0.7105	0.7702	0.8660	0.8879	**0.9244**	**0.9356**	0.6297	0.7057	0.7345	0.7830	0.8053	0.8378
K=600	0.7095	0.7701	0.8561	0.8805	**0.9155**	**0.9284**	0.6280	0.7039	0.7322	0.7807	0.8033	0.8367
K=700	0.6991	0.7635	0.8453	0.8726	**0.9061**	**0.9209**	0.6230	0.7004	0.7255	0.7765	0.7946	0.8300
K=800	0.6872	0.7561	0.8383	0.8674	**0.8965**	**0.9133**	0.6187	0.6991	0.7159	0.7703	0.7866	0.8247
K=900	0.6751	0.7488	0.8276	0.8593	**0.8893**	**0.9075**	0.6083	0.692	0.7064	0.7639	0.7791	0.8193
K=1000	0.6648	0.7430	0.8157	0.8506	**0.8795**	**0.9000**	0.6047	0.6907	0.7038	0.7631	0.7727	0.8153

Tabla A.41: Utilizando un modelo basado en contenido, las K palabras más frecuentes del corpus y DV-SA.

LIAR	Palabras						Caracteres					
DV-SA_Contenido	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score	F1-fake	F1-score
Baseline	**0.5890**	**0.4093**	0.1024	0.4033	0.0062	0.3620	0.5529	0.5109	0.5192	0.5168	0.4910	0.5245
K=100	0.7640	0.8028	0.9308	0.9401	**0.9804**	**0.9828**	0.6862	0.7384	0.7810	0.8145	0.8567	0.8780
K=200	0.7553	0.7977	0.9066	0.9200	**0.9636**	**0.9683**	0.6730	0.7306	0.7687	0.8058	0.8401	0.8642
K=300	0.7480	0.7936	0.9000	0.9146	**0.9553**	**0.9612**	0.6668	0.7264	0.7619	0.8012	0.8341	0.8593
K=400	0.7415	0.7907	0.8883	0.9050	**0.9457**	**0.9531**	0.6587	0.7224	0.7586	0.7986	0.8247	0.8521
K=500	0.7364	0.7876	0.8776	0.8968	**0.9856**	**0.9447**	0.6565	0.7219	0.7544	0.7960	0.8184	0.8473
K=600	0.7256	0.7806	0.8667	0.8882	**0.9260**	**0.9368**	0.6427	0.7131	0.7457	0.7907	0.8135	0.8442
K=700	0.7158	0.7741	0.8553	0.8794	**0.9155**	**0.9283**	0.6364	0.7091	0.7376	0.7852	0.8062	0.8486
K=800	0.7085	0.7689	0.8484	0.8741	**0.9055**	**0.9202**	0.6301	0.7106	0.7318	0.7810	0.8016	0.8350
K=900	0.6999	0.7639	0.8381	0.8664	**0.8995**	**0.9153**	0.6361	0.7102	0.7274	0.7782	0.7965	0.8313
K=1000	0.6912	0.7588	0.8321	0.8621	**0.8923**	**0.9097**	0.6275	0.7048	0.7275	0.7788	0.7895	0.8262

Tabla A.42: Utilizando un modelo basado en contenido, las K palabras más frecuentes del FCE del vocabulario del corpus y DV-SA.

A.3.2 Etapa de Evaluación y comparación con el Estado del Arte

En esta sección mostramos nuestra mejor configuración comparada con el estado del arte para el conjunto de datos LIAR.

Model-LIAR	Feature	Accuracy	Precision	Recall	F1-score
SVM	Lexical	0.56	0.56	0.56	0.48
SVM	Lexical+Sentimiento	0.56	0.57	0.56	0.48
LR	Lexical+Sentimiento	0.56	0.56	0.56	0.51
Decision Tree	Lexical+Sentimiento	0.51	0.51	0.51	0.51
Adaboost	Lexical+Sentimiento	0.56	0.56	0.56	0.54
Naive Bayes	**Bigram (TF-IDF)**	**0.60**	**0.59**	**0.60**	**0.59**
k-NN	Empath Features	0.53	0.53	0.53	0.53
Nuestro Modelo: SVM, DV-MA, Estilo, K=1000, FCE	**Trigramas de Palabras**	**0.59**	**0.59**	**0.57**	**0.56**

Tabla A.43: Comparación LIAR con el estado del arte

A.4 Conjunto de Datos: CoAID (Inglés)

Según el artículo de referencia realizan un 5-cross-validation con una partición 75-25 para entrenamiento y prueba respectivamente y reportan la medida promedio.

Para nuestro enfoque, Realizamos la siguiente distribución de los datos: 75-25, para entrenamiento, y prueba respectivamente, y del 75 por ciento del entrenamiento utilizamos el 80-20 para entrenamiento y validación respectivamente, realizando el proceso de división de forma aleatoria cuatro veces, obteniendo así cuatro colecciones independientes: De igual forma realizamos las cuatro corridas y reportamos la medida promedio, solo que validamos primero nuestro modelo y la mejor representación obtenida es la que evaluamos en el conjunto de prueba para reportar y compararnos con el estado del arte.

A.4.1 Etapa de Validación

En este subepígrafe presentamos todo la experimentación realizada en el conjunto de datos CoAID sobre la partición de validación.

CoAID K=0	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
DV-MA	0.0000	0.4859	0.0000	0.4859	**0.1105**	**0.5421**	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859
DV-SA	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859

Tabla A.44: Utilizando el enmascaramiento total del texto

CoAID DV-MA_Estilo palabras_frecuentes_idioma	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	**0.1927**	**0.5838**	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.0000	0.4859	0.0000	0.4859	**0.1749**	**0.5748**	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859
K=200	0.0385	0.5054	0.0385	0.5054	**0.1927**	**0.5838**	0.0000	0.4859	0.0000	0.4859	0.0172	0.4947
K=300	0.0192	0.4957	0.192	0.4957	**0.1927**	**0.5838**	0.0000	0.4859	0.0172	0.4947	0.0172	0.4947
K=400	0.0365	0.5044	0.0365	0.5044	**0.1927**	**0.5838**	0.0172	0.4947	0.0172	0.4947	0.0172	0.4947
K=500	0.0735	0.5232	0.0.0735	0.5232	**0.1927**	**0.5838**	0.0172	0.4947	0.0172	0.4947	0.0172	0.4947
K=600	0.1213	0.5475	0.1213	0.5475	**0.1927**	**0.5838**	0.0172	0.4947	0.0172	0.4947	0.0172	0.4947
K=700	0.1022	0.5378	0.1022	0.5378	**0.1927**	**0.5838**	0.0172	0.4947	0.0172	0.4947	0.0365	0.5044
K=800	0.1348	0.5544	0.1348	0.5544	**0.1927**	**0.5838**	0.0365	0.5044	0.0172	0.4947	0.0526	0.5126
K=900	0.1589	0.5666	0.1589	0.5666	**0.1927**	**0.5838**	0.0365	0.5044	0.0516	0.5121	0.0526	0.5126
K=1000	**0.2008**	**0.5880**	**0.2008**	**0.5880**	0.1927	0.5838	0.0365	0.5044	0.0324	0.5024	0.0365	0.5044

Tabla A.45: Utilizando un modelo basado en Estilo las k palabras más frecuentes del idioma y DV-MA

CoAID DV-MA_Estilo palabras_frecuentes_corpus	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.0344	0.5034	0.0344	0.5034	**0.1749**	**0.5848**	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859
K=200	0.1233	0.5486	0.1233	0.5486	**0.1927**	**0.5838**	0.0000	0.4859	0.0000	0.4859	0.0172	0.4859
K=300	0.1752	0.5750	0.1752	0.5750	**0.1927**	**0.5838**	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859
K=400	**0.2678**	**0.6220**	**0.2678**	**0.6220**	0.1927	0.5838	0.0000	0.4859	0.0000	0.4859	0.0000	0.4859
K=500	**0.2768**	**0.6265**	**0.2768**	**0.6265**	0.1927	0.5838	0.0000	0.4859	0.0000	0.4859	0.0152	0.4936
K=600	**0.3042**	**0.6404**	**0.3042**	**0.6404**	0.1927	0.5838	0.0000	0.4859	0.0000	0.4859	0.0485	0.5106
K=700	**0.3481**	**0.6627**	**0.3481**	**0.6627**	0.1927	0.5838	0.0000	0.4859	0.0152	0.4936	0.0485	0.5106
K=800	**0.3498**	**0.6634**	**0.3498**	**0.6634**	0.1927	0.5838	0.0000	0.4859	0.0324	0.5024	0.0677	0.5203
K=900	**0.3623**	**0.6698**	**0.3623**	**0.6698**	0.1927	0.5838	0.0000	0.4859	0.0324	0.5024	0.0970	0.5352
K=1000	**0.3806**	**0.6794**	**0.3806**	**0.6794**	0.1927	0.5838	0.0000	0.4859	0.0324	0.5024	0.0970	0.5352

Tabla A.46: Utilizando un modelo basado en Estilo las k palabras más frecuentes del corpus y DV-MA

CoAID DV-MA_Estilo	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.0000	0.4859	0.0000	0.4859	**0.1749**	**0.5747**	0.0000	0.4859	0.0000	0.4859	0.0172	0.4947
K=200	0.0000	0.4859	0.0000	0.4859	**0.1749**	**0.5747**	0.0000	0.4859	0.0000	0.4859	0.0172	0.4947
K=300	0.0486	0.5106	0.0486	0.5106	**0.1927**	**0.5838**	0.0000	0.4859	0.0000	0.4859	0.0365	0.5044
K=400	0.1081	0.5409	0.1081	0.5409	**0.1927**	**0.5838**	0.0000	0.4859	0.0000	0.4859	0.0516	0.5121
K=500	0.1722	0.5734	0.1722	0.5734	**0.1927**	**0.5838**	0.0000	0.4859	0.0000	0.4859	0.0344	0.5034
K=600	0.1685	0.5716	0.1685	0.5716	**0.1927**	**0.5838**	0.0000	0.4859	0.0000	0.4859	0.0869	0.5300
K=700	0.1846	0.5796	0.1846	0.5796	**0.1927**	**0.5838**	0.0000	0.4859	0.0152	0.4936	0.1161	0.5449
K=800	**0.2122**	**0.5937**	**0.2122**	**0.5937**	0.0000	0.4859	0.0000	0.4859	0.0152	0.4936	0.1339	0.5539
K=900	**0.2011**	**0.5880**	**0.2011**	**0.5880**	0.1927	0.5838	0.0000	0.4859	0.0485	0.5106	0.1339	5539
K=1000	0.1829	0.5786	0.1829	0.5786	**0.1927**	**0.5838**	0.0000	0.4859	0.0827	0.5280	0.1482	0.5612

Tabla A.47: Utilizando un modelo basado en Estilo las k palabras más frecuentes del FCE del vocabulario del propio corpus y DV-MA

CoAID DV-SA_Estilo	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.0000	0.4859	0.0000	0.4859	**0.2032**	**0.5890**	0.0192	0.4957	0.0536	0.5131	0.1158	0.5446
K=200	0.0000	0.4859	0.0000	0.4859	**0.2339**	**0.4048**	0.0192	0.4957	0.0670	0.5197	0.1441	0.5590
K=300	0.0000	0.4859	0.0000	0.4859	**0.2092**	**0.5922**	0.0192	0.4957	0.1092	0.5411	0.1722	0.5734
K=400	0.0192	0.4957	0.0192	0.4957	**0.2219**	**0.5986**	0.0365	0.5044	0.1313	0.5524	0.1894	0.5820
K=500	0.0192	0.4957	0.0192	0.4957	**0.1927**	**0.5838**	0.0676	0.5201	0.1728	0.5736	0.1770	0.5757
K=600	0.0192	0.4957	0.0192	0.4957	0.1927	0.5838	0.0823	0.5276	**0.2143**	**0.5947**	**0.2143**	**0.5947**
K=700	0.0365	0.5044	0.0365	0.5044	0.1927	0.5838	0.1048	0.5391	**0.2009**	**0.5877**	0.1902	0.5824
K=800	0.0365	0.5044	0.0365	0.5044	0.1927	0.5838	0.1043	0.5387	**0.2158**	**0.5953**	0.2027	0.5888
K=900	0.0516	0.5121	0.0516	0.5121	0.1927	0.5838	0.1208	0.5471	0.2009	0.5877	**0.2449**	**0.6102**
K=1000	0.0516	0.5121	0.0516	0.5121	0.1794	0.5770	0.1047	0.5389	**0.2422**	**0.6088**	**0.2422**	**0.6088**

Tabla A.48: Utilizando un modelo basado en Estilo las k palabras más frecuentes del idioma y DV-SA

CoAID DV-SA_Estilo	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.0000	0.4859	0.0000	0.4859	0.1727	0.5735	0.0192	0.4957	0.1029	0.5382	**0.1996**	**0.5874**
K=200	0.0000	0.4859	0.0000	0.4859	0.2069	0.5909	0.0813	0.5272	0.2561	0.6161	**0.3883**	**0.6632**
K=300	0.0192	0.4957	0.0192	0.4957	0.2182	0.5965	0.0117	0.5455	0.3449	0.6613	**0.3755**	**0.6769**
K=400	0.0192	0.4957	0.0192	0.4957	0.2189	0.0.5962	0.1853	0.5801	**0.4159**	**0.6974**	0.3958	0.6872
K=500	0.0192	0.4957	0.0192	0.4957	0.2344	0.6047	0.2495	0.6123	**0.4296**	**0.7041**	0.3918	0.6849
K=600	0.0677	0.5203	0.0677	0.5203	0.2189	0.5969	0.2428	0.6088	**0.4118**	**0.6946**	0.4150	0.6968
K=700	0.1147	0.5441	0.1147	0.5441	0.2355	0.6053	0.3303	0.6532	**0.4797**	**0.7294**	0.4150	0.6968
K=800	0.1274	0.5506	0.1274	0.5506	0.2204	0.5978	0.2867	0.6309	**0.4673**	**0.7232**	0.3913	0.6846
K=900	0.1565	0.5654	0.1565	0.5654	0.2369	0.6061	0.3393	0.6577	**0.4432**	**0.7107**	0.3861	0.6820
K=1000	0.1743	0.5745	0.1743	0.5745	0.2176	0.5961	0.3532	0.6649	**0.4577**	**0.7182**	0.4025	0.6903

Tabla A.49: Utilizando un modelo basado en Estilo las k palabras más frecuentes del corpus y DV-SA

CoAID DV-SA_Estilo	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.0000	0.4859	0.0000	0.4859	**0.1419**	**0.5579**	0.0192	0.4957	0.0887	0.5307	0.1325	0.5532
K=200	0.0000	0.4859	0.0000	0.4859	0.1749	0.5748	0.0192	0.4957	0.1167	0.5450	**0.1876**	**0.5812**
K=3000	0.0192	0.4957	0.0192	0.4957	0.1927	0.5838	0.0536	0.5131	0.2109	0.5932	**0.2607**	**0.6184**
K=400	0.0192	0.4957	0.0192	0.4957	0.1927	0.5838	0.0940	0.5336	0.2309	0.6034	**0.2685**	**0.6224**
K=500	0.0192	0.4957	0.0192	0.4957	0.1927	0.5838	0.0940	0.5336	0.2432	0.6095	**0.2748**	**0.6257**
K=600	0.0516	0.5121	0.0516	0.5121	0.1927	0.5838	0.1707	0.5727	**0.2893**	**0.6330**	0.2639	0.6202
K=700	0.0851	0.5291	0.0851	0.5291	0.1927	0.5838	0.1413	0.5573	**0.2502**	**0.6126**	0.2399	0.6079
K=800	0.0851	0.5291	0.0851	0.5291	0.2092	0.5922	0.1452	0.5596	**0.2539**	**0.6149**	0.2518	0.6139
K=900	0.1163	0.5449	0.1163	0.5449	0.2092	0.5922	0.1314	0.5525	0.2564	0.6161	**0.2861**	**0.6314**
K=1000	0.1163	0.5449	0.1163	0.5449	0.2092	0.5922	0.1808	0.5776	0.2745	0.6252	**0.2861**	**0.6314**

Tabla A.50: Utilizando un modelo basado en Estilo las k palabras más frecuentes del FCE del vocabulario del corpus y DV-SA

CoAID DV-MA_Contenido	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.2204	0.5977	0.2204	0.5977	0.1927	0.5838	0.1048	0.5391	**0.2639**	**0.6198**	0.2549	0.6153
K=200	0.2329	0.6041	0.2329	0.6041	0.1927	0.5838	0.0851	0.5291	0.2401	0.6077	**0.2409**	**0.6082**
K=3000	0.2204	0.5978	0.2204	0.5978	0.1927	0.5838	0.0709	0.5219	0.1870	0.5808	**0.2389**	**0.6072**
K=400	0.1780	0.5763	0.1780	0.5763	0.1927	0.5838	0.0709	0.5219	0.1538	0.5638	**0.2493**	**0.6125**
K=500	**0.1780**	**0.5763**	**0.1780**	**0.5763**	0.1927	0.5838	0.0536	0.5131	0.1147	0.5441	0.1714	0.5727
K=600	**0.1780**	**0.5763**	**0.1780**	**0.5763**	0.1927	0.5838	0.0536	0.5131	0.1012	0.5373	0.1688	0.5714
K=700	**0.1780**	**0.5763**	**0.1780**	**0.5763**	0.1927	0.5838	0.0192	0.4957	0.1012	0.5373	0.1688	0.5714
K=800	**0.1653**	**0.5648**	**0.1653**	**0.5648**	0.1927	0.5838	0.0192	0.4957	0.0869	0.5300	0.1452	0.5594
K=900	**0.1653**	**0.5648**	**0.1653**	**0.5648**	0.1927	0.5838	0.0192	0.4957	0.0869	0.5300	0.1475	0.5608
K=1000	**0.1653**	**0.5648**	**0.1653**	**0.5648**	0.1927	0.5838	0.0192	0.4957	0.0869	0.5300	0.1475	0.5608

Tabla A.51: Utilizando un modelo basado en Contenido las k palabras más frecuentes del idioma y DV-SA

CoAID DV-MA_Contenido	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_corpus	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	0.3440	0.6608	0.3249	0.6510	0.2483	0.6120
K=100	0.1959	0.5854	0.1959	0.5854	0.1927	0.5838	0.0728	0.5238	**0.2348**	**0.6049**	0.2048	0.5898
K=200	0.1825	0.5786	0.1825	0.5786	**0.1927**	**0.5838**	0.0563	0.5145	0.1358	0.5546	0.1912	0.5828
K=3000	**0.2087**	**0.5919**	**0.2087**	**0.5919**	0.1927	0.5838	0.0385	0.5054	0.1513	0.5626	0.1917	0.5831
K=400	**0.1960**	**0.5854**	**0.1960**	**0.5854**	0.1927	0.5838	0.0385	0.5054	0.1061	0.5398	0.1791	0.5769
K=500	0.1819	0.5782	0.1819	0.5782	**0.1927**	**0.5838**	0.0385	0.5054	0.0557	0.5142	0.1213	0.5475
K=600	0.1819	0.5782	0.1819	0.5782	**0.1927**	**0.5838**	0.0385	0.5054	0.0557	0.5142	0.1061	0.5398
K=700	0.1668	0.5706	0.1668	0.5706	**0.1927**	**0.5838**	0.0385	0.5054	0.0557	0.5142	0.1212	0.5474
K=800	0.1819	0.5782	0.1819	0.5782	**0.1927**	**0.5838**	0.0370	0.5047	0.0896	0.5314	0.0896	0.5314
K=900	0.1668	0.5706	0.1668	0.5706	**0.1927**	**0.5838**	0.0563	0.5145	0.0735	0.5232	0.0735	0.5232
K=1000	0.1312	0.5475	0.1312	0.5475	**0.1927**	**0.5838**	0.0563	0.5145	0.0735	0.5232	0.0735	0.5232

Tabla A.52: Utilizando un modelo basado en Contenido las k palabras más frecuentes del corpus y DV-MA

CoAID DV-MA_Contenido	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
FCE	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	0.3440	0.6608	0.3249	0.6510	0.2483	0.6120
K=100	0.2204	0.5978	0.2204	0.5978	0.1927	0.5838	0.0563	0.5145	**0.2481**	**0.6116**	0.2166	0.5957
K=200	0.2329	0.6041	0.2329	0.6041	0.1927	0.5838	0.0192	0.4957	**0.2359**	**0.6056**	0.2223	0.5986
K=300	0.1960	0.5854	0.1960	0.5854	0.1927	0.5838	0.0192	0.4957	0.1652	0.5696	**0.2033**	**0.5889**
K=400	**0.1960**	**0.5854**	**0.1960**	**0.5854**	0.1927	0.5838	0.0192	0.4957	0.0900	0.5315	0.1769	0.5757
K=500	**0.1960**	**0.5854**	**0.1960**	**0.5854**	0.1927	0.5838	0.0192	0.4957	0.0735	0.5232	0.1506	0.5624
K=600	**0.1946**	**0.5847**	**0.1946**	**0.5847**	0.1927	0.5838	0.0192	0.4957	0.0563	0.5145	0.1363	0.5551
K=700	**0.1960**	**0.5854**	**0.1960**	**0.5854**	0.1927	0.5838	0.0192	0.4957	0.0385	0.9724	0.1061	0.5398
K=800	**0.1960**	**0.5854**	**0.1960**	**0.5854**	0.1927	0.5838	0.0192	0.4957	0.0385	0.9724	0.0735	0.5232
K=900	0.1819	0.5782	0.1819	0.5782	**0.1927**	**0.5838**	0.0192	0.4957	0.0385	0.5054	0.0735	0.5232
K=1000	0.1503	0.5622	0.1503	0.5622	**0.1927**	**0.5838**	0.0192	0.4957	0.0563	0.5145	0.0385	0.5054

Tabla A.53: Utilizando un modelo basado en Contenido las k palabras más frecuentes del FCE del vocabulario del corpus y DV-MA

CoAID DV-SA_Contenido	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
palabras_frecuentes_idioma	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	0.3440	0.6608	0.3249	0.6510	0.2483	0.6120
K=100	0.1190	0.5463	0.1190	0.5463	0.1927	0.5838	0.2079	0.5915	**0.2741**	**0.6250**	0.2626	0.6193
K=200	0.0869	0.5300	0.0869	0.5300	0.1927	0.5838	0.1629	0.5685	**0.2855**	**0.6309**	0.2712	0.6236
K=300	0.0708	0.5219	0.0708	0.5219	0.1927	0.5838	0.1463	0.5601	**0.2731**	**0.6246**	0.2587	0.6173
K=400	0.0708	0.5219	0.0708	0.5219	0.1794	0.5770	0.1154	0.5444	0.2441	0.6099	**0.2639**	**0.6200**
K=500	0.0708	0.5219	0.0708	0.5219	0.1653	0.5698	0.1163	0.5449	**0.2011**	**0.5880**	0.1957	0.5853
K=600	0.0708	0.5219	0.0708	0.5219	0.1653	0.5698	0.1163	0.5449	0.1870	0.5809	**0.1978**	**0.5864**
K=700	0.0708	0.5219	0.0708	0.5219	0.1653	0.5698	0.1020	0.5377	**0.2300**	**0.6027**	0.1994	0.5871
K=800	0.0708	0.5219	0.0708	0.5219	0.1475	0.5608	0.1020	0.5377	0.1957	0.5853	**0.1994**	**0.5871**
K=900	0.0708	0.5219	0.0708	0.5219	0.1475	0.5608	0.0869	0.5300	**0.1921**	**0.5835**	0.1780	0.5763
K=1000	0.0536	0.5131	0.0536	0.5131	0.1640	0.5692	0.0708	0.5219	**0.1743**	**0.5745**	0.1722	0.5734

Tabla A.54: Utilizando un modelo basado en Contenido las k palabras más frecuentes del idioma y DV-SA

CoAID DV-SA_Contenido palabras_frecuentes_corpus	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	0.3440	0.6608	0.3249	0.6510	0.2483	0.6120
K=100	0.0887	0.5309	0.0887	0.5309	0.1794	0.5770	0.1367	0.5553	**0.2727**	**0.6243**	0.2048	0.5898
K=200	0.0385	0.5054	0.0385	0.5054	0.1794	0.5770	0.028	0.5228	**0.2048**	**0.5898**	**0.2048**	**0.5898**
K=300	0.0385	0.5054	0.0385	0.5054	0.1660	0.5702	0.0900	0.5316	**0.2180**	**0.5966**	0.2056	0.5902
K=400	0.0385	0.5054	0.0385	0.5054	0.1482	0.5612	0.0714	0.5222	**0.1936**	**0.5842**	0.1782	0.5764
K=500	0.0385	0.5054	0.0385	0.5054	0.1339	0.5539	0.0557	0.5142	0.1517	0.5629	**0.1647**	**0.5695**
K=600	0.0385	0.5054	0.0385	0.5054	0.1198	0.5467	0.0385	0.5054	0.1052	0.5393	**0.1506**	**0.5624**
K=700	0.0385	0.5054	0.0385	0.5054	0.1198	0.5467	0.0385	0.5054	0.1363	0.5551	**0.1647**	**0.5695**
K=800	0.0385	0.5054	0.0385	0.5054	0.1048	0.5391	0.0385	0.5054	0.1363	0.5551	**0.1506**	**0.5624**
K=900	0.0385	0.5054	0.0385	0.5054	0.1048	0.5391	0.0385	0.5054	0.1213	0.5475	**0.1363**	**0.5551**
K=1000	0.0385	0.5054	0.0385	0.5054	0.1198	0.5467	0.0385	0.5054	0.1213	0.5475	**0.1505**	**0.5623**

Tabla A.55: Utilizando un modelo basado en Contenido las k palabras más frecuentes del corpus y DV-SA

CoAID DV-SA_Contenido FCE	Palabras						Caracteres					
	Unigramas		Bigramas		Trigramas		Trigramas		Cuatrigramas		Pentagramas	
	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score	F1-Fake	F1-score
Baseline	0.2850	0.6308	0.1927	0.5838	0.1927	0.5838	**0.3440**	**0.6608**	0.3249	0.6510	0.2483	0.6120
K=100	0.1368	0.5553	0.1368	0.5553	0.1927	0.5838	0.2241	0.5997	**0.2818**	**0.6288**	0.2449	0.6102
K=200	0.0708	0.5219	0.0708	0.5219	0.1794	0.5770	0.2114	0.5932	**0.2711**	**0.6233**	0.2358	0.6056
K=300	0.0536	0.5131	0.0536	0.5131	0.1794	0.5770	0.1355	0.547	**0.2405**	**0.6079**	0.2226	0.5988
K=400	0.0385	0.5054	0.0385	0.5054	0.1794	0.5770	0.1054	0.5394	0.2148	0.5948	**0.2165**	**0.5958**
K=500	0.0385	0.5054	0.0385	0.5054	0.1482	0.612	0.0900	0.5315	**0.2165**	**0.5958**	0.2062	0.5907
K=600	0.0385	0.5054	0.0385	0.5054	0.1482	0.612	0.0557	0.5142	**0.2063**	**0.5907**	0.2029	0.5890
K=700	0.0385	0.5054	0.0385	0.5054	0.1339	0.5539	0.0735	0.5232	**0.1782**	**0.5764**	**0.1782**	**0.5764**
K=800	0.0385	0.5054	0.0385	0.5054	0.1339	0.5539	0.0557	0.5142	**0.1782**	**0.5764**	**0.1782**	**0.5764**
K=900	0.0385	0.5054	0.0385	0.5054	0.1198	0.5467	0.0563	0.5144	0.1355	0.5547	**0.1640**	**0.5692**
K=1000	0.0385	0.5054	0.0385	0.5054	0.1198	0.5467	0.0563	0.5144	0.1195	0.5465	**0.767**	**0.5757**

Tabla A.56: Utilizando un modelo basado en Contenido las k palabras más frecuentes del FCE del vocabulario del corpus y DV-SA

A.4.2 Etapa de Evaluación y comparación con el Estado del Arte

En esta sección comparamos nuestro mejor resultado con el estado del arte para este conjunto de datos.

Model-CoAID	Precision	Recall	F1-score
SVM	0.4036	0.1322	0.1986
LR	0.4287	0.0690	0.1143
RF	0.6056	0.0581	0.045
CNN	0.9653	0.1238	0.1983
BiGRU	0.7476	0.0524	0.0930
CSI	0.6814	0.2109	0.2283
SAME\v	0.8922	0.2991	0.3400
HAN	0.6965	0.4659	0.5471
dEFEND	**0.8965**	**0.4847**	**0.5814**
Nuestro Modelo: SVM, DV-SA, Estilo, K=700, palabras_frecuentes del corpus, cutarigramas de caracteres	**0.8903**	**0.6315**	**0.6882**

Tabla A.57: Comparación CoAID con el estado del arte

Apéndice B

Resultados de los experimentos realizados. Utilizando CNN

A continuación, se muestran los resultados de todos los experimentos realizados sobre cada uno de los conjuntos de datos utilizando esta vez, las redes neuronales convolucionales (CNN). De igual forma que la fase de experimentación anterior, donde utilizábamos las máquinas de soporte vectorial como clasificador, en la experimentación con las redes neuronales (CNN) evaluamos para ambos modelos: estilo y contenido y cada uno de ellos para los tres enfoques de selección de las k palabras más frecuentes, con una variación del parámetro k desde 0 hasta 1000, pero esta vez pasamos a la red solamente texto enmascarado con simples asteriscos (algoritmo $DV - SA$).

Se realizaron los experimentos sobre los mismos cuatro conjuntos de datos y se utilizó la misma distribución de los datos para cada uno de ellos teniendo en cuenta para su extracción: las palabras más frecuentes del idioma, las palabras más frecuentes extraídas del vocabulario del propio corpus y las palabras más frecuentes teniendo en cuenta el FCE calculado del propio vocabulario del corpus.

B.1 Conjunto de Datos: MEX-A3T (Español)

Para el conjunto de datos (MEX-A3T); presentamos a continuación todos los experimentos realizados entrenando y probando sobre nuestro conjunto de validación y a partir del mejor resultado obtenido en esta fase de experimentación; (mostramos la tabla), entonces evaluamos nuestro modelo para el conjunto de prueba y comparamos nuestro resultado con el estado del arte para este conjunto de datos.

B.1.1 Etapa de Validación

La validación de los experimentos se llevó a cabo sobre una división del 80-20, para las particiones de entrenamiento y validación respectivamente.

MEX-A3T	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
DV-SA	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.70	0.64	0.68	0.61	0.66	0.73	0.72	0.76
K=0	0.64	0.59	**0.70**	**0.73**	0.68	0.72	0.62	0.70

Tabla B.1: Utilizando el enmascaramiento total del texto (K=0).

	MEX-A3T							
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_idioma	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.70	0.64	0.68	0.61	0.66	0.73	**0.72**	**0.76**
K=100	0.74	0.74	0.69	0.75	**0.72**	**0.75**	0.60	0.70
K=200	0.73	0.74	**0.76**	**0.78**	0.74	0.73	0.71	0.70
K=300	0.73	0.77	0.74	0.76	0.74	0.74	**0.76**	**0.78**
K=400	0.76	0.75	**0.73**	**0.76**	0.68	0.62	0.56	0.41
K=500	**0.76**	**0.77**	0.76	0.76	0.74	0.72	0.66	0.57
K=600	0.76	0.75	0.75	0.76	**0.76**	**0.77**	0.73	0.72
K=700	0.76	0.75	0.76	0.76	**0.78**	**0.77**	0.76	0.76
K=800	0.74	0.74	0.74	0.74	**0.72**	**0.77**	0.71	0.70
K=900	**0.76**	**0.76**	0.74	0.74	0.70	0.66	0.67	0.72
K=1000	**0.81**	**0.80**	0.76	0.75	0.67	0.73	0.67	0.72

Tabla B.2: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes del idioma y DV-SA.

	MEX-A3T							
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_corpus	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.70	0.64	0.68	0.61	0.66	0.73	**0.72**	**0.76**
K=100	0.78	0.75	0.75	0.76	0.78	0.77	**0.78**	**0.79**
K=200	0.78	0.76	**0.79**	**0.77**	0.78	0.75	0.75	0.76
K=300	0.77	0.78	0.76	0.78	0.78	0.75	**0.79**	**0.79**
K=400	0.76	0.78	**0.83**	**0.81**	0.74	0.77	0.81	0.80
K=500	0.79	0.80	0.81	0.78	0.79	0.79	**0.82**	**0.82**
K=600	0.76	0.74	**0.82**	**0.81**	0.80	0.81	0.82	0.80
K=700	0.77	0.77	**0.82**	**0.81**	**0.82**	**0.81**	**0.82**	**0.81**
K=800	0.79	0.79	**0.83**	**0.82**	0.78	0.75	0.79	0.79
K=900	0.82	0.82	0.82	0.81	**0.83**	**0.82**	0.75	0.78
K=1000	0.79	0.78	0.82	0.80	**0.85**	**0.85**	0.81	0.81

Tabla B.3: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes del propio corpus y DV-SA.

MEX-A3T								
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
FCE	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.70	0.64	0.68	0.61	0.66	0.73	**0.72**	**0.76**
K=100	**0.79**	**0.79**	0.73	0.75	0.76	0.74	0.74	0.76
K=200	**0.74**	**0.74**	0.71	0.67	0.73	0.71	0.71	0.73
K=300	0.75	0.75	0.73	0.70	0.75	0.76	**0.76**	**0.77**
K=400	0.77	0.74	0.76	0.74	0.72	0.76	**0.76**	**0.75**
K=500	0.72	0.75	**0.75**	**0.76**	0.73	0.70	0.70	0.75
K=600	0.73	0.72	**0.72**	**0.75**	0.70	0.75	0.75	0.74
K=700	0.72	0.76	**0.76**	**0.78**	0.79	0.75	0.75	0.77
K=800	0.76	0.74	**0.78**	**0.78**	0.76	0.72	0.72	0.73
K=900	0.64	0.72	**0.76**	**0.77**	0.68	0.73	0.73	0.76
K=1000	0.76	0.74	0.74	0.71	0.77	0.73	**0.73**	**0.76**

Tabla B.4: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del cálculo del FCE del propio corpus y DV-SA.

MEX-A3T								
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_idioma	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.70	0.64	0.68	0.61	0.66	0.73	**0.72**	**0.76**
K=100	0.71	0.70	0.74	0.73	**0.73**	**0.75**	0.73	0.73
K=200	0.72	0.72	**0.70**	**0.76**	0.73	0.73	0.73	0.72
K=300	0.72	0.72	0.70	0.67	0.71	0.74	**0.76**	**0.76**
K=400	0.71	0.76	**0.76**	**0.78**	0.69	0.75	0.76	0.75
K=500	**0.74**	**0.78**	**0.74**	**0.78**	0.73	0.77	0.74	0.73
K=600	**0.76**	**0.79**	0.70	0.67	0.75	0.75	0.75	0.75
K=700	0.65	0.74	0.73	0.72	**0.75**	**0.75**	0.73	0.75
K=800	0.73	0.73	0.68	0.74	**0.79**	**0.79**	0.74	0.74
K=900	0.62	0.72	0.74	0.74	**0.76**	**0.78**	0.73	0.74
K=1000	**0.78**	**0.80**	0.77	0.80	0.72	0.77	0.73	0.72

Tabla B.5: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del idioma y DV-SA.

MEX-A3T								
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_corpus	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.70	0.64	0.68	0.61	0.66	0.73	**0.72**	**0.76**
K=100	0.74	0.74	**0.76**	**0.78**	0.74	0.76	0.73	0.75
K=200	0.73	0.75	0.74	0.72	**0.70**	**0.76**	0.70	0.69
K=300	0.71	0.77	0.73	0.75	**0.78**	**0.79**	0.73	0.75
K=400	0.68	0.75	0.69	0.66	0.69	0.66	**0.71**	**0.75**
K=500	0.72	0.69	0.67	0.63	**0.76**	**0.78**	0.63	0.55
K=600	0.71	0.70	0.74	0.76	0.73	0.75	**0.74**	**0.78**
K=700	0.65	0.60	0.69	0.67	**0.72**	**0.74**	0.74	0.73
K=800	0.67	0.65	0.67	0.63	**0.72**	**0.71**	0.69	0.66
K=900	0.68	0.71	0.74	0.75	**0.74**	**0.78**	0.71	0.69
K=1000	0.68	0.75	0.68	0.67	**0.77**	**0.79**	0.73	0.75

Tabla B.6: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes extraídas del vocabulario del propio corpus y DV-SA.

MEX-A3T								
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
FCE	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.70	0.64	0.68	0.61	0.66	0.73	**0.72**	**0.76**
K=100	0.73	0.75	**0.78**	**0.79**	0.75	0.75	0.76	0.78
K=200	0.72	0.75	**0.77**	**0.79**	0.70	0.69	0.70	0.75
K=300	0.74	0.73	0.68	0.65	0.70	0.70	**0.68**	**0.74**
K=400	0.70	0.74	**0.73**	**0.75**	0.67	0.71	0.71	0.69
K=500	0.73	0.74	0.73	0.74	**0.70**	**0.75**	0.72	0.70
K=600	**0.72**	**0.75**	0.71	0.72	0.71	0.73	0.68	0.68
K=700	0.72	0.73	0.71	0.71	0.67	0.70	**0.75**	**0.75**
K=800	0.69	0.74	**0.74**	**0.75**	0.73	0.73	0.68	0.75
K=900	**0.74**	**0.76**	0.75	0.75	0.71	0.72	0.70	0.67
K=1000	**0.77**	**0.79**	0.73	0.71	0.70	0.68	0.66	0.60

Tabla B.7: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes extraídas del cálculo del FCE del vocabulario del propio corpus y DV-SA.

B.1.2 Etapa de Evaluación y comparación con el Estado del Arte

El mejor enfoque obtenido a partir de la etapa de validación, es el que se evalúa en el conjunto de prueba, este último resultado es el que se reporta para este conjunto de datos utilizando las redes neuronales convolucionales y el cual comparamos con el estado del arte.

Model_MEX-A3T	**Fake**	**True**	**F1-macro**	**Accuracy**
Idiap-UAM-1	0.8444	0.8688	0.8566	0.8576
Idiap-UAM-2	0.8406	0.8599	0.8502	0.8508
Nuestro modelo: CNN, DV-SA, Modelo basado en Estilo, K=1000, palabras más frecuentes del corpus Kernel=4	**0.8251**	**0.8153**	**0.8202**	**0.8203**
Ares	0.8188	0.8151	0.8169	0.8169
CIMAT-1	0.7943	0.8117	0.8030	0.8034
Baseline (BoW-RF)	0.7850	0.7879	0.7864	0.7864
Intensos-2	0.7703	0.7883	0.7793	0.7797
Intensos-1	0.7597	0.7376	0.7487	0.7492
Baseline (INGEOTEC)	0.7596	0.7723	0.7659	0.7661
ITCG-SD	0.7464	0.7771	0.7617	0.7627

Tabla B.8: Comparación MEX-A3T con el estado del arte

B.2 Conjunto de Datos: RAW-CovidES (Español)

Según el artículo de referencia realizan un 5-cross-validation con una partición 80-20 para entrenamiento y prueba respectivamente y reportan la medida promedio. Para nuestro enfoque, Realizamos la siguiente distribución de los datos: 60-20-20, para entrenamiento, validación y prueba respectivamente. Validamos primero nuestro modelo y la mejor representación obtenida es la que evaluamos en el conjunto de prueba para reportar y compararnos con el estado del arte.

B.2.1 Etapa de validación

En esta fase de validación, la mejor configuración obtenida, la podemos ver en la Tabla: B.15.

RAW-CovidES								
DV-SA	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
	F1-score	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.58	0.65	0.59	0.58	0.55	0.46	**0.36**	**0.72**
K=0	**0.56**	**0.73**	0.58	0.71	0.61	0.68	0.59	0.61

Tabla B.9: Utilizando el enmascaramiento total del Texto (K=0).

RAW-CovidES								
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_idioma	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.58	0.65	0.59	0.58	0.55	0.46	**0.36**	**0.72**
K=100	0.65	0.62	0.61	0.74	0.30	0.00	**0.56**	**0.77**
K=200	0.49	0.35	**0.56**	**0.77**	0.56	0.56	0.30	0.00
K=300	**0.50**	**0.75**	0.46	0.70	0.62	0.67	0.53	0.42
K=400	0.36	0.72	0.53	0.65	**0.56**	**0.73**	0.36	0.72
K=500	0.53	0.42	0.30	0.00	**0.65**	**0.62**	0.35	0.10
K=600	**0.48**	**0.72**	0.62	0.57	0.36	0.72	0.52	0.36
K=700	0.30	0.00	0.42	0.71	**0.36**	**0.72**	0.30	0.00
K=800	0.49	0.35	**0.59**	**0.77**	0.55	0.46	0.56	0.77
K=900	0.42	0.71	0.36	0.11	0.72	0.73	**0.75**	**0.76**
K=1000	0.56	0.43	0.68	0.74	**0.65**	**0.76**	0.61	0.74

Tabla B.10: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes del idioma Inglés y DV-SA.

RAW-CovidES								
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_corpus	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.58	0.65	0.59	0.58	0.55	0.46	**0.36**	**0.72**
K=100	**0.59**	**0.76**	0.62	0.57	0.65	0.69	0.36	0.72
K=200	0.58	0.52	**0.74**	**0.79**	0.59	0.76	0.65	0.70
K=300	0.68	0.64	0.72	0.68	0.43	0.73	**0.72**	**0.81**
K=400	0.36	0.11	0.56	0.73	0.49	0.35	**0.75**	**0.78**
K=500	**0.40**	**0.68**	0.66	0.65	0.51	0.40	0.68	0.64
K=600	0.43	0.73	0.63	0.73	**0.74**	**0.79**	0.35	0.69
K=700	0.30	0.00	0.58	0.71	**0.81**	**0.83**	0.62	0.60
K=800	0.72	0.72	**0.72**	**0.74**	0.43	0.73	0.62	0.60
K=900	0.56	0.73	0.52	0.36	0.64	0.72	**0.75**	0.78
K=1000	**0.78**	**0.81**	0.55	0.63	0.36	0.11	0.30	0.00

Tabla B.11: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del vocabulario del propio corpus y DV-SA.

RAW-CovidES								
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
FCE	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.58	0.65	0.59	0.58	0.55	0.46	**0.36**	**0.72**
K=100	0.52	0.44	**0.65**	**0.62**	0.30	0.00	0.30	0.00
K=200	0.44	0.32	0.55	0.63	**0.56**	**0.70**	0.30	0.00
K=300	0.49	0.43	0.59	0.63	0.42	0.71	**0.45**	**0.72**
K=400	0.51	0.62	0.47	0.48	**0.63**	**0.63**	0.56	0.53
K=500	0.51	0.71	0.66	0.67	0.53	0.42	**0.36**	**0.72**
K=600	0.44	0.67	**0.71**	**0.76**	0.72	0.73	0.36	0.72
K=700	**0.61**	**0.74**	0.58	0.52	0.36	0.72	0.49	0.35
K=800	0.56	0.56	0.64	0.56	**0.67**	**0.75**	0.47	0.29
K=900	0.52	0.59	**0.71**	**0.76**	0.36	0.72	0.57	0.67
K=1000	0.60	0.50	0.65	0.60	0.42	0.20	**0.65**	**0.69**

Tabla B.12: Utilizando un Modelo basado en Estilo, las k palabras más frecuentes extraídas del cálculo del FCE del vocabulario del propio corpus y DV-SA.

RAW-CovidES								
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_idioma	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.58	0.65	0.59	0.58	0.55	0.46	**0.36**	**0.72**
K=100	0.56	0.43	**0.36**	**0.72**	0.57	0.67	0.29	0.00
K=200	0.36	0.11	0.43	0.73	0.35	0.69	**0.56**	**0.77**
K=300	0.36	0.11	0.52	0.36	0.56	0.50	**0.36**	**0.72**
K=400	0.64	0.56	**0.52**	**0.67**	0.50	0.50	0.56	0.59
K=500	0.65	0.59	0.52	0.36	0.58	0.65	**0.72**	**0.74**
K=600	0.36	0.72	0.52	0.59	**0.43**	**0.73**	0.56	0.61
K=700	0.64	0.72	**0.56**	**0.77**	0.61	0.54	0.47	0.33
K=800	0.56	0.43	0.36	0.72	**0.61**	**0.74**	0.29	0.00
K=900	0.47	0.29	0.57	0.48	0.65	0.62	**0.36**	**0.72**
K=1000	0.65	0.62	0.66	0.67	0.53	0.42	**0.63**	**0.73**

Tabla B.13: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del idioma y DV-SA.

RAW-CovidES								
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_corpus	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.58	0.65	0.59	0.58	0.55	0.46	**0.36**	**0.72**
K=100	0.56	0.43	**0.36**	**0.72**	0.57	0.67	0.29	0.00
K=200	0.36	0.11	0.43	0.73	0.35	0.69	**0.56**	**0.77**
K=300	0.36	0.11	0.52	0.36	0.56	0.50	**0.36**	**0.72**
K=400	0.64	0.56	**0.52**	**0.67**	0.50	0.50	0.56	0.59
K=500	0.65	0.59	0.52	0.36	0.58	0.65	**0.72**	**0.74**
K=600	0.36	0.72	0.52	0.59	**0.43**	**0.73**	0.56	0.61
K=700	0.64	0.72	**0.56**	**0.77**	0.61	0.54	0.47	0.33
K=800	0.56	0.43	0.36	0.72	**0.61**	**0.74**	0.29	0.00
K=900	0.47	0.29	0.57	0.48	0.65	0.62	**0.36**	**0.72**
K=1000	0.65	0.62	0.66	0.67	0.53	0.42	**0.63**	**0.73**

Tabla B.14: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del corpus y DV-SA.

RAW-CovidES								
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
FCE	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.58	0.65	0.59	0.58	0.55	0.46	**0.36**	**0.72**
K=100	**0.36**	**0.72**	0.68	0.64	0.54	0.70	0.69	0.67
K=200	**0.84**	**0.86**	0.75	0.76	0.71	0.67	0.72	0.81
K=300	**0.84**	**0.87**	0.81	0.80	0.72	0.71	0.77	0.82
K=400	0.65	0.59	**0.59**	**0.76**	0.52	0.36	0.72	0.73
K=500	**0.75**	**0.78**	0.65	0.62	0.47	0.29	0.50	0.75
K=600	**0.81**	**0.84**	0.74	0.79	0.68	0.74	0.43	0.73
K=700	**0.72**	**0.81**	0.68	0.79	0.59	0.76	0.66	0.80
K=800	0.47	0.29	**0.63**	**0.73**	0.68	0.62	0.69	0.69
K=900	0.30	0.00	**0.78**	**0.79**	0.42	0.20	0.50	0.75
K=1000	0.60	0.50	**0.70**	**0.77**	0.51	0.40	0.36	0.11

Tabla B.15: Utilizando un Modelo basado en Contenido, las k palabras más frecuentes del cálculo del FCE del propio corpus y DV-SA.

B.2.2 Etapa de Evaluación y comparación con el Estado del Arte.

La mejor configuración obtenida en la etapa de validación, es evaluada entonces en el conjunto de prueba y son los resultados que reportamos en este acápite y comparamos con el estado del arte.

Model RAW-CovidES	**F1-Fake**	**F1-True**	**F1-macro**	**Accuracy**
Baseline (Random)	0.497	0.548	0.522	0.526
Baseline (TF-IDF)	0.494	0.715	0.605	0.637
Full pipeline	**0.690**	**0.790**	**0.740**	**0.750**
Nuestro Modelo: CNN, DV-SA, Modelo Basado en Contenido Kernel = 2 y FCE, K=300	**0.579**	**0.438**	**0.509**	**0.519**

Tabla B.16: Comparación RAW-CovidES con el estado del arte

B.3 Conjunto de Datos: LIAR (Inglés)

Para este conjunto de datos, contamos con tres particiones: entrenamiento, validación y prueba. De igual forma que con los conjuntos de datos anteriores, entrenamos y probamos en el conjunto de validación, para de ahí obtener nuestra mejor configuración y esta última es la que evaluamos finalmente en el de prueba y el resultado obtenido es el que comparamos con el estado del arte.

B.3.1 Etapa de Evaluación y comparación con el Estado del Arte

En esta sección mostramos nuestra mejor configuración comparada con el estado del arte para el conjunto de datos LIAR.

Model-LIAR	Feature	Accuracy	Precision	Recall	F1-score
SVM	Lexical	0.56	0.56	0.56	0.48
SVM	Lexical+Sentimiento	0.56	0.57	0.56	0.48
LR	Lexical+Sentimiento	0.56	0.56	0.56	0.51
Decision Tree	Lexical+Sentimiento	0.51	0.51	0.51	0.51
Adaboost	Lexical+Sentimiento	0.56	0.56	0.56	0.54
Naive Bayes	**Bigram (TF-IDF)**	**0.60**	**0.59**	**0.60**	**0.59**
k-NN	Empath Features	0.53	0.53	0.53	0.53
Nuestro Modelo: SVM, DV-MA, Estilo, K=1000, FCE	**Trigramas de Palabras**	**0.59**	**0.59**	**0.57**	**0.56**

Tabla B.17: Comparación LIAR con el estado del arte

B.4 Conjunto de Datos: CoAID (Inglés)

Según el artículo de referencia realizan un 5-cross-validation con una partición 75-25 para entrenamiento y prueba respectivamente y reportan la medida promedio. Para nuestro enfoque, realizamos la siguiente distribución de los datos: 75-25, para entrenamiento, y prueba respectivamente, y del 75 por ciento del entrenamiento utilizamos el 80-20 para entrenamiento y validación respectivamente. Validamos primero nuestro

modelo y la mejor representación obtenida es la que evaluamos en el conjunto de prueba para reportar y compararnos con el estado del arte.

B.4.1 Etapa de Validación

En este subepígrafe presentamos todo la experimentación realizada en el conjunto de datos CoAID sobre la partición de validación.

	CoAID							
DV-SA	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
	F1-score	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.79	0.60	0.67	0.37	**0.87**	**0.75**	0.57	0.18
K=0	0.48	0.00	0.48	0.00	0.48	0.00	0.48	0.00

Tabla B.18: Utilizando el enmascaramiento total del texto

	CoAID							
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_idioma	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.79	0.60	0.67	0.37	**0.87**	**0.75**	0.57	0.18
K=100	0.71	0.45	0.61	0.25	**0.72**	**0.46**	0.68	0.38
K=200	0.75	0.52	**0.76**	**0.55**	0.69	0.41	0.67	0.37
K=300	0.70	0.42	**0.75**	0.52	0.55	0.12	0.65	0.32
K=400	0.73	0.49	0.73	0.48	**0.77**	**0.57**	0.67	0.36
K=500	0.69	0.40	**0.75**	**0.52**	0.62	0.26	0.64	0.32
K=600	0.69	0.40	0.73	0.48	**0.74**	**0.50**	0.61	0.24
K=700	**0.73**	**0.48**	0.67	0.36	0.60	0.24	0.63	0.28
K=800	0.73	0.49	**0.78**	**0.59**	0.75	0.52	0.61	0.26
K=900	0.76	0.56	0.74	0.50	0.69	0.41	**0.77**	**0.57**
K=1000	0.69	0.40	**0.75**	**0.52**	0.67	0.36	0.70	0.42

Tabla B.19: Utilizando un modelo basado en Estilo las k palabras más frecuentes del idioma y DV-SA

CoAID								
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabras_frecuentes_corpus	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.79	0.60	0.67	0.37	**0.87**	**0.75**	0.57	0.18
K=100	0.72	0.47	**0.79**	**0.60**	0.76	0.55	0.78	0.59
K=200	0.74	0.50	0.79	0.61	0.72	0.47	**0.81**	**0.65**
K=300	0.73	0.48	**0.77**	**0.57**	0.75	0.52	0.74	0.52
K=400	**0.80**	**0.65**	0.73	0.48	0.77	0.56	0.76	0.55
K=500	0.74	0.51	0.72	0.47	0.75	0.53	**0.83**	**0.69**
K=600	0.76	0.55	0.81	0.64	0.80	0.63	**0.81**	**0.65**
K=700	0.79	0.61	0.80	0.62	**0.83**	**0.68**	0.71	0.45
K=800	**0.81**	**0.65**	0.71	0.46	0.80	0.62	0.67	0.37
K=900	0.66	0.35	0.73	0.48	**0.81**	**0.64**	0.78	0.59
K=1000	0.79	0.61	0.76	0.55	**0.82**	**0.66**	0.81	0.65

Tabla B.20: Utilizando un modelo basado en Estilo las k palabras más frecuentes del corpus y DV-SA

CoAID								
DV-SA_Estilo	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
FCE	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.79	0.60	0.67	0.37	**0.87**	**0.75**	0.57	0.18
K=100	0.69	0.41	0.60	0.22	**0.71**	**0.45**	0.60	0.22
K=200	0.60	0.25	0.67	0.36	**0.77**	**0.58**	0.64	0.32
K=300	0.57	0.18	0.65	0.33	**0.65**	**0.34**	0.64	0.31
K=400	0.68	0.40	**0.71**	**0.46**	0.57	0.17	0.62	0.27
K=500	0.65	0.33	0.60	0.23	0.54	0.12	**0.73**	**0.50**
K=600	0.63	0.28	**0.78**	**0.59**	0.68	0.40	0.72	0.47
K=700	**0.69**	**0.42**	0.69	0.41	0.63	0.30	0.66	0.35
K=800	**0.71**	**0.45**	0.62	0.27	0.64	0.31	0.70	0.42
K=900	0.67	0.36	0.79	0.61	**0.81**	**0.65**	0.67	0.36
K=1000	0.68	0.40	**0.77**	**0.57**	0.67	0.36	0.67	0.36

Tabla B.21: Utilizando un modelo basado en Estilo las k palabras más frecuentes del cálculo del FCE del vocabulario del corpus y DV-SA

CoAID								
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabrs_frecuentes_idioma	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.79	0.60	0.67	0.37	**0.87**	**0.75**	0.57	0.18
K=100	0.74	0.52	**0.77**	**0.57**	**0.77**	**0.57**	0.69	0.41
K=200	0.63	0.28	0.72	0.47	0.69	0.42	**0.82**	**0.67**
K=300	0.74	0.52	0.74	0.52	0.73	0.50	**0.82**	**0.66**
K=400	0.71	0.45	0.70	0.43	0.73	0.48	**0.86**	**0.74**
K=500	0.73	0.50	0.72	0.47	**0.83**	**0.67**	**0.83**	**0.67**
K=600	0.69	0.41	0.68	0.39	**0.76**	**0.55**	0.73	0.50
K=700	0.70	0.44	**0.72**	**0.46**	0.69	0.41	0.71	0.46
K=800	0.74	0.51	0.72	0.46	**0.75**	**0.53**	0.73	0.50
K=900	**0.82**	**0.66**	0.73	0.50	0.71	0.46	0.70	0.43
K=1000	**0.75**	**0.53**	0.74	0.51	0.69	0.41	0.71	0.45

Tabla B.22: Utilizando un modelo basado en Contenido las k palabras más frecuentes del idioma y DV-SA

	CoAID							
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
palabrs_frecuentes_corpus	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.79	0.60	0.67	0.37	**0.87**	**0.75**	0.57	0.18
K=100	0.71	0.46	0.77	0.56	0.81	0.63	**0.88**	**0.77**
K=200	0.82	0.66	0.71	0.46	0.78	0.59	**0.83**	**0.67**
K=300	0.78	0.58	0.73	0.50	**0.78**	**0.59**	0.75	0.52
K=400	**0.80**	**0.62**	0.70	0.42	0.79	0.60	0.65	0.33
K=500	0.69	0.42	0.74	0.52	0.80	0.62	**0.81**	**0.65**
K=600	0.76	0.56	**0.81**	**0.64**	0.70	0.43	0.60	0.23
K=700	0.78	0.59	**0.83**	**0.69**	0.73	0.50	0.75	0.52
K=800	0.77	0.57	0.74	0.514	**0.81**	**0.63**	0.69	0.41
K=900	0.73	0.50	0.68	0.40	**0.77**	**0.56**	0.69	0.41
K=1000	**0.83**	**0.69**	0.73	0.48	0.71	0.45	0.73	0.50

Tabla B.23: Utilizando un modelo basado en Contenido las k palabras más frecuentes del corpus y DV-SA

	CoAID							
DV-SA_Contenido	**Kernel=2**		**Kernel=3**		**Kernel=4**		**Kernel=5**	
FCE	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**	**F1-score**	**F1-fake**
Baseline	0.79	0.60	0.67	0.37	**0.87**	**0.75**	0.57	0.18
K=100	0.73	0.48	0.78	0.59	0.77	0.52	**0.83**	**0.68**
K=200	0.72	0.47	**0.79**	**0.61**	0.73	0.50	0.76	0.55
K=300	0.78	0.58	0.75	0.53	0.77	0.57	**0.78**	**0.59**
K=400	0.74	0.51	0.74	0.52	**0.81**	**0.64**	0.69	0.42
K=500	0.76	0.54	0.77	0.56	**0.77**	**0.57**	0.72	0.47
K=600	0.78	0.58	0.71	0.45	0.71	0.46	**0.80**	**0.62**
K=700	**0.83**	**0.69**	0.73	0.50	0.75	0.52	0.75	0.52
K=800	0.70	0.43	0.76	0.54	**0.79**	**0.60**	0.63	0.28
K=900	**0.81**	**0.64**	0.67	0.37	0.71	0.45	0.65	0.33
K=1000	0.78	0.59	**0.83**	**0.68**	0.75	0.53	0.73	0.50

Tabla B.24: Utilizando un modelo basado en Contenido las k palabras más frecuentes del cálculo del FCE del vocabulario del corpus y DV-SA

B.4.2 Etapa de Evaluación y comparación con el Estado del Arte

En esta sección comparamos nuestro mejor resultado con el estado del arte para este conjunto de datos.

Model-CoAID	Precision	Recall	F1-score
SVM	0.4036	0.1322	0.1986
LR	0.4287	0.0690	0.1143
RF	0.6056	0.0581	0.045
CNN	0.9653	0.1238	0.1983
BiGRU	0.7476	0.0524	0.0930
CSI	0.6814	0.2109	0.2283
SAME\v	0.8922	0.2991	0.3400
HAN	0.6965	0.4659	0.5471
dEFEND	**0.8965**	**0.4847**	**0.5814**
Nuestro Modelo: CNN, DV-SA, Contenido, K=100, palabras_frecuentes del corpus, Kernel=5	**0.9134**	**0.7868**	**0.8371**

Tabla B.25: Comparación CoAID con el estado del arte