



**Instituto Politécnico Nacional**  
**Centro de Investigación en Computación**

Detección automática de texto engañoso mediante  
algoritmos de modelos basados en tópicos

**T E S I S**

Que para obtener el grado de  
Doctorado en Ciencias de la Computación

**Presenta**

M. en C. Ángel Hernández Castañeda

**Director de tesis**

Dr. Francisco Hiram Calvo Castro



Ciudad de México, diciembre del 2017



INSTITUTO POLITÉCNICO NACIONAL  
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

SIP-14

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 12:00 horas del día 03 del mes de julio de 2017 se reunieron los miembros de la Comisión Revisora de la Tesis, en la sala designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

**Centro de Investigación en Computación**

para examinar la tesis titulada:

**"Detección automática de texto engañoso mediante algoritmos de modelos basados en tópicos"**

Presentada por el alumno:

**HERNÁNDEZ**  
Apellido paterno

**CASTAÑEDA**  
Apellido materno

**ÁNGEL**  
Nombre(s)

Con registro: 

A	1	4	0	0	9	2
---	---	---	---	---	---	---


aspirante de: **DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**


Después de intercambiar opiniones los miembros de la Comisión manifestaron **APROBAR LA TESIS**, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

**LA COMISIÓN REVISORA**

Director de Tesis


  
Dr. Francisco Hiram Calvo Castro

  
Dr. Edgardo Manuel Felipe Riverón

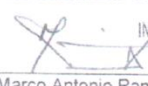
  
Dr. Sergio Suárez Guerra

  
Dr. Alexander Gelbukh

  
Dr. Grigori Sidorov

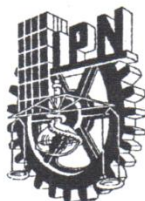
  
Dr. Miguel Jesús Torres Ruiz

**PRESIDENTE DEL COLEGIO DE PROFESORES**

  
Dr. Marco Antonio Ramírez Saldaña



INSTITUTO POLITÉCNICO NACIONAL  
CENTRO DE INVESTIGACIÓN  
EN COMPUTACIÓN



**INSTITUTO POLITÉCNICO NACIONAL**  
**SECRETARÍA DE INVESTIGACIÓN Y POSGRADO**

*CARTA CESIÓN DE DERECHOS*

En la Ciudad de México el día 05 del mes diciembre del año 2017, el (la) que suscribe Ángel Hernández Castañeda alumno (a) del Programa de Doctorado en Ciencias de la Computación con número de registro A140092, adscrito a Laboratorio de Inteligencia Artificial, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de Dr. Francisco Hiram Calvo Castro y cede los derechos del trabajo intitulado Detección de texto engañoso mediante algoritmos de modelos basados en tópicos, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección angelhc2305@gmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

  
Ángel Hernández Castañeda

Nombre y firma

# Resumen

En este trabajo abordamos diferentes casos de estudio. En primer lugar, dado que las características basadas en LDA mostraron un buen desempeño cuando fueron evaluadas en cada conjunto de datos por separado (ver Sección 5.1.1), se realizaron experimentos mezclando todos los conjuntos de datos clasificados en el presente estudio para verificar si las características obtenidas de una mezcla de dominios (ver Sección 5.1.2) mejoraban los resultados o, por el contrario, los empeoraban. Además, exploramos si un conjunto de características puede ser suficientemente general para ser usado en la clasificación de un conjunto de datos de un tema diferente al tema del conjunto de datos utilizado para su formación, lo que permitiría crear detectores de texto engañoso de propósito general independientes del dominio (ver Sección 5.1.3).

Para ello, hemos generado características mediante el uso de varios métodos, tales como la latent Dirichlet allocation (LDA), n-gramas sintácticos (ns-gramas), linguistic inquiry and word count (LIWC), y un modelo de espacio de palabras (MTD), así como las combinaciones de características generadas por los diferentes métodos. Para probar la eficiencia de cada método, utilizamos tres conjuntos de datos sobre diferentes temas, específicamente: OpSpam, que consiste en opiniones sobre hoteles, DeRev, que consiste en opiniones sobre libros comprados en Amazon, y el conjunto de datos de tópicos controversiales, compuesto de opiniones sobre tres temas (aborto, pena de muerte y mejor amigo). Con base en los conjuntos de datos obtenidos, se investiga qué método es mejor en un único dominio, donde tanto el conjunto de entrenamiento y de prueba son del mismo tema, en un entorno de dominio mixto, donde el conjunto de entrenamiento y de prueba consisten en una mezcla de temas, y en un entorno de dominio cruzado, donde los conjuntos de entrenamiento y prueba están formados de diferentes temas (sin intersección entre prueba y entrenamiento).

# Abstract

We identify deceptive text by using different kinds of features: a continuous semantic space model based on latent Dirichlet allocation topics (LDA), one-hot representation (OHR), syntactic information from syntactic n-grams (SN), and lexicon-based features using the linguistic inquiry and word count dictionary (LIWC). Several combinations of these features were tested to assess the best source(s) for deceptive text identification. By selecting the appropriate features, we were able to obtain a benchmark-level performance using a Naïve Bayes classifier. We tested on three different available corpora: a corpus consisting of 800 reviews about hotels, a corpus consisting of 600 reviews about controversial topics, and a corpus consisting of 236 book reviews. We found that the merge of both LDA features and OHR yielded the best results, obtaining accuracy above 80% in all tested datasets. Additionally, this combination of features has the advantage that language-specific-resources are not required (*e.g.* SN, LIWC), compared to other reference works. Additionally, we present an analysis on which features lead to either deceptive or truthful texts, finding that certain words can play different roles (sometimes even opposing ones) depending on the task being evaluated.

# *Agradecimientos*

*A Díos.*

*A mis padres, Lucía y Felipe, por todo el apoyo incondicional.*

*A mis hermanos Néstor, Emmanuel, Brisa, Brenton y Jeshua por haber superado junto conmigo cada obstáculo que se presentó.*

*A mi esposa Claudia, por estar a mi lado en todo momento.*

*A mi abuela Andrea Guadalupe, por cuidar mis pasos.*

*A mi asesor, el Dr. Hiram Calvo, por brindarme su experiencia y conocimiento.*

*Al Dr. Salvador Godoy por invertir su valioso tiempo en el seminario de inteligencia artificial el cual beneficia a todos los estudiantes que participamos en él.*

*A mis sinodales, Dr. Edgardo, Dr. Alexander, Dr. Grigori, Dr. Sergio, y Dr. Miguel por toda la retroalimentación que indudablemente ayudó a mejorar el presente trabajo de investigación.*

*Al Centro de Investigación en Computación y al CONACYT.*

# Índice

1	Capítulo I. Introducción.....	1
1.1	La detección del texto engañoso.....	2
1.2	Motivación .....	5
1.3	Descripción del problema .....	6
1.4	Objetivos.....	7
1.4.1	Objetivo general .....	7
1.4.2	Objetivos específicos .....	7
1.5	Aportaciones .....	8
1.6	Organización de la tesis .....	9
2	Capítulo II. Procesamiento del lenguaje natural .....	10
2.1	Procesamiento de texto.....	11
2.2	Tokenización .....	13
2.3	Análisis sintáctico .....	14
2.3.1	Gramáticas libres del contexto .....	15
2.3.2	Árboles sintácticos .....	16
2.3.3	Robusto.....	18
2.3.4	Desambiguación .....	19
2.4	Análisis semántico .....	20
2.4.1	Teorías y enfoques de la representación semántica.....	22
2.4.2	Enfoques lógicos.....	22
2.4.3	Teoría de representación de discurso .....	23
3	Capítulo III. Estado del arte .....	27
3.1	Introducción .....	28

3.2	El detector de engaño, explorando en el reconocimiento automático del texto engañoso .....	30
3.3	Identificación de engaño por medio de algún estiramiento de la imaginación ...	34
3.4	El uso de diferentes conjuntos de sujetos homogéneos en el análisis de texto engañoso .....	36
3.5	Detección intercultural de engaño.....	38
3.6	Identificando opiniones engañosas mediante el método de aprendizaje de multitudes .....	42
4	Capítulo IV. Enfoque propuesto .....	46
4.1	Fuentes de generación de características .....	47
4.1.1	Matriz término-documento (MTD) .....	48
4.1.2	N-gramas sintácticos .....	48
4.1.3	Linguistic Inquiry and Word Count.....	51
4.1.4	Modelo de espacio semántico continuo - Latent Dirichlet Allocation (LDA)	52
4.2	Conjuntos de datos .....	54
4.2.1	Conjunto de datos DeRev .....	55
4.2.2	Conjunto de datos OpSpam .....	56
4.2.3	Conjunto de datos de tópicos controversiales .....	58
4.2.4	Análisis de los conjuntos de datos .....	59
4.3	Construcción de los vectores de características .....	60
4.4	Identificación del número de tópicos óptimo .....	64
5	Capítulo V. Resultados .....	65
5.1	Detección de engaño.....	66
5.1.1	Clasificación de un dominio específico.....	66
5.1.2	Clasificación de dominios mezclados .....	73
5.1.3	Clasificación de dominio cruzado .....	74



5.1.4	Palabras y tópicos predominantes .....	76
5.1.5	Comparación de resultados y significancia estadística .....	79
5.2	Impacto de la polaridad de textos en la detección de engaño .....	82
5.2.1	Clasificación de polaridad para mejorar la detección de engaño .....	84
6	Conclusiones y trabajo futuro .....	88
	Referencias .....	90

# Índice de figuras

Figura 1. Ejemplo de árbol sintáctico .....	16
Figura 2. Ejemplo de construcción de vectores mediante un modelo de espacio de palabras .....	49
Figura 3. Proceso de generación de bigramas sintácticos dada una oración .....	51
Figura 4. Ejemplo de algunos grupos de palabras, en LIWC, con su respectiva etiqueta. Es mostrada una muestra de solo 7 palabras por etiqueta en este ejemplo. ....	52
Figura 5. Ejemplo de generación de tópicos, de textos del tema “pena de muerte”, mediante LDA .....	53
Figura 6. Ejemplo de documento procesado por LDA. Se muestra la distribución de tópicos. ....	54
Figura 7. Curva de aprendizaje de los diferentes corpora .....	73

# Índice de tablas

Tabla 1. Clases de LIWC dominantes del texto engañoso y veraz .....	32
Tabla 2. Exactitud obtenida por tópico usando una validación cruzada de 10 pliegues .....	33
Tabla 3. Exactitud, precisión (P), cobertura (R), y medida F (F) de la identificación de texto engañoso realizada por humanos.....	34
Tabla 4. Resultados de la aplicación de diferentes fuentes de generación de características y aplicación de dos clasificadores. ....	35
Tabla 5. Resultados de la clasificación de corpus DeCour. Se muestran: precisión (P), exhaustividad (R) y medida F (F), asimismo el promedio de medidas F.....	38
Tabla 6. Experimentos realizados sobre los corpora de cada país, por separado .....	40
Tabla 7. Clasificación de cultura cruzada con LIWC y unigramas como fuente de características.....	41
Tabla 8. Clases de LIWC predominantes en el texto engañoso y en el texto veraz...	42
Tabla 9. Tabla de clasificación del conjunto DeRev que muestra el rendimiento de los métodos de etiquetado.....	44
Tabla 10. Comparación entre bigramas sintácticos y convencionales en la tarea de identificación de autoría .....	50
Tabla 11. Comparación entre trigramas sintácticos y convencionales en la tarea de identificación de autoría .....	50
Tabla 12. Se muestran la cantidad de tipos y tokens en los corpora .....	59
Tabla 13. Promedio de tipos compartidos en los diferentes corpora .....	60
Tabla 14. Comparación de exactitud con respecto a los valores binarios y la selección de atributos (SeAt) .....	61
Tabla 15. Exactitud obtenida con diferentes valores del número de tópicos.....	63
Tabla 16. Clasificación del corpus OpSpam mediante el uso de diferentes fuentes de generación de características .....	67
Tabla 17. Número de características relevantes obtenidas con la combinación de LDA y MTD .....	68

Tabla 18. Clasificación del corpus DeRev mediante el uso de diferentes fuentes de generación de características .....	69
Tabla 19. Clasificación del tópico Aborto mediante el uso de diferentes fuentes de generación de características .....	70
Tabla 20. Clasificación del tópico "mejor amigo" mediante el uso de diferentes fuentes de generación de características .....	71
Tabla 21. Clasificación del tópico "pena de muerte" mediante el uso de diferentes fuentes de generación de características .....	72
Tabla 22. Exactitud, precisión (P), exhaustividad (R) y medida F (F) obtenidas con los corpora mezclados mediante SVM.....	74
Tabla 23. Exactitud, precisión (P), exhaustividad (R) y medida F (F) obtenidas con los corpora mezclados mediante NB .....	74
Tabla 24. Medida F obtenida en la clasificación de dominio cruzado. Se muestran resultados para los clasificadores NB y SVM .....	75
Tabla 25. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos OpSpam .....	77
Tabla 26. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos "mejores amigos" .....	77
Tabla 27. Los diez tópicos más relevantes y palabras representantes del texto engañoso y veraz en el conjunto de datos "Aborto" .....	78
Tabla 28. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos "pena de muerte" .....	79
Tabla 29. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos "DeRev" .....	80
Tabla 30. Comparación de nuestros resultados con otros estudios de los mismos corpora.....	81
Tabla 31. Significancia estadística .....	81
Tabla 32. Resultados de la clasificación excluyendo palabras de polaridad .....	84
Tabla 33. Clasificación del corpus OpSpam, etiquetado por los autores (OAP) y el etiquetado realizado en este estudio (PTW), basado en la polaridad del texto.....	85

Tabla 34. Clasificación de los conjuntos de datos agregando característica de polaridad .....	85
Tabla 35. Comparación de la medida F obtenida antes (F-B) y después (Avg. F) de la clasificación de polaridad. Las medidas precision (P), Recall (R), medida F (F), y el número de opiniones por clase (# docs) son también mostrados .....	86

# Capítulo I. Introducción

## *Contenido*

La detección de texto engañoso  
Motivación  
Descripción del problema  
Objetivo general  
Aportaciones de la tesis  
Organización de la tesis

## **1.1 La detección del texto engañoso**

El estudio del engaño se ha perfeccionado con el tiempo, desde el análisis del rostro, por ejemplo, mediante el movimiento de los ojos y las expresiones faciales; hasta el monitoreo de las respuestas fisiológicas del ser humano mediante un polígrafo. Este último enfoque tiene como objetivo medir variables, tales como la presión arterial, el ritmo cardíaco y la frecuencia respiratoria con lo cual ha logrado resultados óptimos (con una exactitud por encima del 95%) en la detección de un engaño. Los principios del funcionamiento de estos detectores de engaño se basan principalmente en la hipótesis que mentir produce cierto estado emocional; y para cada estado emocional existe una reacción fisiológica identificable por el detector de engaño.

De manera similar, investigadores del procesamiento de lenguaje natural han tratado de identificar el engaño en el texto. Con la hipótesis que cuando la gente trata de engañar, mediante la escritura, usa ciertas palabras que en un texto real no usaría. Es decir, el engaño propicia un estado emocional que cambia nuestra forma de escribir. Por ejemplo, cuando al estar realizando una tarea en especial pasamos de nuestro estado normal a una situación muy estresante es posible que no recordemos o que omitamos ciertos detalles de la tarea en marcha debido al momento de estrés en el que estamos. Escribir un texto que sabemos que tiene que engañar a una persona o conjunto de personas puede generar, de forma consciente o inconsciente, un cambio en el comportamiento; o mejor dicho en la forma habitual en que realizamos una acción. Este cambio se genera debido a que el cerebro siempre está preparado para decir la verdad, sin embargo, para mentir debe incrementar el rendimiento de los procesos cognitivos.

“Un mentiroso puede elegir no mentir. Engañar a la víctima es un acto deliberado; el mentiroso pretende desinformar a la víctima. El mentiroso puede estar o no estar justificado, bajo su opinión o bajo la opinión de la comunidad. El mentiroso puede ser buena o mala persona, agradable o desagradable. Sin embargo, la persona que ha mentido pudo elegir engañar o ser veraz, y sabe la diferencia entre ambas” [8].

Se puede dar un debate filosófico de lo que puede significar el concepto mentir, sin embargo, para fines de este estudio, consideramos que el engaño se presenta cuando el mentiroso pretende transferir a la víctima información total o parcialmente falsa, es decir, hacer creer a la víctima sobre un hecho que no sucedió o que sucedió de manera diferente.

Hoy en día, los usuarios realizan una variedad de actividades en línea tales como comprar y vender artículos, difundir ideas a través de blogs e intercambiar información en general. Dicha información no siempre es confiable: algunas personas utilizan Internet para transmitir información con el propósito de manipular y engañar a otros usuarios. Por ejemplo, cuando un usuario quiere comprar un artículo en línea, la principal manera de saber si el producto es bueno, o no, es leer la sección de opiniones sobre el mismo en la página web del vendedor. Tales opiniones han demostrado tener un gran impacto en la decisión final de adquirir o no, el artículo. Debido a esto, algunos vendedores contratan personas, algunas veces expertas, para escribir opiniones positivas con el fin de aumentar las ventas de un producto, incluso si esas personas no tienen una idea real sobre la calidad o cualidad del artículo. En otros casos, las opiniones engañosas pretenden desacreditar los productos ofrecidos por los competidores.

Aparte de los textos engañosos escritos para manipular las decisiones de compra de los usuarios, también existen textos engañosos que pretenden cambiar la opinión o el punto de vista de las personas sobre un tema determinado, como un candidato político o un tema de debate público.

Esto hace muy importante el estudio de la detección de textos engañosos. La tarea, básicamente, puede definirse como la identificación de aquellas opiniones escritas en las que el autor pretende transmitir información en la que no cree [24]. Los estudios sobre textos engañosos han demostrado empíricamente que la comunicación veraz es cualitativamente diferente de la comunicación engañosa [7], [41]. Debido a esto, se han emprendido diferentes proyectos con el objetivo de identificar los textos engañosos con la mayor precisión posible. Para ello, se han creado varios conjuntos



de datos. Dichos conjuntos de datos consisten en textos etiquetados como verídicos o engañosos.

En el enfoque de aprendizaje automático, una parte de los textos del conjunto de datos se utiliza como un conjunto de entrenamiento para un clasificador y el resto como un conjunto de prueba. Por lo tanto, una comparación directa entre diferentes clasificadores y métodos de selección de características es posible mediante la aplicación de estos sobre el mismo conjunto de datos.

En este trabajo abordamos diferentes casos de estudio. En primer lugar, dado que las características basadas en LDA mostraron un buen desempeño cuando fueron evaluadas en cada conjunto de datos por separado (ver Sección 5.1.1), se realizaron experimentos mezclando todos los 3 conjuntos de datos clasificados en el presente estudio para verificar si las características obtenidas de una mezcla de dominios (ver Sección 5.1.2) mejoraban los resultados o, por el contrario, los empeoraban. Además, exploramos si un conjunto de características puede ser suficientemente general para ser usado en la clasificación de un conjunto de datos que trate de un tema diferente al del conjunto de datos utilizado para su formación, lo que permitiría crear detectores de texto de engaño de propósito general independientes del dominio (ver Sección 5.1.3).

Para ello, hemos generado características mediante el uso de varios métodos, tales como la latent Dirichlet allocation (LDA), linguistic inquiry and word count (LIWC), y un modelo de espacio de palabras (MTD), así como combinaciones de características generadas por los diferentes métodos. Para probar la eficiencia de cada método, utilizamos tres conjuntos de datos sobre diferentes temas, específicamente: OpSpam, que consiste en opiniones sobre hoteles; DeRev, que consiste en opiniones sobre libros comprados en Amazon; y el conjunto de datos de tópicos controversiales, compuesto de opiniones sobre tres temas (aborto, pena de muerte y mejor amigo). Con base en los conjuntos de datos obtenidos, se investiga qué método es mejor en un único dominio, donde tanto el conjunto de entrenamiento y de prueba son del mismo tema, en un entorno de dominio mixto, donde tanto el conjunto de entrenamiento y de prueba consisten en una mezcla de temas, y en un

entorno de dominio cruzado, donde los conjuntos de entrenamiento y prueba están formados de diferentes temas (sin intersección entre prueba y entrenamiento).

Con estos experimentos, evaluamos la posibilidad de utilizar conjuntos de datos existentes para detectar textos engañosos sobre un tema para el que no hay un conjunto de datos disponible, es decir, la posibilidad de desarrollar un detector de texto engañoso independiente del dominio.

## **1.2 Motivación**

Actualmente la venta de productos y servicios está migrando a la web. A diferencia de una venta en donde el producto está físicamente presente para poder revisarlo, en una venta online sólo se cuenta con un conjunto de imágenes para revisar lo que se está comprando.

Los vendedores consideraron que las imágenes no bastaban para que el comprador tuviera la seguridad de que el producto cumpliría sus expectativas. Por lo tanto, se añadió una sección en donde los compradores pueden agregar sus comentarios después de tener y probar el producto. Estos comentarios servirían para que los futuros compradores pongan en tela de juicio su compra, es decir, si las opiniones son positivas en su mayoría, entonces el comprador sabrá que el producto es bueno, de otra forma, si existen demasiadas opiniones negativas, el comprador tendrá los argumentos necesarios para tomar la decisión de comprar, o no, el producto.

Esto fue una buena idea hasta que los compradores lo vieron como una forma de sacar ventaja y aumentar sus ventas. Como resultado, diferentes páginas web que ofrecen productos buscan personas que se dedican a escribir opiniones engañosas acerca de los productos, estas opiniones dirán que el producto es muy bueno cuando en realidad la persona que lo escribió tendrá como fin ayudar al vendedor a generar más ventas.

Las opiniones engañosas no son necesariamente un argumento positivo hacia el producto. Puede existir el caso en que la competencia quiera desprestigiar a

determinado vendedor, por lo cual tendrá que agregar un comentario que no favorezca al producto en cuestión. Esta opinión será por tanto negativa, y también engañosa.

En un artículo<sup>1</sup> publicado en la revista *The Economist* en el 2015 se concluyó que más de 1,000 usuarios en amazon<sup>2</sup> se dedican a escribir opiniones engañosas. Además de amazon, en muchos otros sitios web existen este tipo de opiniones debido a que la motivación es clara, un aumento en las ventas. Por ejemplo, de acuerdo al artículo, un restaurante con una estrella más en Yelp (un sitio popular de opiniones) aumentará sus ingresos de un 5% a un 9%.

El problema se centra en que los humanos tenemos problemas para detectar el engaño. Por ejemplo, en la investigación de Pérez-Rosas et al. [34] se verifica el rendimiento de tres jueces humanos en la detección de engaño en el texto, audio y video. Asimismo, se muestra que el rendimiento oscila entre 54.2-65.3%, 58.5-70.3%, 63.0-71.0% para texto, audio y video, respectivamente. Se puede notar que en el caso del texto es más difícil poder diferenciar el engaño de la verdad. Esto puede deberse a que no podemos observar al sujeto que está mintiendo, por lo tanto, no se puede obtener cierta información, por ejemplo, si el sujeto está nervioso, con voz agitada, está sudando o alguna señal que nos ayude a mejorar el rendimiento en la identificación del engaño.

Dado lo anterior, es necesaria la investigación de nuevos enfoques que permitan detectar, con mayor eficiencia, el engaño en los textos. De esta manera, se proveerán mecanismos que ayuden a los compradores de distintas páginas web a que puedan realizar compras más seguras y confiables.

### **1.3 Descripción del problema**

La detección de engaño en un texto es un reto aún mayor que detectar engaño en una persona debido a que solo contamos con las palabras incluidas en el documento

---

<sup>1</sup> <http://www.economist.com/news/business/21676835-evolving-fight-against-sham-reviews-five-star-fakes>

<sup>2</sup> [www.amazon.com](http://www.amazon.com)

o colección de documentos. Es decir, no tenemos a la persona para monitorear sus reacciones fisiológicas, sus expresiones faciales o incluso su comportamiento.

En distintos estudios se han tratado de buscar palabras que identifiquen al texto engañoso, sin embargo, podemos formular las siguientes preguntas, ¿estas palabras se mantienen para distintos tipos de engaño?, ¿las palabras que identifican al texto engañoso pueden cambiar si el tema al que se refiere el texto cambia también?, ¿cuál es el método de generación de características que genera un modelo que represente de la mejor manera al texto engañoso?, ¿la combinación de distintos métodos puede generar un mejor resultado en la detección de texto engañoso?.

Es difícil saber si las personas tienen un comportamiento predeterminado al realizar cierta acción tal como mentir. Sin embargo, con base en distintos trabajos que se realizaron previamente (ver Sección 3), se ha logrado identificar que los mentirosos siguen un patrón de uso de palabras; por ejemplo, al momento de mentir pueden ser poco específicos, usar en mayor o menor cantidad ciertas palabras. En otras palabras, mediante algoritmos de reconocimiento de patrones y aprendizaje automático, es posible distinguir un texto engañoso de un texto veraz; sin embargo, la falta de conjuntos de datos (o los conjuntos generalmente pequeños de datos) para realizar experimentos hace que los enfoques propuestos solo sean probados en entorno demasiado específico.

## **1.4 Objetivos**

### **1.4.1 Objetivo general**

Mejorar la detección de engaño utilizando vectores de características compuestos de múltiples métodos.

### **1.4.2 Objetivos específicos**

- Experimentar con modelos basados en tópicos, particularmente LDA (latent Dirichlet allocation) con el fin de detectar el engaño en el texto.

- Combinar características generadas mediante diferentes enfoques con el objetivo de mejorar la eficiencia de la detección de engaño.
- Evaluar el desempeño de enfoques seleccionados (por separado y combinados) en diferentes casos de estudio.
- Evaluar el rendimiento de los métodos y combinación de métodos en la clasificación de un dominio específico.
- Evaluar el rendimiento de los métodos y combinación de métodos en la clasificación de un dominio mezclado.
- Evaluar el rendimiento de los métodos y combinación de métodos en la clasificación de un dominio cruzado.
- Evaluar el impacto de la polaridad en la clasificación de textos engañosos.

## 1.5 Aportaciones

Mediante el presente estudio se han logrado las siguientes aportaciones.

- Se encontró una combinación de características usando métodos que no requieren información a priori para generar las mismas, con lo cual se pueden procesar textos en diferentes idiomas sin necesidad de requerir corpus para cada caso, siempre y cuando existan herramientas para generar los tokens.
- Fue encontrada una combinación de métodos de generación de características basado en LDA y una matriz de palabras que mejora la eficiencia en la detección de texto engañoso con respecto al uso de relaciones sintácticas y LIWC.
- Se demostró que el uso de LDA es una alternativa viable y que genera una exactitud mayor que el uso de LIWC. Además de ser gratuita.
- Se publicó un artículo en la revista indizada *SoftComputing*. Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., & Flores, J. J. G. (2017). Cross-domain deception detection using support vector networks. *Soft Computing*, 21(3), 585-595.
- Se publicó un artículo en la revista indizada *Intelligent Data Analysis*. Hernández-Castañeda, Á., & Calvo, H. (2017). Deceptive text detection using continuous semantic space models. *Intelligent Data Analysis*, 21(3), 679-695.

- Se publicó un artículo en la revista *Computación y Sistemas*.  
Hernández-Castañeda, Á., & Calvo, H. (2017). Author Verification Using a Semantic Space Model. *Computación y Sistemas*, 21(2).
- Se publicó un artículo en la revista indizada *Journal of Intelligent & Fuzzy Systems*  
Hernández-Castañeda, Á., Calvo, H., Juárez Gambino O. Impact of polarity in deception detection. *Journal of Intelligent & Fuzzy Systems*, to appear.

## 1.6 Organización de la tesis

La presente tesis consta de seis capítulos. En el primer capítulo se da una breve introducción del texto engañoso y de los casos de estudio manejados en la presente tesis, además de especificar la motivación, objetivos y aportaciones de la misma. En el segundo capítulo mostramos un breve fundamento teórico que sirvió como base para las herramientas de generación de características usadas. En el tercer capítulo, mostramos un resumen de diferentes estudios que proponen enfoques, mediante aprendizaje automático, para la identificación de texto engañoso. En el capítulo cuatro, se aborda el enfoque propuesto en el presente estudio, además de los conjuntos de datos analizados y clasificados. En el capítulo cinco, mostramos los resultados obtenidos en diferentes casos de estudio. Finalmente, en el capítulo seis mostramos las conclusiones.

## Capítulo II. Procesamiento del lenguaje natural

### *Contenido*

En este capítulo se presenta un breve fundamento teórico [22] que brinda el conocimiento básico para comprender las diferentes herramientas de generación de características usadas en el presente estudio.

## **2.1 Procesamiento de texto**

Los lenguajes naturales contienen ambigüedades inherentes, además los lenguajes escritos generan aún más ambigüedad.

El procesamiento de texto es la tarea de convertir una cantidad de archivos de texto en una secuencia bien definida de unidades lingüísticas significativas. El procesamiento de texto es una parte esencial de cualquier sistema de procesamiento de lenguaje natural (PLN) debido a que los caracteres, palabras y sentencias identificadas en esta fase, son las unidades fundamentales que pasan a las siguientes fases de procesamiento. Estas fases pueden ser de análisis y de etiquetado, por ejemplo, análisis morfológico y etiquetado de categorías gramaticales, por medio de aplicaciones como recuperación de información y sistemas de traducción automática.

El procesamiento de texto puede dividirse en dos etapas. En primer lugar, el proceso de convertir un conjunto de archivos digitales en un conjunto de documentos de texto bien definidos. Esto implica, también, determinar los algoritmos para analizar el lenguaje específico de los documentos; la identificación del lenguaje determina entonces, el lenguaje natural del cual están formados los documentos. Este paso está estrechamente relacionado con la codificación de caracteres (aunque esto no lo determina únicamente).

En segundo lugar, está la segmentación de texto que es el proceso de convertir un corpus de texto bien definido en componentes como palabras y sentencias. La segmentación consiste en partir o cortar las secuencias de caracteres en el texto mediante la localización de los límites de las palabras, es decir, los puntos donde comienza y termina una palabra. Para los propósitos de la lingüística computacional, a las palabras identificadas, con la segmentación de texto, se les da el nombre de tokens, además el proceso de segmentación también es conocido como tokenización.

La normalización del texto es un proceso relacionado que incluye todas las mezclas de la escritura de un token en una forma canónica normalizada; por ejemplo, las diferentes formas del token: Señor, Sr., señor, SEÑOR. Por otra parte, también se



puede incluir el proceso de segmentación de sentencias que identifica dónde comienza y termina una sentencia.

El tipo de sistema de escritura usado por un lenguaje es el factor más importante para determinar el mejor enfoque de preprocesamiento. Los sistemas de escritura pueden ser logográficos, donde un extenso número de símbolos individuales representan palabras; silábicos, donde los símbolos individuales representan sílabas; alfabéticos, donde los símbolos individuales representan sonidos. En la práctica, no hay sistemas modernos de escritura que usen símbolos de un solo tipo, por lo tanto, ningún sistema de lenguaje natural de escritura puede ser clasificado como puramente logográfico, silábico o alfabético. El inglés, por ejemplo, está basado en el alfabeto Romano y utiliza símbolos logográficos incluyendo los números arábigos, símbolos de moneda, entre otros. El mismo, sin embargo, es predominantemente alfabético, y muchos de los otros sistemas de escritura están compuestos por símbolos que son mayormente de un tipo.

La tokenización está inevitablemente relacionada con la subyacente codificación de caracteres del texto que se va a procesar, y la identificación de caracteres es un primer paso esencial. Mientras la cabecera de un documento digital puede contener la información correspondiente con su codificación de caracteres, esta información no está siempre presente o incluso disponible, en este caso, la codificación debe ser determinada automáticamente.

Además de la variedad de tipos de símbolos (logográficos, silábicos, o alfabéticos) usados en sistemas de escritura, existen un conjunto de convenciones ortográficas en los lenguajes escritos para denotar los límites entre unidades lingüísticas tal como sílabas, palabras o sentencias. El inglés maneja los espacios en blanco entre palabras y signos de puntuación como límite de sentencia; sin embargo, incluso esto no es suficiente para segmentar completamente el texto solventando las ambigüedades.

Muchos de los algoritmos de segmentación de lenguajes naturales existentes son para lenguajes específicos y dependientes de corpus, desarrollados para manejar las ambigüedades predecibles en un texto bien formado. Dependiendo del origen y del propósito del texto, el uso de letras mayúsculas y las reglas de puntuación pueden

ser respetadas como en los documentos científicos, medianamente respetadas como en algunas páginas web informativas, y prácticamente descuidadas como en las redes sociales.

Aunque la segmentación de palabras y sentencias es necesaria, en realidad, no existe una definición absoluta sobre lo que constituye cada una de estas. Ambas tienen distinciones relativamente arbitrarias que cambian fuertemente en los diferentes lenguajes escritos. Sin embargo, para propósitos de la computación lingüística, se necesita definir exactamente qué se necesita para un procesamiento posterior; en muchos casos, el lenguaje y la aplicación determinan las convenciones necesarias.

## **2.2 Tokenización**

La tokenización, actualmente, está bien establecida y bien entendida para los lenguajes artificiales, tal como los lenguajes de programación. Sin embargo, tales lenguajes artificiales pueden ser estrictamente definidos para eliminar las ambigüedades léxicas y estructurales; en el caso de los lenguajes naturales no se puede contar con la misma definición, en este tipo de lenguajes una palabra puede servir para diferentes propósitos y la sintaxis no está estrictamente definida.

Existen lenguajes delimitados por espacios, tal como muchos lenguajes europeos en los cuales los límites de palabra son indicados por la inserción de un espacio en blanco. Por otra parte, también hay lenguajes no segmentados, como el chino y thai donde las palabras son escritas consecutivamente sin indicar ningún límite entre ellas.

Existen tres categorías principales en las cuales las estructuras de las palabras pueden ser colocadas, y cada categoría existe en ambos sistemas de escritura: delimitados por espacio y no segmentados. La morfología de las palabras en un lenguaje puede ser aislado, cuando las palabras no se dividen en pequeñas unidades; aglutinante, cuando las palabras se dividen en pequeñas unidades, llamadas morfemas, con límites claros entre ellos; o flexional, cuando los límites

entre morfemas no son claros y los morfemas componentes pueden expresar más de un sentido gramatical.

Aun contando con un corpus de sentencias bien formado, existen varios tópicos a resolver en la tokenización. Mucha de la ambigüedad, al tokenizar, se presenta por el uso de signos de puntuación debido a que el mismo signo de puntuación puede tener diferentes funciones en una sola sentencia.

Una idea inicial en la tokenización de un lenguaje delimitado por espacios podría ser considerar como un token separado cualquier secuencia de caracteres precedidos y seguidos de un espacio. Esto tokeniza exitosamente palabras que son una secuencia de caracteres alfabéticos, pero no toma en cuenta caracteres de puntuación. En muchos casos, los caracteres tal como comas, punto y coma y puntos deben ser tratados como tokens separados.

La tarea de tokenización en los lenguajes no segmentados necesita un enfoque más elaborado que un simple análisis léxico.

### **2.3 Análisis sintáctico**

Es conveniente considerar los aspectos en los cuales difieren los analizadores sintácticos aplicados a lenguajes naturales y aquellos aplicados a lenguajes de programación. Una de esas diferencias respecta al poder de los formalismos de la gramática usada, es decir, la capacidad generativa de la gramática. Los lenguajes de programación son usualmente diseñados para permitir codificar mediante una gramática no ambigua. Para lograrlo se restringen cuidadosamente las subclases de la gramática libre del contexto (CFG, por sus siglas en inglés *context-free grammar*). Por el contrario, los lenguajes naturales requieren normalmente instrumentos más poderosos. Uno de los casos más destacables del poder expresivo han sido las dependencias de larga distancia, por ejemplo, en el inglés, las palabras interrogativas.

Otra diferencia tiene que ver con la estructura fuertemente ambigua del lenguaje natural. En el recorrido de la sentencia, habrá típicamente reglas gramáticas que podrían aplicar. Un ejemplo clásico es el siguiente: “Pon el cuadro en la caja sobre la

mesa". Si asumimos que "pon" subcategoriza a dos objetos, entonces existen dos posibles análisis: "Pon el cuadro [en la caja sobre la mesa]" y "Pon [el cuadro en la caja] sobre la mesa".

El propósito de una gramática general podría ser capturar lo que es posible en cualquier contexto. Mucho del trabajo en el análisis sintáctico trata con las formas en las cuales los potenciales espacios de búsqueda enormes pueden ser tratados eficientemente, y cómo se puede elegir el análisis más apropiado.

La tercera diferencia se basa en el hecho de que los textos en lenguaje natural son inherentemente ruidosos. A diferencia del lenguaje natural, un lenguaje de computadora tiene una sintaxis específica completa, lo cual significa que, por definición, toda cadena de entrada correcta, es analizable. En el análisis del lenguaje natural, es notoriamente difícil distinguir si una falla que produce el resultado del análisis es debido a un error de entrada o a la falta de cobertura de la gramática.

### **2.3.1 Gramáticas libres del contexto**

Desde su introducción por Chomsky [4], las gramáticas libres del contexto (CFG) han sido el formalismo gramatical más influyente para describir la sintaxis de un lenguaje. Esto debido a que muchos de los formalismos gramaticales se derivan o de alguna manera pueden estar relacionados con las CFG.

La forma estándar de definir una CFG es mediante la tupla  $G = (\Sigma, N, S, R)$ , donde  $\Sigma$  y  $N$  son conjuntos finitos disjuntos con símbolos terminales y no terminales, respectivamente, y el conjunto  $V = \Sigma \cup N$  contiene los símbolos de la gramática. Por otra parte,  $R$  es un conjunto finito de reglas de producción de la forma  $A \rightarrow \alpha$ , donde  $A \in N$  es un no terminal y  $\alpha \in V^*$  es una secuencia de símbolos.

Se usan letras mayúsculas para identificar a los símbolos no terminales y letras minúsculas para identificar a los símbolos terminales.

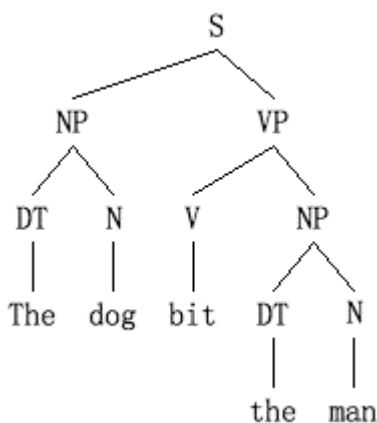
Una gramática está en forma normal de Chomsky (CNF, por sus siglas en inglés Chomsky normal form) cuando cada regla es una regla unaria de terminal de la forma  $A \rightarrow w$ , o una regla binaria de no terminales de la forma  $A \rightarrow BC$ . Siempre es posible transformar una gramática en la CNF tal que acepte el mismo lenguaje. Sin embargo,

la transformación puede cambiar radicalmente la estructura de la gramática. Por ejemplo, si la gramática original tiene  $n$  reglas, la gramática transformada puede tener, en el peor de los casos,  $O(n^2)$  reglas.

Se puede atenuar esta forma normal por medio de permitir reglas unitarias de no terminales de la forma  $A \rightarrow B$ . La transformación para esta forma es más simple y la gramática transformada es estructuralmente parecida.

### 2.3.2 Árboles sintácticos

La forma estándar de representar las estructuras sintácticas de una sentencia gramática es mediante un árbol sintáctico (ver Figura 1), el cual es una representación de todos los pasos de la derivación de una sentencia desde su nodo raíz. Esto significa que cada nodo interno en el árbol representa una aplicación de una regla gramatical.



**Figura 1. Ejemplo de árbol sintáctico**

En la práctica, las CFG no son ampliamente usadas para implementar gramáticas. Una de las razones es que las CFG no son lo suficientemente expresivas, es decir, no describen todas las peculiaridades del lenguaje natural, pero la razón principal es que son muy difíciles de usar.

En el análisis sintáctico participan diferentes procedimientos. Un reconocedor es una función que determina si una sentencia de entrada se encuentra acorde con la

gramática establecida. Asimismo, un analizador sintáctico es un reconocedor que produce análisis estructurales de acuerdo a la gramática. Un analizador robusto produce salidas, tanto como análisis parciales, incluso cuando la entrada no está totalmente cubierta por la gramática.

Podemos pensar que una gramática induce a un espacio de búsqueda que consiste en un conjunto de estados representando etapas de reescritura de reglas sucesivas, y un conjunto de transiciones entre esos estados. Cuando se analiza una sentencia, el analizador debe reescribir las reglas gramáticas en cierta secuencia. Una secuencia que conecta el estado S, la cadena que consiste de la categoría de inicio de la gramática, y el estado que consiste de la cadena exacta formada por las palabras de entrada, es llamada derivación. Cada estado en la secuencia consiste entonces de un estado en V y es llamada forma sentencial. Si tal secuencia existe, se dice que la sentencia es acorde a la gramática.

Los analizadores sintácticos pueden ser clasificados en diversas dimensiones, de acuerdo a la forma en que se hacen las derivaciones. Una de esas dimensiones se refiere a la regla de invocación. Estas pueden ser arriba-abajo, donde cada forma sentencial es producida por su predecesor mediante el reemplazo de un símbolo no terminal A por una cadena de símbolos terminales o no terminales  $X_1...X_d$ , donde  $A \rightarrow X_1...X_d$  es una regla gramatical. De forma inversa, en una derivación abajo-arriba, cada forma sentencial es producida al reemplazar  $X_1...X_d$  por A, dada la misma regla gramatical, y así sucesivamente aplicando reglas gramaticales en la dirección inversa.

Otra dimensión se refiere al modo en el cual el analizador trata con la ambigüedad, en particular, si el proceso es determinístico o no determinístico. Para el primer caso, sólo una e irrevocable opción puede ser tomada cuando el analizador se enfrenta a una ambigüedad local. Esta elección es típicamente basada en alguna forma de *mirar hacia adelante* o alguna preferencia sistemática.

Una tercera dimensión se refiere a si el analizador procede desde la derecha o desde la izquierda a través de la entrada o en algún otro orden.

### 2.3.3 Robusto

En un analizador sintáctico es natural esperar que las entradas corresponden con aquellas cadenas que están en el lenguaje formal  $L(G)$  generado mediante la gramática  $G$ . Sin embargo, un analizador de lenguaje natural siempre podrá estar expuesto a alguna cantidad de entradas que no estarán en  $L(G)$ . Una fuente de este problema es la subgeneración, la cual es causada por la falta de cobertura de  $G$  referente a el lenguaje natural  $L$ . Otro problema es que la entrada puede contener errores, es decir, que puede estar mal formada. Pero independientemente de por qué la entrada no está en  $L(G)$ , es usualmente deseable tratar de recuperar tanta información relevante como sea posible, en lugar de regresar ningún resultado en absoluto. En cierto sentido, el análisis robusto significa que desviaciones pequeñas de la entrada esperada sólo causarán pequeños inconvenientes en el resultado del análisis, mientras que desviaciones extremas pueden causar grandes inconvenientes.

Claramente, un sistema robusto requiere métodos que sacrifiquen algo de la tradicional idea de recuperar un análisis completo y exacto. Para evitar la situación donde el analizador puede solo detener y reportar la falla en el análisis de una entrada, una opción es aliviar algunas de las restricciones gramaticales hasta tal forma que una potencial sentencia no gramatical obtiene un análisis completo.

El problema clave de este enfoque es que, como el número de errores crece, el número de alternativas de alivio de restricciones que son compatibles con los análisis de toda la entrada pueden salirse de control, y que la búsqueda por la mejor solución sea, por lo tanto, muy difícil de controlar.

Uno puede entonces centrarse en el diseño de la gramática, haciéndola menos rígida con la esperanza que esto permitirá procesar con más tolerancia. La cantidad de información contenida en las representaciones estructurales producidas por el analizador es usualmente concerniente a la distinción entre análisis a fondo y análisis superficial. Un análisis profundo típicamente captura las dependencias de larga distancia. En contraste, un análisis superficial hace uso de una representación más esquelética. Esto se hace mediante primero asignar todos las posibles etiquetas

gramaticales y sintácticas a todas las palabras. Aplicando entonces reglas de reconocimiento de patrones para desambiguar las etiquetas, y como resultado disminuir el número de análisis. El resultado constituye una estructura de dependencia en el sentido que esto sólo provee relaciones entre palabras, y puede ser ambiguo debido a que las entidades de dependientes no son completamente especificadas.

Una segunda opción es sacrificar las entradas completas, es decir, la cobertura de la entrada, esto mediante pasar solamente los fragmentos que están bien formados de acuerdo a la gramática.

Una tercera opción es sacrificar la noción tradicional de un análisis constructivo, es decir, el análisis de sentencias mediante la construcción de representaciones sintácticas impuestas por reglas de una gramática. En lugar de eso, poder usar un análisis eliminativo, el cual consiste en inicializar primero un máximo de condiciones, y después reducir gradualmente los análisis que son ilegales de acuerdo a las restricciones dadas, hasta tener sólo análisis legales.

#### **2.3.4 Desambiguación**

Una observación básica es que, aunque una gramática general permitiría un gran número de análisis de casi cualquier sentencia no trivial, muchos de estos análisis serían extremadamente implausibles en el contexto de un dominio particular.

Basados en un corpus suficientemente extenso analizado por la gramática original, es entonces posible identificar combinaciones comunes de reglas de una gramática y guardarlas en un único conjunto de reglas. El resultado sería una gramática especializada, la cual, comparada con la gramática original, tendría un número más grande de reglas, pero una estructura más simple, reduciendo la ambigüedad y permitiendo un rápido procesamiento usando un análisis izquierda-derecha (LR, por sus siglas en inglés *Left-Right*).

Otra posibilidad es usar un método híbrido que le dé una jerarquía a un conjunto de análisis con base en su probabilidad en un determinado dominio, esto basándose en



datos de un conjunto de entrenamiento. En consecuencia, la desambiguación permite, naturalmente en un sentido u otro, la aplicación de inferencia estadística.

## **2.4 Análisis semántico**

Para textos numerosos, las aplicaciones específicas de NLP para análisis semántico pueden incluir recuperación de información, extracción de información, generación de resúmenes, minería de datos, y traducción automática, entre otras.

El análisis semántico es también muy relevante en el esfuerzo de generar ontologías Web y sistemas de representación del conocimiento. Asimismo, existe un gran valor en llevar a cabo análisis semánticos tan bien como sea posible, en el sentido en que reflejen la realidad cognitiva de los hablantes. Esto hace más fácil modelar la intuición de hablantes nativos y simular sus procesos de inferencia, sin embargo, existe incertidumbre en saber hasta dónde esto es posible.

En la lingüística en general, el análisis semántico se refiere a analizar el significado de las palabras, expresiones fijas, y el contexto de los enunciados. En la práctica, esto significa convertir las expresiones originales en algún tipo de metalenguaje semántico.

Muchos enfoques bajo la influencia de la lógica psicológica se han enfocado al significado condicional de la verdad, pero tales análisis son demasiados limitados para permitir la comprensión completa del uso del lenguaje ordinario o para solventar las diferentes aplicaciones prácticas requeridas, especialmente aquellas que involucran una interfaz humano-computadora o de razonamiento ingenuo de usuarios comunes.

Existe una división natural hecha entre semántica léxica, la cual trata con el significado de las palabras y de las combinaciones fijas de las mismas, y la semántica supraléxica, la cual tiene que ver con los significados del ampliamente indefinido número de combinaciones de palabras, como frases y sentencias, permitidas por la gramática.

Muchos lingüistas ahora hablan de lexicogramática, en lugar de mantener una distinción estricta entre lexicón y gramática. En parte, esto es debido a que es

evidente que la potencial combinatoria de palabras es fuertemente determinada por sus significados, en cierto sentido porque es claro que muchas construcciones gramaticales tienen significados específicos de construcción.

Un punto importante en la teoría semántica es referente a si es posible dibujar una línea bien marcada entre el contenido semántico, en el sentido del contenido codificado en la lexicogramática, y el conocimiento enciclopédico general.

En términos generales, la primera evidencia de la semántica lingüística viene de las interpretaciones de los hablantes nativos del uso de las expresiones lingüísticas en contexto, de la observación natural del lenguaje en uso, y de la distribución de expresiones lingüísticas; esto último se refiere a el patrón de uso, colocación, y frecuencia.

Un requerimiento frecuentemente identificado para el análisis semántico en el NLP se trata de la resolución de ambigüedad. Desde el punto de vista del procesamiento computacional de textos, muchos enunciados hechos por humanos están abiertos a múltiples interpretaciones, debido a que las palabras pueden tener más de un significado (ambigüedad léxica), o debido a que ciertas palabras, tal como los cuantificadores, adverbios de modo, u operadores negativos pueden darle un significado no determinista a una oración (ambigüedad de alcance), o debido a que la referencia pretendida de los pronombres puede no ser clara (ambigüedad de referencia).

En relación a las ambigüedades léxicas, es usual distinguir entre homonimia, la cual consiste en la relación de semejanza en la manera de escribirse o pronunciarse, es decir, que presentan dos palabras de significado diferente; y la polisemia, cuando una palabra tiene más de un significado.

Otros problemas para la semántica léxica se plantean por la existencia de expresiones figurativas o las unidades multipalabra; estas últimas son palabras cuyo significado no es predecible mediante el significado de las palabras individuales.

### 2.4.1 Teorías y enfoques de la representación semántica

Diferentes teorías y enfoques de la representación semántica pueden ser posicionados en dos categorías: formal vs. cognitiva y composicional vs. léxica.

Las teorías formales han sido fuertemente definidas desde finales de 1960, mientras que los enfoques cognitivos se han vuelto populares en las últimas tres décadas, movidas bajo la influencia de la ciencia cognitiva y la psicología.

La semántica composicional tiene que ver con la construcción, de lo menor a lo mayor, del significado, empezando con los componentes léxicos. Por otra parte, la semántica léxica tiene como objetivo analizar el significado de los componentes léxicos, por medio de analizar su estructura interna y contenido, o por medio de representar las relaciones con otros elementos en el lexicón.

### 2.4.2 Enfoques lógicos

Los enfoques lógicos del significado generalmente se enfocan en problemas composicionales, bajo la hipótesis que el significado de expresiones supraléxicas está determinado por el significado de sus componentes y la forma en que estas se combinan.

Diferentes sistemas lógicos han sido y están siendo desarrollados en el área de la semántica lingüística y el NLP. Uno de esos sistemas es *predicate logic*, en el cual las propiedades de conjunto de objetos pueden ser expresadas por medio de predicados, conectivas lógicas y cuantificadores. Esto, mediante proveer sintaxis especificando cómo los elementos del lenguaje lógico pueden ser combinados para formar expresiones lógicas correctas, y mediante proveer semántica especificando que significan tales expresiones dentro del sistema lógico.

Aunque algunas veces se presentó como una teoría general de conocimiento, *predicate logic* no es suficientemente robusto como para representar las complejidades del significado semántico y es fundamentalmente diferente del razonamiento humano.

### **2.4.3 Teoría de representación de discurso**

La teoría de representación de discurso fue desarrollada a principios de los 80s por Kamp [23] para la captura de la semántica de los discursos o textos, es decir, las secuencias coherentes de las sentencias o enunciados. La idea básica es que a medida que se desarrolla un discurso o texto el oyente construye una representación mental, representada por la llamada estructura de representación del discurso, y que por cada sentencia entrante solicita adiciones a esta representación. De esta manera es un enfoque dinámico para la semántica del lenguaje natural.

La teoría de representación de discurso requiere los siguientes componentes: una definición formal de la representación del lenguaje que consiste de una definición recursiva del conjunto de todas las estructuras de representación del discurso bien formadas, y una semántica modelo-teórica para los miembros de este conjunto; un procedimiento de construcción que especifique cómo una estructura de representación del discurso puede ser agregada cuando llega a estar disponible nueva información.

El enfoque de la teoría de representación de discurso está bien adaptado para tratar, por ejemplo, con resolución de anáfora.

#### **2.4.3.1 Lexicón generativo de Pustejovsky**

Otro enfoque dinámico de la semántica, pero que se enfoca en elementos léxicos, es la teoría del lexicón generativo de Pustejovsky [35].

Pustejovsky postula que, dentro de contextos particulares, los componentes léxicos asumen diferentes sentidos. Él desarrolló la idea de un lexicón en el cual los sentidos de las palabras en el contexto pueden ser flexiblemente derivados con base en una representación multinivel que involucra al menos los siguientes niveles:

1. Estructura del argumento: Especificación del número y tipo de los argumentos lógicos y cómo ellos se desenvuelven sintácticamente.
2. Estructura de evento: Definición del tipo de evento de un elemento léxico y una frase. El tipo de evento incluye estados, procesos y transiciones; la estructuración de subeventos es posible.

3. Estructura de cualidades subjetivas: Son modos de explicación, que comprenden las cualidades subjetivas en cuatro tipos: constitutiva, ¿de qué está hecho el objeto?; formal, ¿qué es el objeto? -cómo se distingue en un amplio dominio; de fin definido, ¿cuál es el propósito o función del objeto?; y de naturaleza, ¿cómo surgió el objeto? -factores que influyeron en su surgimiento.
4. Estructura de herencia léxica: Identificación de cómo una estructura léxica está relacionada con otras estructuras en el lexicón y su contribución con la organización global del lexicón.

Los procesos generativos del sistema de Pustejovsky interconectan los niveles de representación con el objetivo de proveer una interpretación composicional de los elementos léxicos en el contexto.

El lexicón generativo tiene diferentes propósitos en los enfoques lógico-inspirados para la semántica. Se orienta a un enfoque léxico descomposicional detallado para la semántica lingüística, y al mismo tiempo provee herramientas que permiten conocer el significado en el contexto mediante un cálculo composicional.

#### **2.4.3.2 Metalenguaje semántico natural**

El metalenguaje semántico natural es un sistema descomposicional basado en primas semánticas establecidas empíricamente, es decir, significados simples e indefinibles que parecen estar presentes como significados de palabra en todos los lenguajes.

El sistema de metalenguaje semántico natural usa un metalenguaje, el cual es esencialmente un subconjunto estandarizado del lenguaje natural. Este consiste en un subconjunto que contiene palabras con su significado, junto con un subconjunto con sus propiedades sintácticas asociadas.

El modo formal de la representación del significado en el enfoque de metalenguaje semántico natural es la explicación semántica. Esto es una paráfrasis reductiva, es decir, un intento de expresar en otras palabras lo que el hablante está diciendo cuando pronuncia la expresión que es explicada. Originado con Wierzbicka (1972), el

metalenguaje semántico natural ha sido desarrollado y refinado por 35 años; sin embargo, aunque este enfoque es posiblemente la mejor teoría de semántica léxica en la escena contemporánea, tiene pocas aplicaciones para el NLP.

Los investigadores de este enfoque reconocen que, para muchas palabras en un vocabulario concreto, no es posible producir explicaciones plausibles directamente en términos de primas semánticas únicas. En vez de eso, las explicaciones típicamente requieren una combinación de primas semánticas y de ciertos significados léxicos conocidos en la teoría del metalenguaje semántico natural, como moléculas semánticas. Las moléculas semánticas funcionan como bloques de construcción para lograr conceptos más complejos. Estas moléculas también se pueden anidar, una dentro de otra, creando cadenas de dependencia semántica.

#### **2.4.3.3 Semántica orientada a objetos**

La semántica orientada a objetos es relativamente un nuevo campo en la semántica lingüística. Aunque es bastante restringida en los dominios de aplicaciones semánticas, esta es principalmente aplicada a la representación de significado verbal.

La motivación básica detrás del esparcimiento del paradigma computacional orientado a objetos en la semántica lingüística es su accesibilidad intuitiva. El sistema cognitivo humano se basa en entidades y en qué son ellas, cómo se relacionan con otras entidades, qué les pasa y qué hacen, y cómo ellas interactúan con otras. Esto corresponde al enfoque orientado a objetos, en el cual el concepto de objeto es central, cuyas características, relaciones con otras entidades, comportamiento, e interacciones son modeladas de forma rigurosa.

#### **2.4.3.4 Hechos relacionales y ontologías**

Las relaciones semánticas entre elementos léxicos (relaciones de sentido) generan la base para las redes de palabras, tal como la base de datos digital llamada WordNet (Fellbaum 1998) y de enfoques similares para otros lenguajes diferentes al inglés, por ejemplo, la base de datos multilenguaje EuroWordNet (Vossen 1998, 2001). El

uso de ontologías y la estructura lingüística se está convirtiendo en un área sobresaliente en la representación de conocimiento y el NLP.

Las relaciones de sentido pueden ser vistas como un revelador de la estructura semántica de un lexicón. Existen relaciones de sentido horizontales y verticales. Las relaciones horizontales incluyen la sinonimia, es decir, palabras con el mismo significado, pero con diferente forma lingüística. También existen dos relaciones verticales principales que incluyen la hiponimia y la meronimia. La hiponimia ocurre cuando el significado de un elemento léxico, el hipónimo, es más específico que el significado del otro, el hiperónimo. Por otra parte, la meronimia ocurre cuando el significado de un elemento léxico especifica que forma parte de otro elemento léxico, por ejemplo, dedo es merónimo de mano.

Las ontologías han sido desarrolladas y empleadas en el área de inteligencia artificial y representación de conocimiento, y más generalmente en las ciencias de la computación.

## Capítulo III. Estado del arte

### *Contenido*

En este capítulo se presentan algunas investigaciones que han propuesto nuevos enfoques para la detección de texto engañoso mediante aprendizaje automático.



### 3.1 Introducción

La detección de texto engañoso mediante algoritmos de aprendizaje automático trata de identificar las características que definen a estos textos de acuerdo a un conjunto de datos previo llamado conjunto de entrenamiento. De esta manera, los algoritmos pueden analizar un nuevo texto (no contenido en el conjunto de entrenamiento) e identificar si pertenece a alguna clase específica.

Los investigadores han estudiado la detección de engaño mediante el uso de diferentes fuentes de características. Algunos de estos métodos están basados en herramientas de bolsa de palabras (BoW), en las cuales no se considera el orden de las palabras; otros añaden información sintáctica como características. En algunos casos, se buscan algunas señales generales de engaño [6], tales como el uso de palabras únicas, autoreferencias o modificadores, entre otros.

En general, los enfoques de estilo lingüístico (como ns-gramas) analizan las relaciones entre las palabras; en cambio, los enfoques de BoW (LDA, MTD, LIWC, n-gramas) ignoran la gramática e incluso el orden de las palabras, pero siguen contando el número de instancias de cada palabra.

Para detectar textos engañosos, una técnica comúnmente aplicada es usar n-gramas. Este método puede extraer características de un texto basado en diferentes elementos; por ejemplo, palabras, sílabas, fonemas, letras, etc.

En el trabajo de Donato et al. [13] la palabra n-gramas se comparó con la letra n-gramas. Estos últimos han demostrado tener un mejor rendimiento en el conjunto de datos OpSpam. Aunque los n-gramas logran resultados aceptables por sí mismos, generalmente se complementan con otras técnicas de PLN debido al hecho de que la combinación de características ha demostrado mejorar los resultados.

Otro enfoque de BoW consiste en utilizar el diccionario linguistic inquiry and word count (LIWC), que incluye una clasificación de palabras y una herramienta de recuento. Newman et al. [27], por ejemplo, al analizar las categorías de palabras de LIWC, encontraron que los mentirosos usan menos autoreferencias y usan palabras

de emociones negativas. Este trabajo sentó las bases para que la herramienta LIWC sea ampliamente utilizada por otros investigadores [38], [40].

Hauch et al. [19] introdujo un meta-análisis de varios trabajos de investigación de identificación de textos engañosos. Este meta-análisis se centró en categorías lingüísticas específicas, por ejemplo, las contenidas en LIWC. Los hallazgos de la investigación sugieren que los mentirosos usan ciertas categorías lingüísticas en una tasa diferente que los que dicen la verdad.

La detección de engaño se ha aplicado en diversas situaciones particulares. En Williams et al. [42], se hizo una comparación entre las mentiras dichas por niños y las mentiras dichas por adultos. La investigación se llevó a cabo con el objetivo de detectar el engaño en los tribunales donde los niños testifican. Los autores eligieron 48 niños y 28 adultos para generar un conjunto de datos; la mitad de los niños y adultos contaron mentiras y la mitad de ellos dijeron la verdad. De esta forma, se utilizó la herramienta LIWC para generar las muestras para la clasificación. Los resultados mostraron que existen diferencias significativas entre los textos verdaderos y los falsos, principalmente en variables lingüísticas tales como las autoreferencias singulares (por ejemplo, I, my, me), autoreferencias (por ejemplo, we, our, us) y las emociones negativas. Además, los resultados de la investigación mostraron que las variables lingüísticas se encontraron en distintas proporciones dependiendo de si la mentira fue contada por un niño o por un adulto.

Los estudios desarrollados mediante el uso de herramientas de BoW han tenido éxito; sin embargo, en un esfuerzo por mejorar los resultados, se ha tenido en cuenta el contexto de las oraciones, al analizar las relaciones sintácticas de las palabras con el uso de árboles de dependencia [43]. En general, el uso de las relaciones sintácticas no ha mostrado un rendimiento sobresaliente en la tarea de clasificar texto engañoso. Aunque complementar este método con un enfoque de BoW puede mejorar los resultados.

En el estudio de Mihalcea y Pérez-Rosas [33], las características se generaron utilizando diferentes enfoques (por ejemplo, etiquetas gramaticales (PoS), gramáticas

libres del contexto (CFG), unigramas, LIWC y combinaciones de estos). Los autores predijeron, con una exactitud de entre 60% y 70%, si una persona de sexo femenino o masculino había escrito un texto engañoso. En los resultados mostrados, el uso de PoS y CFG no mostró una mejora significativa en la exactitud con respecto a los unigramas y LIWC. Esto sugiere que los enfoques de BoW tienen un desempeño similar a los enfoques de estilo lingüístico.

Para probar la eficiencia de los algoritmos de aprendizaje se recurre a conjuntos de datos llamados corpus (singular) o corpora (plural). Los conjuntos de datos se pueden dividir en dos subconjuntos: el ya mencionado conjunto de entrenamiento que le da el conocimiento previo al clasificador; y el conjunto de prueba, el cual consiste en documentos que se clasificaran con base en el conocimiento previo.

### **3.2 El detector de engaño, explorando en el reconocimiento automático del texto engañoso**

En el trabajo de investigación de Mihalcea y Strapparava [26] se abordó la detección de texto engañoso y texto veraz mediante el enfoque de aprendizaje automático. De esta manera, se pretendían capturar características subyacentes, de forma automática en el texto que permitieran identificar si un texto es engañoso o veraz. Los autores se basaron solamente en la obtención de características lingüísticas, aunque otros trabajos incluyen características no lingüísticas (por ejemplo, características acústicas), esto debido a que la mayor parte de información que se encuentra en la web son textos.

El objetivo de la investigación fue saber si los textos engañosos y veraces realmente eran separables, es decir, si existen características que solo se pueden encontrar o que se encuentran con mayor probabilidad en los textos engañosos. Para resolver lo anterior, los autores crearon un conjunto de datos de tópicos controversiales; este conjunto contiene textos cortos engañosos y veraces.

De forma específica el conjunto de datos que generaron consta de tres tópicos controversiales: aborto, pena de muerte y mejores amigos. Como en todo caso, y

para poder evaluar el enfoque propuesto, fue necesario que el conjunto de datos fuera etiquetado, es decir, que cada texto tenga su respectivo identificador para saber si es engañoso o veraz. Para realizar la labor de etiquetado optaron por usar el servicio del turco mecánico de Amazon, el cual es una herramienta web que permite a diferentes usuarios del turco mecánico ayudar en una tarea a cambio de una aportación monetaria. El conjunto se formó mediante dos dinámicas. Primero, en el caso de los tópicos acerca de aborto y pena de muerte la tarea consistió en que los usuarios debieron escribir su opinión real acerca de cada uno de estos dos tópicos; en seguida se les pidió escribir la opinión contraria a su opinión real. De esta forma, el resultado es un texto con información veraz y un texto con información engañosa. En segundo lugar, para el caso del tópico referente a los mejores amigos, se les pidió a los usuarios que pensarán en su mejor amigo y describieran las razones de su amistad; después se les pidió que pensarán en una persona a la que no pudieran soportar y la describieran como si fuera su mejor amigo. Como resultado, en el primer caso los participantes mienten acerca de sus creencias, mientras que en el segundo caso los participantes mienten acerca de sus sentimientos hacia otra persona. Finalmente, fueron recolectados 100 textos veraces y 100 engañosos por tópico, sumando un total de 600 textos.

Con el propósito de obtener las características más relevantes del texto engañoso y del texto veraz idearon un método para valorar en qué medida las palabras se inclinan hacia una u otra clase. Tal método se basa en las siguientes medidas: la cobertura del texto engañoso (fórmula ( 3.1), la cual obtiene la fracción del número de *tokens* de la palabra  $W_i$  en el corpus de textos engañosos (D) entre el total de *tokens* del corpus D. De la misma forma, la cobertura del texto veraz (fórmula ( 3.2 )) realiza el mismo cálculo, pero esta vez para el corpus de texto veraz (T).

$$Cobertura_D(C) = \frac{\sum_{W_i \in C} Frecuencia(W_i)}{Tamaño_T} \quad (3.1)$$

$$Cobertura_T(C) = \frac{\sum_{W_i \in C} Frecuencia(W_i)}{Tamaño_T} \quad (3.2)$$

$$Dominancia_D C = \frac{Cobertura_D(C)}{Cobertura_T(C)} \quad (3.3)$$

**Tabla 1. Clases de LIWC dominantes del texto engañoso y veraz**

Clase	Puntaje	Ejemplo de palabras
Texto engañoso		
Metáfora	1.71	god, die, sacred, mercy, sin, dead,hell, soul, lord.
Tú/Ustedes	1.53	you, thou
Otros	1.47	she, her, they, his, them, him, herself, himself, themselves
Humanos	1.31	person, child, human, baby, man, girl,humans, individual, male, person, adult
Certero	1.24	always, all, very, truly, completely, totally
Texto veraz		
Óptimo	0.57	best, ready, hope, accepts, accept, determined, accepted, won, super
Yo	0.59	I, myself, mine
Amigos	0.63	friend, companion, body
Sí mismo	0.64	Our, myself, mine, ours
Visión	0.65	Believe, think, know, see, understand, found, thought, feels, admit

Finalmente, se calcula el valor de dominancia (fórmula ( 3.3 ) de la clase C (clases incluidas en linguistic inquiry and word count, en su versión 2001), el cual es la proporción entre la cobertura del texto veraz y la cobertura del texto engañoso, calculado para la misma clase. Los tres tópicos fueron mezclados para formar dos conjuntos de datos generales: el conjunto T y el conjunto D, de texto veraz y engañoso, respectivamente. Sobre estos conjuntos generales fueron calculados los valores de dominancia de cada clase C. De esta manera, un resultado cercano a 1 indicará que existe una distribución similar de palabras en la clase C tanto de texto veraz como de texto engañoso. Por otra parte, si el resultado menor a 1, significará

que la clase es predominante del texto veraz, en cambio, si el resultado es superior a 1, significará que la clase es predominante del texto engañoso. En la Tabla 1 se muestran algunos valores de clases dominantes obtenidos por los autores.

En los experimentos se clasificaron los textos mediante un clasificador naïve Bayes y una máquina de vectores de soporte (SVM, por sus siglas en inglés *Support Vector Machine*). Primero, se clasificaron los tres tópicos por separado; los resultados son mostrados en la Tabla 2, en la cual se especifican los valores de exactitud obtenidos por cada clasificador; asimismo, la exactitud promedio de los tres tópicos. Se observa también que el mejor resultado en promedio fue obtenido por el clasificador naïve Bayes.

**Tabla 2. Exactitud obtenida por tópico usando una validación cruzada de 10 pliegues**

Tópico	Naïve Bayes	SVM
Aborto	70.0%	67.5%
Pena de muerte	67.4%	65.9%
Mejor amigo	75.0%	77.0%
Promedio	70.8%	70.1%

Siguiendo el trabajo de Mihalcea y Strapparava, también se realizó un estudio para detectar texto engañoso escrito en español: Almela et al. [1] recopiló un nuevo conjunto de datos con temas que abarcan la adopción homosexual, opiniones sobre la tauromaquia y sentimientos acerca de un mejor amigo. Cien documentos falsos y cien verdaderos fueron recogidos para cada tema con un promedio de 80 palabras por documento. Distintas dimensiones de LIWC se utilizaron para lograr una clasificación más precisa por medio de una máquina de vectores de soporte (SVM).

### 3.3 Identificación de engaño por medio de algún estiramiento de la imaginación

El objetivo de la investigación de Ott et al. [29] fue identificar las opiniones engañosas que se definen como “opiniones ficticias que han sido deliberadamente escritas para sonar auténticas”.

Para lograr el objetivo los autores exploran tres enfoques, el primero es ver el problema como una tarea de categorización de texto mediante el uso de n-gramas y clasificadores, esto con el fin de etiquetar los documentos como veraces o engañosos. El segundo fue optar por un enfoque psicolingüista, en el cual se pretendía verificar cuales son los estados psicológicos, como las emociones negativas, relacionados con el engaño. Por último, en el tercero se abordó la tarea como un problema de género mediante el cual se percibe la escritura de textos engañosos como de tipo imaginativa y la escritura de textos veraces como de tipo informativa.

**Tabla 3. Exactitud, precisión (P), cobertura (R), y medida F (F) de la identificación de texto engañoso realizada por humanos**

Participante	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
Juez 1	61.9%	57.9	87.5	69.7	74.4	36.3	48.7
Juez 2	56.9%	53.9	95.0	68.8	78.9	18.8	30.3
Juez 3	53.1%	52.3	70.0	59.9	54.7	36.3	43.6

Mediante la clasificación de un conjunto de 400 textos engañosos y 400 textos veraces los autores identificaron que los clasificadores de aprendizaje automático, entrenados con características obtenidas mediante enfoques psicolingüistas y de identificación de género, sobrepasan estadísticamente el nivel de los enfoques de categorización de texto, tal como los n-gramas. Además, remarcaron que las capacidades humanas para diferenciar entre un texto engañoso y uno veraz, como se muestra en la Tabla 3, son limitadas. Esto lo ratificaron mediante una clasificación,

hecha por tres participantes, de un subconjunto de 160 documentos. En la tabla también se muestran medidas menos simplistas que la exactitud, éstas son: precisión (P), exhaustividad (R) y medida F (F).

**Tabla 4. Resultados de la aplicación de diferentes fuentes de generación de características y aplicación de dos clasificadores.**

Características	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
Identificación de género							
POS/SVM	73.0%	75.3	68.5	71.7	71.1	77.5	74.2
Enfoque psicolingüístico							
LIWC/SVM	75.8%	77.2	76.0	76.6	76.4	77.5	76.9
Categorización de texto							
Unigramas/SVM	88.4%	89.9	86.5	88.2	87.0	90.3	88.6
Bigramas/SVM	89.6%	90.1	89.0	89.6	89.1	90.3	89.7
LIWC+bigramas/SVM	89.8%	89.8	89.8	89.8	89.8	89.8	89.8
Trigramas/SVM	89.0%	89.0	89.0	89.0	89.0	89.0	89.0
Unigramas/NB	88.4%	92.5	83.5	87.8	85.0	93.3	88.9
Bigramas/NB	88.9%	89.8	87.8	88.7	88.0	90.0	89.0
Trigramas/NB	87.6%	87.7	87.5	87.6	87.5	87.8	87.6

Los autores también dan importancia a que la identificación de texto tiene que ser afrontada mediante el contexto y la motivación que condujo al engaño, y no solamente mediante la identificación de un conjunto de señales universales de engaño. Por esta razón, proponen generar características basadas en etiquetas gramaticales (POS tags, por sus siglas en inglés part-of-speech tags); esto, con base en que algunas investigaciones [2], [37] han probado que la frecuencia y distribución de las etiquetas gramaticales tienen cierta relación con la identificación de género. De esta manera, los autores construyen vectores de características basadas en la frecuencia de cada etiqueta gramatical.



Por otra parte, y para comparar los resultados del enfoque que propusieron, usaron LIWC en su versión 2007, generando una característica por cada una de las 80 dimensiones incluidas en este diccionario.

Finalmente, con referencia a las fuentes de características, usaron n-gramas debido a que mediante ellos se puede capturar el contenido y contexto de los documentos. Asimismo, generaron características a partir de unigramas, bigramas y trigramas.

Para realizar la clasificación de documentos usaron un clasificador naïve Bayes y una SVM con una validación cruzada de cinco pliegues. Los resultados se muestran en la Tabla 4, en la cual podemos observar que la combinación de características de LIWC y bigramas, clasificados con una SVM, obtuvieron el mejor resultado. Además, con respecto a exactitud y medida F, todas las características generadas y clasificadas con aprendizaje automático superan la clasificación humana (ver Tabla 3).

Aunque los autores proponen la identificación de texto mediante un enfoque de identificación de género, se puede observar en la Tabla 4 que la generación de características mediante la frecuencia de las etiquetas gramaticales no fue la mejor opción para producir vectores suficientemente informativos. Y puesto que los autores argumentan que en los experimentos se probaron todas las combinaciones de fuentes de características, y dado que no se muestran más combinaciones del enfoque POS, se puede inferir que combinar el enfoque de etiquetas gramaticales con otro enfoque resultó perjudicial para la clasificación.

### **3.4 El uso de diferentes conjuntos de sujetos homogéneos en el análisis de texto engañoso**

En la investigación de Fornaciari y Poesio [10] se aborda la identificación de texto engañoso mediante la creación de subconjuntos a partir de un conjunto general o corpus. La agrupación de subconjuntos homogéneos de sujetos se da, por ejemplo, con la separación de género, es decir, si el sujeto que escribió el texto es hombre o mujer. La agrupación de conjuntos fue realizada tanto de forma automática, mediante

técnicas de agrupamiento automático, y también mediante el uso de los metadatos contenidos en los conjuntos de datos. De esta manera, las agrupaciones, como argumentan los autores, podrían generar la eficiencia en la detección de texto engañoso, basados en que el comportamiento similar en la escritura de los autores podría influir en la detección del engaño.

En este estudio se generó un conjunto de datos llamado DeCour (Deception in Court), el cual es un conjunto de datos de transcripciones de treinta y cinco audiencias llevadas a cabo en cuatro cortes de Italia. Los textos se componen de preguntas hechas en la corte a los testigos que defienden al acusado, como resultado el conjunto de datos contiene etiquetas que informan si la oración o respuesta transcrita es veraz o engañosa.

Los autores crearon vectores de características con tres tipos de características: la primera agrega información básica como el tamaño de las sentencias y el número de palabras que tienen más de seis caracteres; la segunda agrega información léxica mediante el uso de LIWC 2001, y agregando al vector de características las 80 dimensiones con las que cuenta dicha versión; la última agrega frecuencias de n-gramas de lemas y de etiquetas gramaticales, los autores tomaron en cuenta, en los experimentos, desde unigramas hasta pentagramas.

Mediante la obtención de n-gramas, con valores de  $n$  desde 1 hasta 5, se generaron dos listas de los n-gramas sobresalientes, una correspondiente al texto engañoso y otra al texto veraz. Dichas listas se compararon para evitar que una característica igual se presente en ambas, de esta manera generar solamente información particular que permita una clasificación eficiente.

Tres experimentos fueron realizados usando el conjunto de datos creado. El primer experimento consistió en usar todo el conjunto de datos por medio de establecer un conjunto de entrenamiento y de prueba.

En el segundo experimento se generaron subconjuntos del corpus completo, de forma que se eliminaron aquellos vectores que estuvieran demasiado alejados de los

grupos (*outliers*). Estos grupos fueron establecidos mediante un método no supervisado de agrupamiento de patrones.

En el tercer experimento se hizo uso de los metadatos que se recabaron con el corpus. En principio el corpus cuenta con la siguiente información: el género (masculino o femenino), el lugar de nacimiento y la edad al momento de la declaración. Finalmente, los autores decidieron tomar en cuenta sólo el género de los sujetos, en específico, sólo aquellos participantes de género masculino. Esta restricción impactó fuertemente sobre el conjunto de entrenamiento y no a tal grado el conjunto de prueba.

El conjunto de prueba consta de 426 oraciones de las cuales 190 son engañosas y 236 veraces. Los resultados de la clasificación en los tres diferentes experimentos se muestra en la Tabla 5. Se puede observar que el peor resultado, respecto al promedio de medida F, se obtiene cuando sólo se contempla el género masculino, esto podría deberse a que el conjunto de entrenamiento se redujo drásticamente. Por otra parte, el mejor resultado se obtuvo en el experimento donde se eliminaron los *outliers*, lo que pudo permitir un grupo más homogéneo.

**Tabla 5. Resultados de la clasificación de corpus DeCour. Se muestran: precisión (P), exhaustividad (R) y medida F (F), asimismo el promedio de medidas F.**

Experimento	Promedio F	Veraz			Engañoso		
		P	R	F	P	R	F
Todo el corpus	60.1	62.9	94.0	75.3	80.8	31.0	44.9
Sin <i>outliers</i>	64.0	66.8	93.8	77.9	81.0	36.2	50.0
Género masculino	59.5	67.8	94.2	78.9	74.4	27.4	40.0

### 3.5 Detección intercultural de engaño

En la sección 3.4 se consideró que la creación de subconjuntos basados en el género y en los *outliers* tendría cierto impacto en la eficiencia de la clasificación. Asimismo, en la investigación de Pérez-Rosas y Mihalcea [32] se realizan

experimentos con textos escritos por personas de diferentes países para verificar hasta qué punto es posible identificar el engaño, por ejemplo, usando los textos de un país para identificar engaño en los textos de otro país diferente.

La mayoría de estudios de detección de engaño están enfocados a documentos de determinado país e idioma, principalmente al idioma inglés. En este sentido los autores argumentan que los textos, principalmente aquellos que se encuentran en los sitios web, tienden, en cierta proporción, a ser escritos por personas que tienen diferente cultura o idioma; esto implica que las personas pueden tener diferentes creencias y valores morales. En consecuencia, se pone en tela de juicio si los enfoques que se basan en un solo idioma y cultura podrán ser igualmente aplicados a una mezcla de textos interculturales e incluso de diferente idioma (este último, usando herramientas de traducción).

En este estudio se abordó la detección de engaño con textos escritos en tres diferentes culturas: Estados Unidos de América, México, y la India. Por lo cual, los autores generaron tres conjuntos de datos que abordan, para cada uno, tres tópicos controversiales: aborto, pena de muerte y mejores amigos. Lo anterior tal y como se manejó en el estudio presentado en la sección 3.2, excepto para los textos recabados en español de México; estos últimos se obtuvieron mediante una página creada por los autores debido a la poca participación recibida mediante el turco mecánico de amazon.

Las fuentes de generación de características para este estudio fueron dos. Como en la mayoría de las investigaciones mostradas en esta sección (estado del arte), los unigramas son requeridos para generar una parte del vector de características, por lo que se puede inferir que las palabras por sí solas son una importante fuente informativa para detectar el engaño. Para complementar el vector de características fue usada una segunda herramienta que se ha vuelto muy popular en la detección de texto engañoso, esta es LIWC, que como se ha dicho antes es un enfoque que brinda un análisis psicolingüístico.

Fueron realizados dos tipos de experimentos: el primero es una clasificación sobre los conjuntos de datos con documentos del mismo país; el segundo se trata de un cruce de dominios, es decir, tomar dos conjuntos (de los tres disponibles) para formar el conjunto de entrenamiento y usar el conjunto restante para generar el conjunto de prueba. Este último experimento se hizo tanto de tópico mezclado, donde se aplica el dominio cruzado sobre los tres tópicos sobre cada cultura por separado; como para cultura cruzada, donde se aplica dominio cruzado sobre las tres culturas. Este último experimento, con el objetivo de probar si las características que sirven para identificar texto engañoso en un determinado lenguaje de un país, son equivalentes o útiles al buscar engaño en el lenguaje manejado por otro país diferente (de diferente cultura).

**Tabla 6. Experimentos realizados sobre los corpora de cada país, por separado**

Tópico	LIWC	Unigramas	
		En dominio	Dominio cruzado
Inglés: EUA			
Aborto	73.0	63.8	80.4
Mejor amigo	73.0	74.5	60.8
Pena de muerte	58.1	58.1	77.2
Inglés: India			
Aborto	56.0	46.0	50.0
Mejor amigo	71.4	60.5	57.2
Pena de muerte	63.5	57.5	54.0
Español: México			
Aborto	62.2	52.5	57.7
Mejor amigo	75.3	66.7	50.5
Pena de muerte	62.2	54.9	63.4

En la Tabla 6 se muestran los resultados de la clasificación, de forma individual, de cada cultura mediante la generación de características con el enfoque psicolingüístico y mediante el uso de unigramas. Sobre este último se aplica la clasificación en dominio, en la cual se clasifica cada tópico por separado (aborto,

mejor amigo, y pena de muerte) y la clasificación de dominio cruzado, en la cual se usan dos tópicos como entrenamiento y un tercer tópico como prueba.

En la Tabla 7 se muestran los resultados de la clasificación de dominio cruzado mediante el uso de dos lenguajes como conjunto de entrenamiento y el lenguaje restante como conjunto de prueba.

**Tabla 7. Clasificación de cultura cruzada con LIWC y unigramas como fuente de características**

Tópico	LIWC	Unigramas
Entrenamiento: Inglés: EUA, prueba: Inglés: India		
Aborto	52.3	57.9
Mejor amigo	59.5	51.0
Pena de muerte	53.5	59.0
Entrenamiento: Inglés: India, prueba: Inglés: EUA		
Aborto	62.5	55.5
Mejor amigo	55.8	53.2
Pena de muerte	39.2	50.7
Entrenamiento: Inglés: EUA, prueba: español: México		
Aborto	53.9	61.5
Mejor amigo	67.7	65.0
Pena de muerte	62.2	59.8
Entrenamiento: Inglés: India, prueba: español: México		
Aborto	43.6	55.1
Mejor amigo	60.8	67.2
Pena de muerte	59.8	51.2

En la Tabla 8 se muestran las clases predominantes tanto del texto engañoso como del texto veraz. Además, algunos ejemplos de palabras incluidas en cada clase. Similar al trabajo mostrado en la Sección 3.2, un puntaje significativamente menor a uno indica que la clase es predominante del texto veraz, mientras que un valor significativamente por encima de uno indica que la clase es predominante del texto engañoso.

**Tabla 8. Clases de LIWC predominantes en el texto engañoso y en el texto veraz**

Clase	Puntaje	Ejemplo	Clase	Puntaje	Ejemplo
Inglés: EUA					
Engañoso			Veraz		
Metáfora	1.77	Die, died, hell, sin, lord	Visión	0.68	Accept, believe, understand
Otros	1.46	He, her, herself, him	Yo	0.66	I, me, my, myself
Tú/Ustedes	1.41	Thou, you	Optimismo	0.65	Accept, hope, top, best
Otras referencias	1.18	He, her, herself, him	Nosotros	0.55	Our, ourselves, us, we
Emociones negativas	1.18	Afraid, agony, awful,bad	Amigos	0.46	Buddies, friend
Inglés: India					
Engañoso			Veraz		
Negación	1.49	Cannot, neither, no, none	Pasado	0.78	Happened, helped, liked, listened
Físico	1.46	Heart, ill, love, loved	Yo	0.66	I, me, mine, my
Futuro	1.42	Be, may, might, will	Optimismo	0.65	Accept, accepts, best, bold
Otros	1.17	He, she, himself, herself	Nosotros	0.55	Our, ourselves, us, we
Humanos	1.08	Adult, baby, children, human	Amigos	0.46	Buddies, companion, friend, pal
Español: México					
Engañoso			Veraz		
Certeza	1.47	Jamás(never), siempre(always)	Optimismo	0.66	Aceptar(accept), animar(cheer)
Humanos	1.28	Bebé(baby), persona(person)	Sí mismo	0.65	Conmigo(me), tengo(have), soy(am)
Tú/Ustedes	1.26	Eres(are), estás(be), su(his/her)	Nosotros	0.58	Estamos(are), somos(be), tenemos(have)
Negación	1.25	Jamás(never), tampoco(neither)	Amigos	0.37	Amigo(friend), amistad(friendship)
Otros	1.22	Es(is), está(are), otro(other)	Pasado	0.32	Compartimos(share), vivimos(lived)

### 3.6 Identificando opiniones engañosas mediante el método de aprendizaje de multitudes

Las opiniones engañosas en la venta de libros, de acuerdo al estudio de Fornaciari y Poesio [12], ocurren “cuando los autores escriben una opinión brillante sobre sus propios libros”. Con el propósito de analizar las opiniones engañosas presentes en la

venta de libros a través de amazon, los autores crearon un nuevo conjunto de datos que etiquetaron con base en algunas señales de engaño, la presencia de dichas señales determinó si la opinión fue considerada engañosa o veraz. El enfoque usado para el etiquetado fue el aprendizaje de multitudes propuesto por Raykar [36]; además fue evaluada la efectividad de diferentes métodos de etiquetado de acuerdo con el rendimiento de los modelos generados para detectar opiniones engañosas.

Los autores argumentan que la gran desventaja de otros trabajos que tratan la detección de engaño es que son carentes de un conjunto de datos con textos reales, en cambio son recreados artificialmente; por ejemplo, cuando se les pide a los participantes mentir acerca de una opinión real. El punto es que, al darle permiso a una persona de mentir, es posible que las características de engaño no se reflejen en el texto, o se reflejen de forma parcial; inclusive cabe la posibilidad de que lo que se esté capturando no sea el mismo fenómeno, por lo tanto, la identificación y uso del modelo no será congruente con los textos engañosos en fenómenos reales.

Por lo anterior, los autores propusieron un nuevo método para identificar opiniones engañosas en amazon. Este sistema se basa en dos procesos que se definen a continuación.

En el primero se colaboró con un experto en detección de opiniones engañosas para identificar una serie de criterios para encontrar opiniones no genuinas. De algunas de estas opiniones ya se tenía, de antemano, la seguridad de que eran engañosas debido a que se obtuvieron confesiones de personas dedicadas a la escritura de las mismas.

En el segundo, los autores desarrollaron un enfoque que identificó la veracidad de las opiniones usando indicadores potenciales de veracidad. Finalmente, usaron el algoritmo propuesto por Raykar para asignar cada opinión del corpus a una clase.

Después de crear el corpus, los autores identificaron un conjunto de señales, las cuales su presencia sugería que los textos eran engañosos. Las señales identificadas fueron las siguientes: (1) libros sospechosos, de acuerdo a la investigación, son libros que seguramente o muy probablemente recibieron opiniones



engañosas; (2) el tiempo, Sandra Parker, una escritora de opiniones engañosas relató que las agencias para las que trabajaba le daban máximo 48 horas para escribir las opiniones, por lo cual sugirió que se pusiera especial atención en libros que recibieran opiniones en cortos periodos de tiempo; (3) registro, amazon permite crear cuentas de usuario usando los nombre reales de los usuarios, por tanto es menos probable que usaran su nombre real escribieran opiniones engañosas; (4) libro comprado, otra información que se puede obtener de amazon es si el libro fue comprado o no mediante dicho sitio, por lo tanto la ausencia de compra del libro fue considerada una señal de engaño.

Fueron llevados a cabo dos experimentos. En el primero la asignación de clases se realizó con base en una votación de las señales de engaño. Aquellas opiniones con hasta 2 señales de engaño fueron consideradas veraces, y aquellas con más de dos señales fueron consideradas engañosas. En el segundo el método de aprendizaje de multitudes fue usado.

**Tabla 9. Tabla de clasificación del conjunto DeRev que muestra el rendimiento de los métodos de etiquetado**

Método	Exactitud	Engañoso		
		P	R	F
Voto de clases	75.4%	83.3	63.6	72.1
Algoritmo de Raykar et al.	76.3%	78.7	72.0	75.2

Para realizar la clasificación fue generado un vector de características por cada opinión. Dichas características se basaron en unigramas, bigramas, y trigramas de lemas y de etiquetas gramaticales. Entonces, se generaron dos listas de las frecuencias de cada lista de las fuentes de generación de características, una lista para opiniones engañosas y otra para opiniones veraces. Con esto se buscaron las características más frecuentes de cada clase. Para esto, solamente los n-gramas que aparecieran más de 300 veces, en cada lista de frecuencias, fueron tomados en cuenta.

En la Tabla 9 se muestra la clasificación del conjunto de datos DeRev mediante una máquina de vectores de soporte. La tabla muestra también la comparación entre métodos de etiquetado, teniendo un mayor rendimiento, aunque no significativo, el uso del algoritmo de Raykar et al.

## Capítulo IV. Enfoque propuesto

### *Contenido*

En esta sección presentamos nuestro método propuesto para la detección de engaño. Primero, en la Sección 2 detallamos las diversas fuentes de características que usamos. A continuación, en la Sección 4.2 describimos los diferentes conjuntos de datos que utilizamos para la evaluación, y finalmente en la Sección 4.3 damos detalles sobre nuestra construcción de vector de características.

## **4.1 Fuentes de generación de características**

Nos centramos específicamente en tres fuentes diferentes de características (LIWC, NS-GRAMAS, MTD) destinadas a comparar y combinar el método propuesto (LDA).

La mayoría de los estudios, presentadas como las normas actuales (ver Sección 3) utilizaron los unigramas como base para agregar nuevas características con el fin de obtener un mejor rendimiento. En cambio, hemos optado por utilizar un modelo de espacio de palabras (MTD, por sus siglas en inglés Word Space Model) (Sección 4.1.1), ya que mostró un rendimiento similar, además de ser una representación más simple.

Dado que el engaño involucra procesos cognitivos, otra fuente de características que seleccionamos fue el diccionario LIWC (Sección 4.1.3). Esta herramienta se ha utilizado para detectar el engaño generando información psicolingüística, debido a que fue creado para abordar estudios psicológicos. LIWC consiste en un conjunto de palabras con relaciones semánticas.

Sin embargo, el principal inconveniente de LIWC es el conjunto de palabras preestablecidas. Por lo tanto, para poder comparar el desempeño de LIWC con respecto a una herramienta que también genera grupos de palabras con relaciones semánticas, optamos por utilizar un modelo de espacio semántico continuo (LDA) (Sección 4.1.4). Este método crea automáticamente un conjunto de tópicos; cada tópico consiste en un conjunto de palabras con ciertas relaciones semánticas. A diferencia del LIWC, el conjunto de palabras de LDA puede cambiar dependiendo del conjunto de datos procesado, lo que sugiere que las categorías generadas son específicas de la colección de documentos, por lo que las características pueden ser más informativas.

Hasta ahora sólo hemos considerado la información léxica, incluida la semántica léxica; sin embargo, no se incluyó información de estilo de texto. Por lo tanto, y dado que distintos trabajos muestran que la información sintáctica aporta información valiosa, hemos explorado un enfoque basado en el estilo de texto (NS-GRAMAS)

(Sección 4.1.2). En contraste con otros métodos (BoW), los NS-GRAMAS puede considerar el contexto de las palabras en las oraciones.

#### 4.1.1 Matriz término-documento (MTD)

Los vectores de palabras se eligen de manera prominente en muchos trabajos debido al hecho de que generan características relevantes que ayudan a clasificar los textos. Por esta razón, hemos decidido analizar el rendimiento de las características basadas en una matriz de palabras, más específicamente, utilizando una representación binaria.

Para obtener una representación única, se forma una lista de todas las palabras  $W_1, W_2, \dots, W_n$  en el conjunto de datos. A continuación, analizamos cada documento buscando si  $W_n$  existe en el texto actual; si ese es el caso, la característica  $n$  ( $F_n$ ) se establece en uno, de lo contrario se establece en cero. En la **¡Error! No se encuentra el origen de la referencia.** se muestra cómo los vectores de características están representados.

#### 4.1.2 N-gramas sintácticos

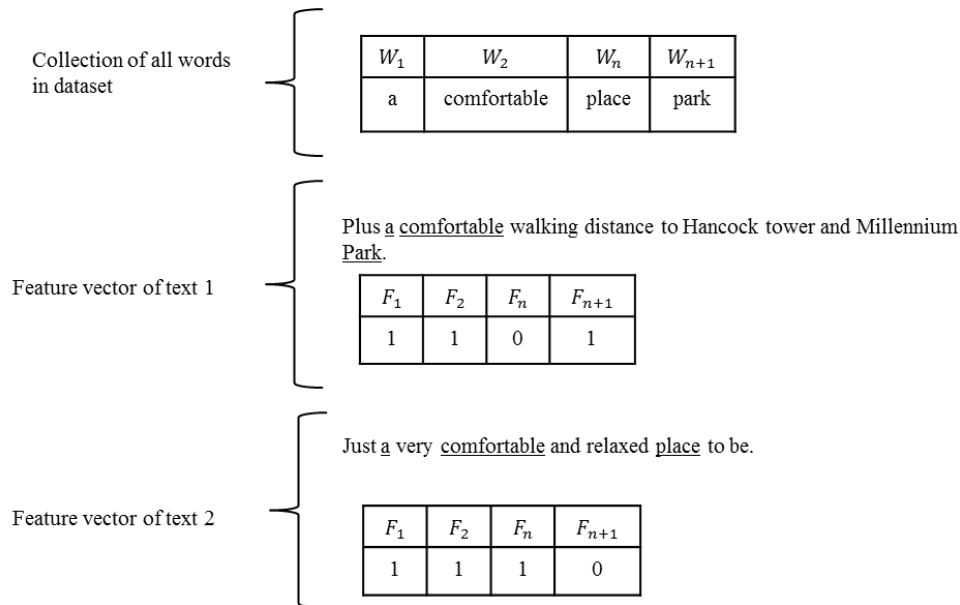
Los n-gramas sintácticos [39] (ns-gramas) son una característica relativamente nueva que surgió después de considerar ciertos inconvenientes de los n-gramas convencionales, siendo la principal desventaja de estos últimos que las relaciones de larga distancia no son correctamente capturadas, produciendo el efecto que los n-gramas convencionales parecían ser generados de una manera demasiado aleatoria.

La principal ventaja de los NS es que como se extraen a partir de una estructura sintáctica, estos pueden ser considerados como un fenómeno lingüístico. Esto no pasa con los n-gramas convencionales, generados de forma arbitraria.

La desventaja de poder generar n-gramas con información sintáctica es que se necesita un procesamiento sintáctico previo. Este procesamiento requiere de recursos adicionales como un analizador sintáctico para el lenguaje específico que

se esté manejando; estos recursos no siempre están disponibles para todos los lenguajes.

De forma similar a los n-gramas convencionales, los ns-gramas-grams pueden generarse a partir de diferentes tipos de elementos, por ejemplo, a partir de palabras, de etiquetas gramaticales, o de relaciones sintácticas.



**Figura 2. Ejemplo de construcción de vectores mediante un modelo de espacio de palabras**

Para probar el rendimiento de los ns-gramas-grams, en el estudio de Sidorov et al. [39] fue clasificado un conjunto de datos de 39 documentos escritos, en inglés, por tres autores diferentes. Asimismo, el 60% de los documentos fue usado como conjunto de entrenamiento y el 40% restante fue usado como conjunto de prueba. Se muestra en la Tabla 10 y la Tabla 11 una comparación de rendimiento obtenida por los autores. En dichas tablas el tamaño del archivo determina el umbral de n-gramas considerados como los más frecuentes; la etiqueta NA significa que no se alcanzó el umbral establecido. Podemos apreciar que tanto para bigramas y trigramas sintácticos logran clasificar correctamente todos los documentos. Con base en los buenos resultados obtenidos por los autores, optamos por usar los ns-gramas como

fuentes de generación de características para conocer su eficiencia en la detección de texto engañoso. Para tal propósito se detalla el procedimiento a continuación.

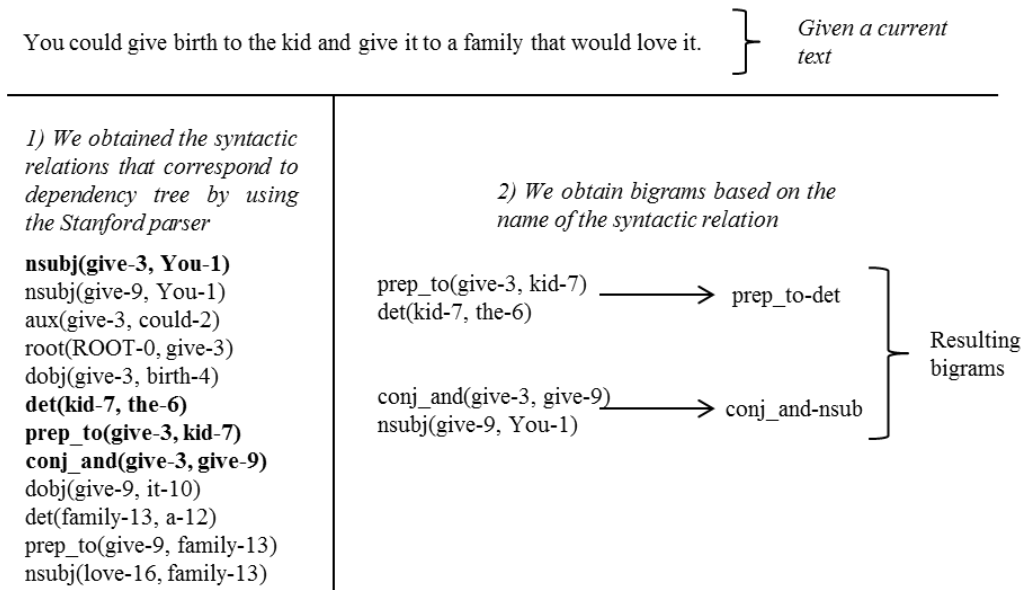
**Tabla 10. Comparación entre bigramas sintácticos y convencionales en la tarea de identificación de autoría**

Tamaño del archivo	Características			
	ns-gramas de rel. sintácticas	n-gramas de POS tags	n-gramas de caracteres	n-gramas de palabras
400	100%	90%	90%	86%
1,000	100%	95%	95%	86%
4,000	100%	NA	90%	86%
7,000	100%	NA	NA	86%
11,000	100%	NA	NA	89%

**Tabla 11. Comparación entre trigramas sintácticos y convencionales en la tarea de identificación de autoría**

Tamaño del archivo	Características			
	ns-gramas de rel. sintácticas	n-gramas de POS tags	n-gramas de caracteres	n-gramas de palabras
400	100%	90%	76%	81%
1,000	100%	90%	86%	71%
4,000	100%	100%	95%	95%
7,000	100%	100%	90%	90%
11,000	100%	95%	100%	90%

Para obtener ns-gramas, se construye un árbol sintáctico para recolectar relaciones sintácticas representadas por las aristas del árbol. Estas aristas son aquellas que unen palabras con la etiqueta apropiada.



**Figura 3. Proceso de generación de bigramas sintácticos dada una oración**

Utilizamos el analizador de Stanford [5] para generar las relaciones sintácticas que corresponden a los documentos analizados. No se utilizó la representación colapsada. Se muestra un ejemplo de la representación de Stanford de las relaciones sintácticas en el paso 1 de la Figura 3. En el paso 2, las relaciones útiles se seleccionan para generar bigramas; Se puede ver que `prep_to` (give-3, kid-7) está relacionado con `det` (kid-7, the-6) porque kid-7 los vincula a ambos. De esta manera, obtenemos el bigrama `prep_to-det`.

Los documentos pueden ser representados por sus relaciones sintácticas mediante el uso de un vector de características. Esto se hace de la misma manera que mostramos en la sección anterior, ver **¡Error! No se encuentra el origen de la referencia.**, pero esta vez reemplazando las palabras por n-gramas sintácticos.

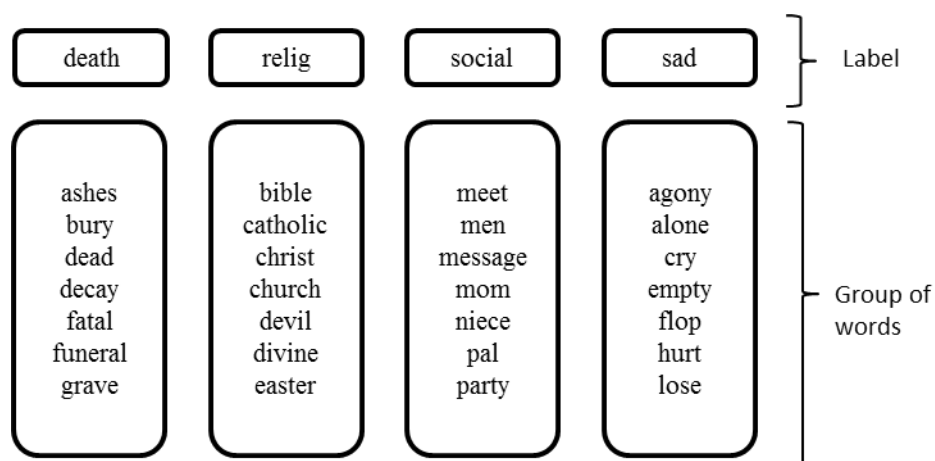
#### 4.1.3 Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) es una herramienta basada en un diccionario de palabras [31]. El diccionario de LIWC contiene grupos de palabras etiquetados por los humanos y fue utilizado originalmente en trabajos relacionados



con el análisis psicológico. Las aplicaciones de este recurso han crecido recientemente; Por ejemplo, el LIWC se ha utilizado en la lingüística computacional como una fuente de características para la identificación de autoría [16] y detección de engaño [25], entre otras aplicaciones.

En este trabajo utilizamos la versión 2007 de LIWC. Esta versión consta de cerca de 4500 palabras agrupadas en 80 categorías o dimensiones. La Figura 4 muestra un ejemplo de algunos grupos de palabras que podemos encontrar en LIWC.

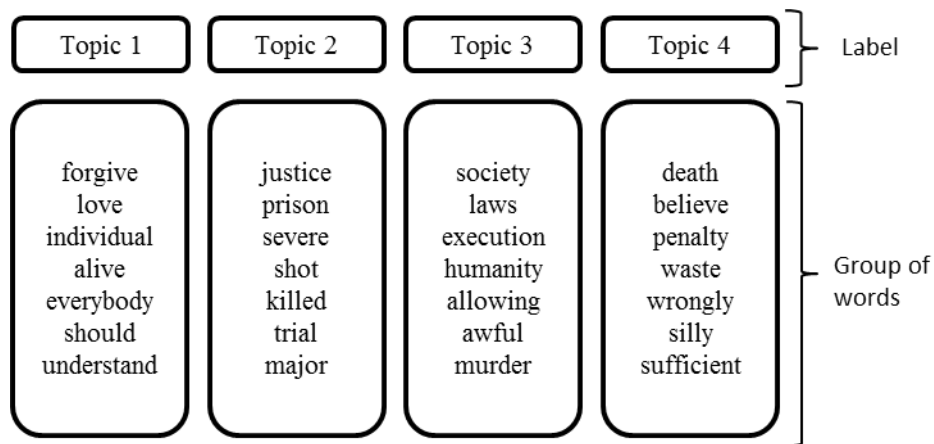


**Figura 4. Ejemplo de algunos grupos de palabras, en LIWC, con su respectiva etiqueta. Es mostrada una muestra de solo 7 palabras por etiqueta en este ejemplo.**

#### 4.1.4 Modelo de espacio semántico continuo - Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) es un modelo probabilístico generativo [3] para colecciones de datos discretos, tales como colecciones de texto.

LDA representa a los documentos como una mezcla de diferentes tópicos; cada tópico consiste en un conjunto de palabras que mantienen algún vínculo semántico entre ellas. Las palabras, a su vez, se eligen en función de una probabilidad. Se repite el proceso de selección de tópicos y palabras para generar un documento o un conjunto de documentos. Como resultado, cada documento generado se compone de diferentes tópicos.



**Figura 5. Ejemplo de generación de tópicos, de textos del tema “pena de muerte”, mediante LDA**

De forma simplificada, el proceso generativo de LDA consta de los siguientes pasos:

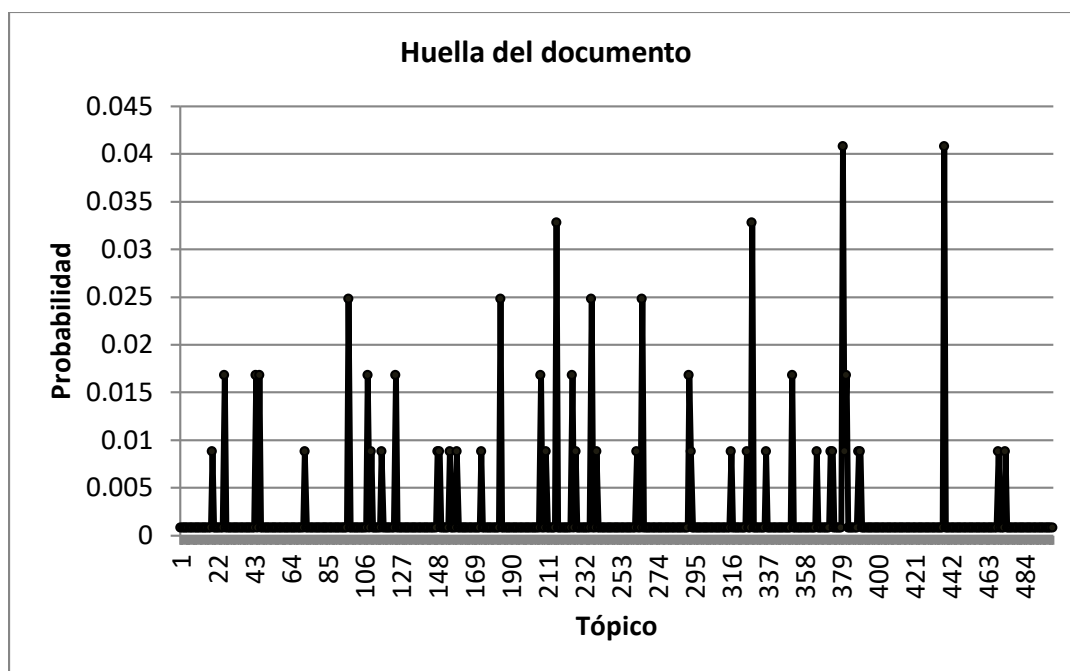
1. La determinación de las N palabras que un documento contendrá de acuerdo con la distribución de Poisson.
2. La elección de una mezcla de tópicos para el documento de acuerdo con la distribución de Dirichlet con respecto a un conjunto fijo de K tópicos.
3. La generación de cada palabra en el documento mediante:
  - a) Elegir un tema
  - b) Utilizar el tema elegido para generar la palabra

Utilizando un modelo generativo de forma inversa, LDA analiza el conjunto de documentos para encontrar el conjunto más probable de tópicos posiblemente tratados en un documento.

Podemos considerar a LDA como una herramienta que genera grupos de palabras similares, tal como LIWC; pero a diferencia de LIWC, LDA genera automáticamente los grupos de palabras (tópicos). Además, los grupos de palabras de LDA no están

etiquetados, y su contenido es diferente dependiendo del corpus donde se entrena LDA.

Vea la Figura 5 para un ejemplo de palabras que se obtienen usando LDA.



**Figura 6. Ejemplo de documento procesado por LDA. Se muestra la distribución de tópicos.**

La Figura 6 muestra un documento procesado mediante LDA (antes de la binarización). Cada tópico mostrado tiene una probabilidad particular de existir dentro del documento actual. De esta manera, una alta probabilidad de un tópico (x) en el documento (y) indica que es más probable que x aparezca en y. Cada documento de la colección tiene una distribución diferente de una cantidad fija de temas.

## 4.2 Conjuntos de datos

Para evaluar nuestro método, se utilizaron tres conjuntos de datos diferentes: el corpus DeRev (DEception in REViews) [11], el corpus OpSpam (Opinion SPAM) [29], [28] y un corpus creado por opiniones sobre tres tópicos controversiales [32]. El primer corpus fue recolectado utilizando el engaño sancionado, mientras que los dos

últimos fueron recolectados usando un engaño no sancionado [14]. Los detalles sobre estos dos tipos de engaño se describen en las siguientes secciones.

#### **4.2.1 Conjunto de datos DeRev**

El conjunto de datos DeRev es un corpus compuesto de opiniones engañosas y veraces obtenidas de una página web de Amazon. Este corpus consiste en opiniones sobre libros. Es un estándar de oro (*gold standard*) que contiene 236 textos de los cuales 118 son veraces y 118 son engañosos. Tanto los textos engañosos como los veraces fueron obtenidos de Amazon.

Para lograr un alto grado de confianza en una colección correctamente etiquetada de textos engañosos, los autores de DeRev consideraron dos investigaciones; la primera, de Sandra Parker, fue publicada en Money Talks News; mientras que el segundo, de David Streitfeld, apareció en The New York Times. Parker afirmó que recibió un pago con el propósito de escribir opiniones sobre 22 libros. Streitfeld, por otra parte, dio a conocer cuatro libros en los que sus autores admitieron haber pagado por escribir opiniones engañosas; a partir de esto, se hizo un análisis y la selección se centró en veinte escritores de opiniones falsas, creando un corpus con 96 opiniones engañosas.

Finalmente, para obtener los 118 textos veraces, los autores del corpus DeRev tomaron en cuenta ciertos aspectos para asegurar una alta probabilidad de que la selección fuera correcta. Los textos no deben tener ninguna señal de engaño. Por esta razón, la selección se centró principalmente en aspectos tales como si las opiniones fueron escritas por usuarios que habían utilizado sus nombres reales, y si las opiniones fueron escritas por los usuarios que compraron el libro a través de Amazon, entre otros.

En este conjunto de datos, los textos engañosos y veraces no se obtuvieron de manera deliberada, es decir, a los participantes no se les pidió que escribieran mentiras; en cambio, los textos se obtuvieron después de que el participante había

mentido. Por esta razón, se dice que la forma de construir este texto es por engaño no sancionado.

### Ejemplo de texto engañoso

Circle of Lies, by Douglas Alan is a fast paced, gritty crime thriller that introduces us to John Delaney, ex-cop turned lawyer who finds himself defending his best friend against embezzlement and murder charges when no one else will believe him. Alan, a retired trial lawyer, gives us a rather unique look at the "men behind the curtains" of the American justice system with this fast paced, exciting read. Highly recommended!

### Ejemplo de texto veraz

This book is well written and very funny, it is a good "guys"; kind of book, and it is an easy read, but don't let that fool you. The story is good and it is in the painful details. I felt like I was working right there with them. However, it is crude and not for the casual reader. I try to read both classic and non-classic writers of all eras, and I am always looking for something weird, funny, and cool to mix it up and this was it.

#### 4.2.2 Conjunto de datos OpSpam

El conjunto de datos OpSpam es un corpus compuesto de opiniones falsas y veraces. Estas son opiniones sobre diferentes hoteles.

Para la generación de opiniones engañosas los creadores del corpus OpSpam utilizaron Amazon Mechanical Turk (AMT). Los autores mostraron a cada turker (trabajador de AMT) el nombre del hotel y su respectivo sitio web para que pudieran completar la tarea. Al utilizar el sitio web de AMT, se les pidió a los turkers que imaginaran que trabajaban en un hotel y que el administrador les pidió que

escribieran una opinión sobre el lugar, como si fueran invitados. La opinión debería parecer real y destacar los aspectos positivos del hotel. Las opiniones se limitaron a una por turker para evitar que la misma persona escribiera más de una. Asimismo, la tarea se limitaba a aquellos que vivían en los Estados Unidos; y, además, los turkers tendrían que alcanzar una calificación de aprobación de al menos el 90%. Los turkers tenían un máximo de treinta minutos para escribir la opinión y se les pagó un dólar por cada opinión aceptada. Con este procedimiento, los autores de OpSpam lograron recopilar 400 textos engañosos.

### Ejemplo de texto engañoso

I had a wonderful time at the James Hotel while on business in Chicago. The rooms are modern, tasteful and well-kept, while the staff was responsive and efficient. This was the perfect place to unwind after working all day, but also provided an ideal atmosphere to do some more work.

### Ejemplo de texto veraz

Despite what other are saying, this was one, if not the best Hotel stay in Chicago I have had. I travel to the Big City about three times a year for pleasure and The James rates up with the best. I was upgraded and the staff made the stay worth my special weekend visit. I don't believe you will be disappointed.

Por otra parte, las opiniones veraces se recogieron de una página web de TripAdvisor. En primer lugar, 6,977 opiniones fueron extraídas de los veinte hoteles más populares. A continuación, los autores eliminaron 3,130 que no tenían cinco estrellas, 41 que no estaban escritas en inglés, 75 que tenían menos de 150 caracteres y 1,607 que habían sido publicadas por personas que opinaban por

primera vez en Amazon. Finalmente, los autores de OpSpam seleccionaron 400 de los textos restantes.

Se recopilaron un conjunto de datos compuesto de 800 textos en total. Para formar este corpus, se pidió a los participantes que escribieran mentiras para obtener el texto engañoso. En consecuencia, el corpus se formó mediante el engaño sancionado.

#### **4.2.3 Conjunto de datos de tópicos controversiales**

##### **Ejemplo de texto engañoso**

Abortion is despicable because we're giving women the choice to kill life. Where is the unborn child's voice? Do we not care about those who can't defend themselves? Abortion is systemic murder that needs to be stopped.

##### **Ejemplo de texto veraz**

Abortion is a choice. It should be made by both parties if possible, but the woman should have final say if there is disagreement. I think that with counseling, and all the other options available, it is not morally wrong for a woman to choose to have an abortion. A fetus is not viable outside the womb before 26 weeks, and then only barely, and with very expensive care.

Se trata de un corpus compuesto por 600 opiniones sobre tres temas controversiales: el aborto (200), la pena de muerte (200) y un mejor amigo (200). El corpus consta de 100 textos engañosos y 100 textos verídicos por cada tópico (dando el total de 600 textos). La recopilación de textos se realizó a través de AMT y la tarea se limitó a los turkers que vivían en los Estados Unidos.

Para obtener textos veraces, los autores solicitaron a los participantes su opinión real sobre cada uno de los tópicos; después, se pidió a los participantes que mintieran acerca de su opinión real, mediante la cual se obtuvieron textos engañosos. El método utilizado para recolectar este corpus fue un engaño sancionado.

Este corpus también contiene textos en inglés de la India y en el español de México; sin embargo, esos textos no fueron utilizados en este trabajo.

#### 4.2.4 Análisis de los conjuntos de datos

Para proporcionar una visión más profunda de los corpora, mostramos en la Tabla 12 un recuento de tokens y tipos de los diferentes conjuntos de datos descritos anteriormente, donde los tokens muestran la cantidad total de palabras contenidas en los documentos y los tipos representan el número total de palabras no repetidas encontradas en documentos. Las *stopwords* (palabras que carecen de significado por sí solas) se mantuvieron para todos los experimentos.

**Tabla 12. Se muestran la cantidad de tipos y tokens en los corpora**

Corpus	Número de textos	Tokens	Tipos	Promedio de tokens por doc.
OpSpam	800	96,793	6,469	121
DeRev	236	29,990	5,162	127
Aborto	200	15,958	1,997	80
Mejor amigo	200	11,717	1,718	59
Pena de muerte	200	15,615	2,034	78

La Tabla 13 muestra el promedio de tipos con respecto a cada corpus. Esto es, el promedio de palabras no repetidas que los documentos del corpus tienen en común. Los resultados se muestran tanto en todos los documentos (engañosos + veraces) como en clases separadas. Los resultados de esta tabla también indican que es menos probable que haya sesgos en los conjuntos, ya que el promedio de palabras compartidas es muy cercano entre clases unidas y clases separadas, lo que sugiere



que no hay demasiadas palabras predominantes por clase. Por el contrario, si hubiera palabras predominantes por clase, inferimos que el número de palabras compartidas debería haber disminuido significativamente, cuando las clases se combinaron, debido al hecho de que tendrían pocas palabras compartidas entre ellas.

**Tabla 13. Promedio de tipos compartidos en los diferentes corpora**

Corpus	Promedio de tipos compartidos		
	Engañoso + veraz	Engañoso	Veraz
OpSpam	19	19	19
DeRev	14	16	13
Aborto	14	12	16
Mejor amigo	9	8	12
Pena de muerte	14	13	15

### 4.3 Construcción de los vectores de características

Se realizaron varios experimentos utilizando diferentes fuentes (descritas en la Sección 2) y combinaciones de estas para encontrar la mejor combinación de características para la detección del engaño (de acuerdo a los conjuntos de datos analizados en este trabajo). Para todos los experimentos, la clasificación se llevó a cabo a través del algoritmo de aprendizaje automático Naïve Bayes multinomial implementado bajo WEKA [18] con una validación cruzada de cinco pliegues.

En un intento de generar un modelo que represente mejor la detección de engaño, implementamos una selección de atributos con el propósito de eliminar características repetitivas e irrelevantes.

La selección de atributos [15] ayuda, en algunos casos, a obtener un modelo para identificar el texto engañoso con mayor eficiencia. También ayuda a reducir las dimensiones de los vectores de las características. Por ejemplo, en el presente

estudio, los vectores compuestos de aproximadamente 4,000 características se redujeron a aproximadamente 60 características.

**Tabla 14. Comparación de exactitud con respecto a los valores binarios y la selección de atributos (SeAt)**

Corpus	Accuracy		
	Binarios + AtSe	No binarios + AtSe	Binary sin AtSe
Aborto	<b>87.5%</b>	75.2%	72.2%
Mejor amigo	<b>87.0%</b>	82.0%	76.8%
Pena de muerte	<b>80.0%</b>	69.3%	61.5%
DeRev	<b>94.9%</b>	74.9%	88.8%
OpSpam	<b>90.9%</b>	88.5%	87.2%

En el proceso de selección de atributos, la herramienta WEKA se utilizó con un evaluador basado en la correlación propuesta por Hall [17]; además, se utilizó un criterio de búsqueda basado en un algoritmo de escalando la colina con retroceso (*backtraking*). Dicha combinación mostró un aumento significativo en la precisión. Se establecieron parámetros específicos para la selección de atributos en Weka de la siguiente manera: el evaluador utilizado fue CfsSubSetEval (numThreads 1, poolSize 1) y la búsqueda de estrategia utilizada fue BestFirst (dirección: Forward, searchTermination 5). Además, hemos intentado con otras estrategias de búsqueda como ExhaustiveSearch, pero la comprobación de todas las posibilidades en un gran conjunto de atributos fue muy demandante en términos de tiempo. Igualmente, hemos utilizado la búsqueda genética, misma que, aunque se desarrolló más rápido, no proporcionó buenos resultados, incluso con el aumento del número de generaciones, tamaño de la población y el ajuste de la probabilidad de

recombinación/mutación. Eventualmente, BestFirst fue más eficiente aplicado a la representación de documentos utilizados en este trabajo. Un estudio más detallado de todos los enfoques de selección de atributos se ha dejado como trabajo futuro.

Conjuntamente, encontramos que la binarización de los vectores de características dio lugar a un modelo más preciso para la detección de textos engañosos: La Tabla 3 muestra los resultados de la detección de engaño con características LDA + MTD y el efecto de la aplicación de binarización y selección de atributos, la binarización se implementó antes del proceso de selección de atributos. Dado que la binarización con la selección de los atributos en todos los casos presentaron los mejores resultados, utilizamos esta configuración en todos los próximos experimentos. Las salidas detalladas (precisión, exhaustividad y medida F) de la binarización con selección de atributos (Binario con SeAt) mostradas en la Tabla 14 se muestran en la Sección 5.

A continuación, mostramos detalles de la conversión de todas las características en valores binarios.

- LIWC generó vectores de 64 características. Los medios para obtener cada vector fueron los siguientes: dado un documento y las 64 categorías, si una palabra de una categoría actual se encontró en el documento, entonces esa característica tenía el valor de uno, de lo contrario tenía el valor de cero.
- Los medios para generar vectores de características usando ns-gramas fueron los siguientes: primero, formamos una lista de ns-gramas no repetidos obtenidos en todos los documentos del conjunto de datos a analizar. A continuación, dado un documento y la lista de ns-gramas, si el ns-grama actual se encontró en el documento, entonces el valor de la característica se estableció en uno, de lo contrario se estableció en cero.
- Como se mencionó anteriormente, LDA muestra, como resultado, vectores de características con valores reales (probabilidades de pertenecer a cada tópico). Por lo tanto, procedimos a convertir valores de las características en valores binarios. Para ello, se calculó un umbral dividiendo la suma de todas

las probabilidades de pertenencia por el número de tópicos establecidos. Cada probabilidad que es igual o mayor que el umbral se convirtió en uno; de lo contrario se convirtió en cero.

**Tabla 15. Exactitud obtenida con diferentes valores del número de tópicos**

Tópicos	Medida F (%)				
	Aborto	Mejor amigo	Pena de muerte	DeRev	OpSpam
100	69.46	78.01	57.75	81.25	87.36
200	75.52	82.61	65.18	86.47	87.93
300	81.17	85.38	71.71	90.20	89.93
400	84.50	87.05	77.22	92.35	90.06
<b>500</b>	<b>86.70</b>	<b>87.58</b>	<b>80.41</b>	<b>94.53</b>	<b>90.10</b>
600	87.40	87.00	81.18	95.54	91.34
700	87.52	86.29	81.86	95.74	89.53
800	87.76	85.03	82.05	95.68	90.15
900	87.38	84.33	81.82	95.97	90.68
1000	86.92	83.29	81.72	96.06	90.25

Hemos intentado usar el esquema de ponderación TF-IDF (Term Frequency-Inverse Document Frequency) y hemos detectado que el TF booleano da más información útil para la clasificación que el TF basado en recuentos de repetición de palabras. Además, la exclusión del IDF de la ecuación no representó un cambio significativo en los resultados. Como resultado, el esquema de ponderación para NS-GRAMAS, MTD y LIWC es dado por TF booleano.

#### **4.4 Identificación del número de tópicos óptimo**

LDA requiere que se especifique el número de tópicos que serán generados; cualquier cambio en este parámetro puede cambiar la precisión de la clasificación. Por esta razón, es necesario encontrar un valor apropiado.

Para encontrar el número de tópicos que permitieron una clasificación óptima, realizamos varios experimentos. Los resultados de estos experimentos se muestran en la Tabla 15; en esta tabla, el número de tópicos se comparan con la medida F obtenida. También, se puede observar que al aumentar el número de tópicos es posible alcanzar un punto óptimo del cual el aumento del número de los mismos no implica una disminución de la medida F (es decir, 500 tópicos).

Aunque es posible aumentar ligeramente la precisión de algunos conjuntos de datos mediante el ajuste del número de tópicos, estábamos interesados en encontrar un método general con características que producen un buen rendimiento con diferentes tipos de engaño. De esta manera, buscamos un conjunto de parámetros comunes que funcionen para todos los conjuntos de datos.

A partir de entonces, en todos los experimentos se utilizaron 500 temas, es decir, cada documento procesado por LDA genera un vector de 500 características y cada uno de ellos está representado por una probabilidad de pertenecer a cada tema (véase la Sección 4.1.4, Figura 6).

## Capítulo V. Resultados

### *Contenido*

En este capítulo mostramos los resultados obtenidos al usar el enfoque propuesto para la detección de texto engañoso en diferentes casos de estudio dependiendo del uso de los conjuntos de datos: en dominio, de dominio mezclado y de dominio cruzado.

## 5.1 Detección de engaño

En esta sección presentamos resultados detallados sobre la identificación del engaño con base en distintos métodos de generación de características. Las tablas de cada corpus clasificado contienen los siguientes valores: exactitud, precisión (P), exhaustividad (R) y medida F (F). Aunque la exactitud es una medida utilizada en muchas investigaciones sobre la detección de engaño y nos proporciona un punto de comparación con otros resultados, también optamos por mostrar precisión, exhaustividad y medida F; esto permite un análisis más profundo de los resultados. De esta manera, la precisión muestra el porcentaje de textos seleccionados que son correctos, mientras que el recuerdo muestra el porcentaje de textos correctos que se seleccionan. Por último, la medida F es la medida combinada para evaluar la compensación P/R.

Se obtuvieron estos valores para los siguientes métodos: latent Dirichlet allocation (LDA), el modelo de espacio de palabras (MTD), n-gramas sintácticos (NS-GRAMAS) y linguistic inquiry and word count (LIWC); así como para las combinaciones de éstos.

### 5.1.1 Clasificación de un dominio específico

El objetivo de combinar diferentes técnicas de PLN es encontrar cuál es la combinación de características que mejor favorece la exactitud y medida F en la clasificación de texto engañoso y veraz. Las palabras contienen mucha información por sí mismas; esto puede ser confirmado por el gran número de investigaciones (ver Sección 3) que utilizan unigramas como base para agregar nuevas características, esto con el objetivo de obtener un mejor rendimiento. De manera similar, usamos LDA para complementar las características de una MTD. Los resultados en esta sección muestran que las características generadas con la MTD en combinación con las de LDA son mejores que otros métodos utilizados (tales como, LIWC, NS-GRAMAS) en la mayoría de los casos. En la clasificación del conjunto de datos OpSpam, que se muestra en la Tabla 16, podemos ver que la combinación de LDA y

MTD muestran un aumento en la precisión con respecto a las mismas técnicas evaluadas por separado. Esto se debe a que las características generadas usando LDA complementan de una manera favorable las generadas por la MTD.

**Tabla 16. Clasificación del corpus OpSpam mediante el uso de diferentes fuentes de generación de características**

Método	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
LDA	72.0%	72.8	70.3	71.5	71.3	73.8	72.5
MTD	87.9%	87.0	89.0	88.0	88.7	86.8	87.7
NS-GRAMAS (bigramas)	68.6%	72.8	59.5	65.5	65.8	77.8	71.2
NS-GRAMAS+MTD	86.0%	88.7	82.5	85.5	83.6	89.5	86.5
LIWC	65.0%	65.7	62.8	64.2	64.4	67.3	65.8
LIWC + MTD	87.9%	89.4	86.0	87.6	86.5	89.8	88.1
LDA+MTD	90.9%	94.8	86.5	90.5	87.6	95.3	91.3
LDA+LIWC	72.3%	73.9	68.8	71.2	70.8	75.8	73.2
LDA+NS-GRAMAS	74.9%	72.6	80.0	76.1	77.7	69.8	73.5
LDA+MTD+LIWC	87.1%	87.2	87.0	87.1	87.0	87.3	87.1
LDA+LIWC+NS- GRAMAS	79.7%	77.8	83.3	80.4	82.0	76.3	79.0
LDA+MTD+NS- GRAMAS	88.9%	88.6	89.3	88.9	89.2	88.5	88.8
LDA+MTD+LIWC+NS- GRAMAS	88.1%	88.2	88.0	88.1	88.0	88.3	88.1
Mile Ott <i>et al.</i> 2011 (LIWC + bigramas)	89.8%	89.8	89.8	89.8	89.8	89.8	89.8
Song Feng <i>et al.</i> 2012 (syntactic rel.+ unigramas)	<b>91.2%</b>	-	-	-	-	-	-
Donato <i>et al.</i> 2015	90.2%	-	-	-	-	-	-

Hay casos en los que hay muchas características que representan el texto veraz, mientras que sólo algunas características representan texto engañoso, o viceversa.



LDA puede ayudar a minimizar este problema aumentando el número de características relevantes que conducen a una clasificación de texto más eficiente.

**Tabla 17. Número de características relevantes obtenidas con la combinación de LDA y MTD**

Método	MTD		MTD+LDA	
	Veraz	Engañoso	Veraz	Engañoso
Aborto	43	12	45	25
Mejor amigo	24	17	30	20
Pena de muerte	32	16	39	29
DeRev	27	52	37	57
OpSpam	31	22	31	24

Para estudiar este efecto, enumeramos el número de atributos, después de la selección del atributo, que están fuertemente vinculados con el texto engañoso o con el texto veraz, respectivamente. Por ejemplo, en 60 características relevantes, hay un número específico de características que están fuertemente correlacionadas con la clase engañosa, y otras características con la clase veraz (ver las dos primeras columnas de MTD en la Tabla 17). Además, el número de características relevantes puede aumentarse combinando métodos de generación de características; en la Tabla 17 se muestra el número de características relevantes obtenidas para cada clase utilizando sólo MTD en comparación con el número de características obtenidas mediante el uso de la combinación LDA y MTD. Se puede ver que en todos los casos se presenta un aumento de características del texto engañoso.

Los vectores de características generadas por la combinación de características de LIWC y MTD mostraron resultados similares, aunque con un rendimiento inferior en comparación con la combinación LDA + MTD. Sin embargo, la principal desventaja de LIWC, con respecto a LDA, es que la herramienta LIWC es de fabricación

humana; esto significa que cada vez que se analiza un idioma diferente, es necesario obtener la herramienta para ese idioma específico. Por otro lado, LDA no requiere ninguna información dependiente del lenguaje para analizar un conjunto de datos.

**Tabla 18. Clasificación del corpus DeRev mediante el uso de diferentes fuentes de generación de características**

Método	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
LDA	80.5%	79.0	83.1	81.0	82.1	78.0	80.0
MTD	91.9%	90.9	93.2	92.1	93.0	90.7	91.8
NS-GRAMAS (bigramas)	75.4%	80.0	67.8	73.4	72.1	83.1	77.2
NS-GRAMAS+MTD	91.1%	89.4	93.2	91.3	92.9	89.0	90.9
LIWC	69.9%	77.0	56.8	65.4	65.8	83.1	73.4
LIWC + MTD	91.5%	92.2	90.7	91.5	90.8	92.4	91.6
LDA+MTD	<b>94.9%</b>	94.2	95.8	95.0	95.7	94.1	94.9
LDA+LIWC	73.3%	81.6	60.2	69.3	68.5	86.4	76.4
LDA+NS-GRAMAS	78.4%	83.2	71.2	76.7	74.8	85.6	79.8
LDA+MTD+LIWC	90.7%	98.0	83.1	89.9	85.3	98.3	91.3
LDA+LIWC+NS- GRAMAS	79.2%	86.3	69.5	77.0	74.5	89.0	81.1
LDA+MTD+NS- GRAMAS	91.5%	90.2	93.2	91.7	93.0	89.8	91.4
LDA+MTD+LIWC+NS- GRAMAS	83.0%	84.8	80.5	82.6	81.5	85.6	83.5
Fornaciari and Poesio 2014	76.27%	-	-	-	-	-	-

Otro enfoque probado fue el uso de ns-gramas. Al igual que LIWC, es necesaria información previa dependiente del lenguaje, debido a que los árboles sintácticos se obtienen basándose en la probabilidad de ocurrencia en un corpus previamente etiquetado. Los bigramas sintácticos mostraron una precisión aceptable, pero la combinación de sus características puede empeorar los resultados. Por ejemplo, en los conjuntos de datos mostrados en la Tabla 18 y Tabla 20, la combinación de

características entre ns-gramas y MTD disminuyó la exactitud de la clasificación. Por lo tanto, esta combinación específica de características era desventajosa.

**Tabla 19. Clasificación del tópico Aborto mediante el uso de diferentes fuentes de generación de características**

Método	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
LDA	79.5%	76.6	85.0	80.6	83.1	74.0	78.3
MTD	82.5%	82.8	82.0	82.4	82.2	83.0	82.6
NS-GRAMAS (bigramas)	65.8%	73.2	52.0	60.8	62.8	81.0	70.7
NS-GRAMAS+MTD	83.0%	83.0	83.0	83.0	83.0	83.0	83.0
LIWC	72.5%	72.7	72.0	72.4	72.3	73.0	72.6
LIWC + MTD	83.5%	81.3	87.0	84.1	86.0	80.0	82.9
LDA+MTD	<b>87.5%</b>	87.9	87.0	87.4	87.1	88.0	87.6
LDA+LIWC	80.0%	76.8	86.0	81.1	84.1	74.0	78.7
LDA+NS-GRAMAS	73.0%	78.8	63.0	70.0	69.2	83.0	75.5
LDA+MTD+LIWC	76.5%	76.2	77.0	76.6	76.8	76.0	76.4
LDA+LIWC+NS-GRAMAS	77.0%	76.5	78.0	77.2	77.6	76.0	76.8
LDA+MTD+NS-GRAMAS	86.5%	90.1	82.0	85.9	83.5	91.0	87.1
LDA+MTD+LIWC+NS-GRAMAS	82.5%	77.8	91.0	83.9	89.2	74.0	80.9
Perez-Rosas and Mihalcea 2014	80.3%	-	-	-	-	-	-
Song Feng <i>et al.</i> 2012(syntactic rel. + unigramas)	77.0%	-	-	-	-	-	-

Es importante señalar que con el método propuesto (LDA + MTD) se mantiene una precisión competitiva para detectar textos engañosos en los diferentes conjuntos de datos clasificados, a diferencia de algunas otras técnicas de PLN, como la combinación entre relaciones sintácticas y unigramas utilizada por Song Feng et al. [9]. El resultado del autor es bueno en la evaluación del conjunto de datos OpSpam, que se muestra en la Tabla 16. Sin embargo, este desempeño no se mantuvo

cuando los autores clasificaron conjuntos de datos sobre el aborto, la pena de muerte y el mejor amigo; como se muestra en la Tabla 19, Tabla 20, y Tabla 21.

**Tabla 20. Clasificación del tópico "mejor amigo" mediante el uso de diferentes fuentes de generación de características**

Método	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
LDA	82.0%	82.7	81.0	81.8	81.4	83.0	82.2
MTD	85.5%	84.5	87.0	85.7	86.6	84.0	85.3
NS-GRAMAS (bigramas)	73.0%	81.1	60.0	69.0	68.3	86.0	76.1
NS-GRAMAS+MTD	84.5%	84.8	84.0	84.4	84.2	85.0	84.6
LIWC	75.5%	74.3	78.0	76.1	76.8	73.0	74.9
LIWC + MTD	83.5%	82.5	85.0	83.7	84.5	82.0	83.2
LDA+MTD	<b>87.0%</b>	85.6	89.0	87.3	88.5	85.0	86.7
LDA+LIWC	78.0%	73.7	87.0	79.8	84.1	69.0	75.8
LDA+NS-GRAMAS	85.0%	88.9	80.0	84.2	81.8	90.0	85.7
LDA+MTD+LIWC	77.0%	72.5	87.0	79.1	83.8	67.0	74.4
LDA+LIWC+NS- GRAMAS	79.5%	78.6	81.0	79.8	80.4	78.0	79.2
LDA+MTD+NS- GRAMAS	85.0%	92.7	76.0	83.5	79.7	94.0	86.2
LDA+MTD+LIWC+NS- GRAMAS	86.5%	94.0	78.0	85.2	81.2	95.0	87.6
Perez-Rosas and Mihalcea 2014	75.9%	-	-	-	-	-	-
Song Feng <i>et al.</i> 2012 (syntactic rel. + unigramas)	85.0%	-	-	-	-	-	-

Con base en el análisis de los resultados obtenidos, se encontró que la mejor combinación de características fue LDA y la matriz término-documento (LDA + MTD). Esta combinación muestra una precisión del 90,9% para el conjunto de datos OpSpam y del 94,9% en el conjunto de datos DeRev. Se obtuvieron exactitudes del

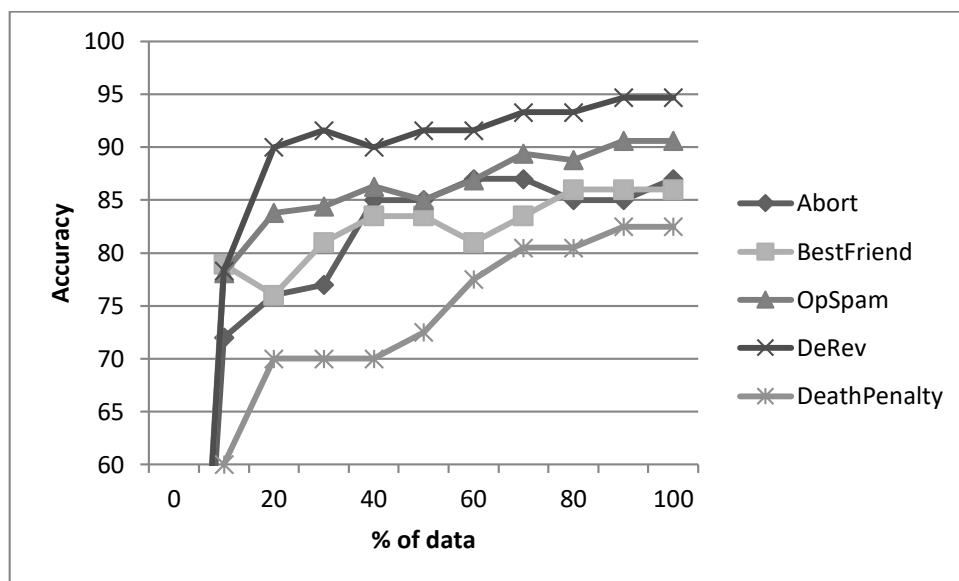
87,5%, 87,0% y 80,0% para los temas de aborto, pena de muerte y mejor amigo, respectivamente.

**Tabla 21. Clasificación del tópico "pena de muerte" mediante el uso de diferentes fuentes de generación de características**

Método	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
LDA	77.0%	76.5	78.0	77.2	77.6	76.0	76.8
MTD	78.5%	77.7	80.0	78.8	79.4	77.0	78.2
NS-GRAMAS (bigramas)	61.5%	66.2	47.0	55.0	58.9	76.0	66.4
NS-GRAMAS+MTD	78.5%	78.8	78.0	78.4	78.2	79.0	78.6
LIWC	65.5%	64.5	69.0	66.7	66.7	62.0	64.2
LIWC + MTD	78.0%	78.6	77.0	77.8	77.5	79.0	78.2
LDA+MTD	<b>80.0%</b>	78.8	82.0	80.4	81.3	78.0	79.6
LDA+LIWC	66.0%	79.6	43.0	55.8	61.0	89.0	72.4
LDA+NS-GRAMAS	70.0%	72.7	64.0	68.1	67.9	76.0	71.7
LDA+MTD+LIWC	61.0%	60.4	64.0	62.1	61.7	58.0	59.8
LDA+LIWC+NS- GRAMAS	75.0%	78.4	69.0	73.4	72.3	81.0	76.4
LDA+MTD+NS- GRAMAS	74.0%	91.4	53.0	67.1	66.9	95.0	78.5
LDA+MTD+LIWC+NS- GRAMAS	79.0%	73.4	91.0	81.3	88.2	67.0	76.1
Perez-Rosas and Mihalcea 2014	77.2%	-	-	-	-	-	-
Song Feng <i>et al.</i> 2012 (syntactic rel. + unigramas)	71.5%	-	-	-	-	-	-

La Figura 7 muestra el proceso de aprendizaje de cada conjunto de datos. Para cada porcentaje de datos muestreados al azar, del 10% al 100%, se utilizó una parte como conjunto de entrenamiento (80%) y el resto (20%) como conjunto de prueba. Por ejemplo, si se utilizó el 60% de los datos, digamos, 600 de 1000 registros, 480 se

utilizarían como entrenamiento y 120 como prueba. Este gráfico sugiere que, para algunos corpora, el uso de más datos podría producir un pequeño aumento en la exactitud, mientras que para otros casi no hay cambio entre el 90% o el 100% de todos los datos.



**Figura 7. Curva de aprendizaje de los diferentes corpora**

### 5.1.2 Clasificación de dominios mezclados

El objetivo principal de los experimentos mostrados a continuación fue investigar hasta qué punto la generación de características y el aprendizaje automático puede utilizarse cuando combinamos los cinco conjuntos de datos en diferentes dominios. Con esto, el conjunto de entrenamiento contendría temas que también incluiría el conjunto de prueba (pero no de textos de un solo dominio). En este caso, nuevamente una combinación de características LDA y MTD produjo el mejor resultado; ver Tabla 22. La SVM genera una exactitud mayor que el clasificador NB, ver Tabla 23.

**Tabla 22. Exactitud, precisión (P), exhaustividad (R) y medida F (F) obtenidas con los corpora mezclados mediante SVM**

Método	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
LDA	65.8	67.6	60.9	64.1	64.4	70.8	67.4
MTD	73.5	73.6	73.5	73.5	73.5	73.6	73.5
<b>LDA + MTD</b>	<b>76.3</b>	<b>76.8</b>	75.6	<b>76.2</b>	<b>75.9</b>	<b>77.1</b>	<b>76.5</b>
LIWC	59.8	60.1	58.4	59.2	59.5	61.1	60.3
LIWC + MTD	73.7	71.9	<b>77.8</b>	74.7	75.8	69.7	72.6
LDA + LIWC	69.7	69.3	70.8	70.1	70.2	68.7	69.4
LIWC + WS + LDA	72.6	72.3	73.5	72.9	73.0	71.9	72.5

**Tabla 23. Exactitud, precisión (P), exhaustividad (R) y medida F (F) obtenidas con los corpora mezclados mediante NB**

Método	Exactitud	Veraz			Engañoso		
		P	R	F	P	R	F
LDA	61.1	61.5	59.3	60.4	60.7	62.8	61.7
MTD	<b>74.3</b>	73.9	75.3	<b>74.6</b>	74.8	73.3	74.1
<b>LDA + MTD</b>	73.5	71.9	<b>77.0</b>	74.4	<b>75.3</b>	69.9	72.5
LIWC	56.4	55.6	64.7	59.8	57.7	48.3	52.6
LIWC + MTD	73.3	73.3	73.3	73.3	73.3	73.3	73.3
LDA + LIWC	64.8	65.1	64.2	64.6	64.7	65.5	65.1
LIWC + WS + LDA	73.3	<b>76.0</b>	68.3	71.9	71.2	<b>78.4</b>	<b>74.6</b>

### 5.1.3 Clasificación de dominio cruzado

A diferencia de la clasificación que combina todos los dominios, para este experimento seleccionamos cada conjunto de datos una vez como conjunto de pruebas y usamos los otros conjuntos de datos restantes como un conjunto combinado de entrenamiento. De esta manera, el dominio del conjunto de prueba no se incluyó en el conjunto de entrenamiento.

**Tabla 24. Medida F obtenida en la clasificación de dominio cruzado. Se muestran resultados para los clasificadores NB y SVM**

		DeRev	OpSpam	Aborto	Mejor amigo	Pena de muerte	Promedio
LDA	SVN	52.1	48.8	57.5	50.0	53.5	50.43
	NB	43.2	49.8	56.0	51.5	56.0	46.50
<b>MTD</b>	<b>SVN</b>	53.3	52.8	55.5	55.5	<b>58.5</b>	53.05
	<b>NB</b>	50.8	<b>53.8</b>	<b>58.5</b>	56.0	48.5	52.30
LIWC	SVN	51.3	49.2	53.1	54.3	52.8	50.25
	NB	50.9	51.1	52.6	56.8	51.4	51.00
LIWC + MTD	SVN	54.6	<b>53.8</b>	54.5	55.0	56.0	54.20
	NB	47.8	52.5	55.5	59.0	51.0	50.15
<b>LDA + MTD</b>	<b>SVN</b>	<b>59.3</b>	50.6	57.5	<b>64.0</b>	55.0	54.95
	NB	58.8	52.3	55.0	59.5	52.5	<b>55.55</b>
LDA + LIWC	SVN	56.3	46.3	54.5	55.5	52.0	51.30
	NB	45.7	48.1	46.5	51.5	49.5	46.90
LDA + LIWC + MTD	SVN	52.1	52.6	57.0	57.0	54.0	52.35
	NB	52.9	53.0	58.0	62.5	53.0	52.95
Mejor:		SVN	NB/SVN	NB	SVN	SVN	NB

La Tabla 24 muestra los resultados de la clasificación inter-dominio. En estos experimentos, a diferencia de los experimentos presentados en las secciones 5.1.1 y 5.1.2, la combinación de características de LDA y MTD no dio consistentemente la mejor exactitud. Mostramos en negrita la mejor precisión obtenida para cada conjunto de datos. En la mayoría de los casos (3 de 5), el SVN superó a NB; sin embargo, con una configuración relativamente simple de una matriz término-documento, NB es capaz de mejorar la detección de engaño con las características aprendidas de otros conjuntos de datos.



#### 5.1.4 Palabras y tópicos predominantes

Detectar mentira en un texto es una tarea que se ha tratado de mejorar mediante distintas técnicas. Dicha tarea no es sencilla debido a que las características que distinguen a un texto engañoso pueden variar entre distintos conjuntos de datos. Algunas palabras que representan el texto falso en un conjunto de datos pueden representar al texto veraz en otro; esto indica que una palabra no es estrictamente representante de una clase u otra en todos los casos. Por ejemplo, se puede observar en la Tabla 27 que el pronombre personal “I” y el adjetivo posesivo “my” son características representantes del texto veraz en el conjunto de datos “Aborto”, en cambio, como se puede ver en la Tabla 25 estas dos palabras funcionan como características representantes del texto engañoso en el conjunto de datos OpSpam.

El hecho de que ciertas palabras sean contradictorias en conjuntos de datos distintos no indica que los métodos automáticos de reconocimiento de texto engañoso sean disfuncionales, sino que estas palabras pueden representar a un conjunto específico de datos, pero no figurarán al tener un conjunto más general, por ejemplo, al unir el conjunto OpSpam con el conjunto “Aborto”.

Se probó que la combinación de características LDA y MTD generó buena precisión en los tres tipos de conjuntos de datos estudiados. Esto implica que los modelos generados mediante esta combinación de características son más fiables para la detección de texto engañoso puesto que han sido probados en distintos casos de engaño y se han mantenido estables en cuanto a la exactitud obtenida. Por otra parte, LDA+MTD puede ser aplicada a otros idiomas debido a que no requiere de algún corpus para generar las características como LIWC o n-gramas sintácticos.

El uso de LDA resultó ser en gran medida un método alternativo más eficiente que el uso de LIWC. Esto se debe a que LDA genera tópicos y toma en cuenta todas las palabras contenidas en los documentos, de tal forma que no habría palabras que escapen del proceso estadístico de generación de grupos de palabras. Por otra parte, LIWC contiene un grupo de palabras preestablecidas de las cuales probablemente algunas no estén incluidas en los documentos a procesar. Por lo

tanto, la generación de un modelo que represente al texto engañoso puede ser más exitoso si se usa LDA.

**Tabla 25. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos OpSpam**

Texto engañoso	Texto veraz
<ul style="list-style-type: none"> <li>▪ my</li> <li>▪ I</li> <li>▪ luxury</li> <li>▪ relax</li> <li>▪ spa</li> <li>▪ visit</li> <li>▪ vacation</li> <li>▪ anyone</li> <li>▪ amaze</li> <li>▪ chicago</li> </ul>	<ul style="list-style-type: none"> <li>▪ location</li> <li>▪ floor</li> <li>▪ block</li> <li>▪ bathroom</li> <li>▪ street</li> <li>▪ large</li> <li>▪ small</li> <li>▪ 2</li> <li>▪ Priceline</li> <li>▪ upgrade</li> </ul>

**Tabla 26. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos "mejores amigos"**

Características engañosas	Características veraces
<ul style="list-style-type: none"> <li>▪ he is guy really up attention overall unintentionally quite</li> <li>▪ this</li> <li>▪ his</li> <li>▪ guy</li> <li>▪ lie</li> <li>▪ nice</li> <li>▪ put</li> <li>▪ trustworthy</li> <li>▪ wonderful</li> <li>▪ how</li> </ul>	<ul style="list-style-type: none"> <li>▪ we</li> <li>▪ each</li> <li>▪ friend</li> <li>▪ have</li> <li>▪ same</li> <li>▪ year</li> <li>▪ my</li> <li>▪ we each other have interests but over one stuff others</li> <li>▪ live</li> <li>▪ good</li> </ul>

Para mostrar como la combinación LDA y MTD se complementan favorablemente, en la Tabla 27 se muestra un ejemplo de las palabras y tópicos más relevantes del texto engañoso y del texto veraz en el conjunto de datos “aborto”. Un punto a destacar de este ejemplo es que los tópicos son características dominantes en la detección de los textos engañosos, es decir, está compuesto de más tópicos que palabras únicas. Mientras que, por otra parte, las palabras son características dominantes en el texto veraz.

**Tabla 27. Los diez tópicos más relevantes y palabras representantes del texto engañoso y veraz en el conjunto de datos "Aborto"**

Texto engañoso	Texto veraz
<ul style="list-style-type: none"> <li>▪ god</li> <li>▪ to of end moment cannot give consequences society perfect return couples chance that</li> <li>▪ is not this womb human piece so there inches carry conception trash</li> <li>▪ chance</li> <li>▪ action</li> <li>▪ necessary</li> <li>▪ we and can this society an may base serve contribute nerves started effect removed his always subject</li> <li>▪ morally</li> <li>▪ they actions believe effectively seriously eyes outcome ask mirror ready</li> <li>▪ who way these perform themselves where example available always out</li> </ul>	<ul style="list-style-type: none"> <li>▪ i</li> <li>▪ but</li> <li>▪ believe</li> <li>▪ choice</li> <li>▪ or</li> <li>▪ me</li> <li>▪ for</li> <li>▪ i feel life me know even people bit ends experience especially unborn control sex mother</li> <li>▪ my</li> <li>▪ without</li> </ul>

Se observó, en los conjuntos de datos analizados, que si una clase carece de características relevantes que acentúen una mayor diferencia entre clases, habrá tópicos que aumenten dichas características y mejoren la exactitud en la clasificación.

En la Tabla 26, Tabla 28, y Tabla 29 se muestran los tópicos y palabras altamente discriminatorias entre la clase de texto engañoso y texto veraz. Los tópicos se pueden identificar por ser un conjunto de palabras independientes, que no deben ser confundidas como una oración.

**Tabla 28. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos "pena de muerte"**

Características engañosas	Características veraces
<ul style="list-style-type: none"> <li>▪ instead_of</li> <li>▪ they wouldn't repetitive capital offenders horrid are have and the</li> <li>▪ him</li> <li>▪ would that cases also upon rape used get someone this give which criminals</li> <li>▪ be an on should arm everybody individual he alive year forgive concern love understand please had</li> <li>▪ it people justice prison first severe shot trials major dictate killed trial</li> <li>▪ lesson</li> <li>▪ possibility</li> <li>▪ the because humanely promotes per population future offenders equivalent didn't murderer shows</li> <li>▪ people as murder that humankind off decide alive instead undone person</li> </ul>	<ul style="list-style-type: none"> <li>▪ I</li> <li>▪ believe</li> <li>▪ I the death believe penalty that am waste sufficient wrongly silly because be</li> <li>▪ put</li> <li>▪ death</li> <li>▪ for</li> <li>▪ opinion</li> <li>▪ sure</li> <li>▪ of</li> <li>▪ current</li> </ul>

### 5.1.5 Comparación de resultados y significancia estadística

La comparación de resultados es importante debido a que demuestra la efectividad del enfoque propuesto, además de mostrar las ventajas y desventajas con respecto a otros enfoques usados. En la Tabla 30 se muestra una comparación de los resultados que obtuvieron diferentes investigadores que usaron los mismos conjuntos de datos con diferentes enfoques. Se puede observar que el enfoque propuesto en el presente estudio obtuvo un mejor rendimiento en la clasificación de la mayoría de conjuntos de datos, excepto para el conjunto de datos OpSpam. Sin

embargo, un mejor rendimiento o mayor exactitud no significa que exista una significancia estadística, es decir, si el incremento en la exactitud es lo suficientemente significativo como para concluir que el enfoque es mejor o simplemente obtiene resultados equivalentes a otros métodos.

**Tabla 29. Los diez tópicos y palabras más relevantes, representantes del texto engañoso y veraz del conjunto de datos "DeRev"**

Características engañosas	Características veraces
<ul style="list-style-type: none"> <li>▪ author</li> <li>▪ your</li> <li>▪ uno/un</li> <li>▪ thriller</li> <li>▪ the and of a to is in that this for s book as it on with an are but</li> <li>▪ and</li> <li>▪ intrigue</li> <li>▪ guide</li> <li>▪ drug</li> <li>▪ issue</li> </ul>	<ul style="list-style-type: none"> <li>▪ the was i they were some which had not during these even when if because character been most several</li> <li>▪ i the it book was t read but reading first all my had this that one so characters love</li> <li>▪ history american native our americans country heart indians west day newspaper the tone brown better known fully resist stolen</li> <li>▪ classic</li> <li>▪ edition</li> <li>▪  </li> <li>▪ the movie book than seen made much movies version like cover film better more do it least half then</li> <li>▪ version</li> <li>▪ history</li> <li>▪ again</li> </ul>

Por tal motivo, se muestra en la Tabla 31 la significancia estadística entre los resultados de esta investigación y los resultados de otras investigaciones. Para fines de comparación, se establece un nivel de significancia ( $\alpha$ ) de 0,05 (5%), lo que significa que la significancia estadística se alcanza si el valor p es menor que  $\alpha$ . Con este nivel de significancia, el rendimiento mostrado por algunos de nuestros resultados no presentaría significancia estadística en comparación con otros métodos existentes, haciéndolos prácticamente equivalentes.

**Tabla 30. Comparación de nuestros resultados con otros estudios de los mismos corpora**

Corpus	Estudios	Exactitud
OpSpam	Este estudio (LDA+MTD)	90.9%
	Mile Ott <i>et al.</i> [29] (LIWC + bigramas)	89.8%
	Song Feng <i>et al.</i> [9] (syntactic rel.+ unigramas)	<b>91.2%</b>
	Donato <i>et al.</i> [13]	90.2%
DeRev	Este estudio (LDA+MTD)	<b>94.9%</b>
	Fornaciari y Poesio [12]	76.27%
Aborto	Este estudio (LDA+MTD)	<b>87.5%</b>
	Perez-Rosas y Mihalcea [32]	80.3%
	Song Feng <i>et al.</i> [9] (syntactic rel. + unigramas)	77.0%
Mejor amigo	Este estudio (LDA+MTD)	<b>87.0%</b>
	Perez-Rosas y Mihalcea [32]	75.9%
	Song Feng <i>et al.</i> [9] (syntactic rel. + unigramas)	85.0%
Pena de muerte	Este estudio (LDA+MTD)	<b>80.0%</b>
	Perez-Rosas y Mihalcea [32]	77.2%
	Song Feng <i>et al.</i> [9] (syntactic rel. + unigramas)	71.5%

**Tabla 31. Significancia estadística**

Corpus	# docs	Este estudio	Otros estudios	Valor-p	S. estadística
OpSpam	800	90.9%	Mile Ott <i>et al.</i> [29] (89.9%)	0.248	No
			Song Feng <i>et al.</i> [9] (91.2%)	0.416	No
			Donato <i>et al.</i> [13] (90.2%)	0.316	No
DeRev	236	94.9%	Fornaciari y Poesio [12] (76.3%)	0.000	Sí
			Perez-Rosas and Mihalcea [32] (80.3%)	0.025	Sí
Aborto	200	87.5%	Song Feng <i>et al.</i> [9] (77.0%)	0.003	Sí
Mejor amigo	200	87.0%	Perez-Rosas and Mihalcea [32] (75.9%)	0.002	Sí
			Song Feng <i>et al.</i> , 2012 (85.0%)	0.283	No
			Perez-Rosas and Mihalcea [32] (77.2%)	0.248	No
Pena de muerte	200	80.0%	Song Feng <i>et al.</i> [9] (71.5%)	0.023	Sí

## **5.2 Impacto de la polaridad de textos en la detección de engaño**

Por lo general, las opiniones engañosas pueden ser positivas o negativas. Las opiniones positivas apuntan a persuadir a los compradores a elegir el producto; mientras que, por otro lado, las opiniones negativas intentan darle mala reputación. Hasta ahora, hay algunos estudios que han tratado de descubrir señales universales de engaño [6]; sin embargo, no existe un conjunto universal de características que puedan identificar el engaño en el texto. Esto sugiere que las señales de engaño podrían variar según la naturaleza de los textos. Por lo anterior, en este apartado se siguen experimentos para capturar características bajo el supuesto de que tomar en cuenta la polaridad del texto puede aumentar la precisión de identificación de texto engañoso.

Las opiniones engañosas se han analizado en diferentes investigaciones, pero generalmente los autores muestran clasificaciones generales. Sin embargo, muchas características podrían descartarse debido a que las opiniones engañosas y veraces pueden ser positivas o negativas. Es decir, las opiniones negativas pueden incluir características particulares que pueden diferenciar solo esa categoría de opiniones, por lo tanto, es posible que la polaridad de los textos afecta directamente las características que ayudan a identificar la verdad o el engaño.

Al analizar diferentes conjuntos de datos, este apartado aborda el impacto de la polaridad en la clasificación del engaño. Se usaron los mismos conjuntos de datos disponibles para realizar las pruebas pertinentes y se realizaron diferentes experimentos con el objetivo de analizar en qué medida ciertas características son específicas de cada polaridad (positiva y negativa).

Asimismo, se implementó un clasificador de polaridad, como un paso de preprocesamiento, para llevar a cabo una clasificación específica. Los resultados muestran que hay señales específicas de engaño que pueden variar según la polaridad del texto. Además, existe una mejora en la mayoría de los conjuntos de datos clasificados cuando se utiliza un clasificador de polaridad.

En los primeros estudios, la combinación de LDA + WSM arrojó los mejores resultados para la detección de engaño general. Sin embargo, debido a que las opiniones falsas suelen ser positivas o negativas, podría haber sido posible que la clasificación dependiera en gran medida de la polaridad de las opiniones. Por ejemplo, un corpus podría consistir en 100 opiniones engañosas y 100 veraces, pero, a su vez, las primeras podrían ser en su mayoría positivas y las posteriores podrían ser en su mayoría negativas. En consecuencia, consideramos la suposición de que el clasificador podría tender a identificar características correspondientes a la polaridad de los textos, en lugar de las que corresponden al engaño en los textos. Para descartar la suposición considerada, se llevó a cabo un experimento simple. Suponiendo que la precisión de la clasificación debería disminuir en el caso de que la presencia de palabras de polaridad estuviera vinculada a la clasificación de opinión, se excluyeron palabras positivas y negativas utilizando un léxico de polaridad [21]. Este lexicon se creó mediante el procesamiento de un conjunto de opiniones sobre ciertos productos, y está en constante crecimiento. En este momento, consta de 2006 palabras positivas y 4783 palabras negativas seleccionadas de opiniones subjetivas. Los autores seleccionaron opiniones subjetivas basadas en el hecho de que la presencia de adjetivos puede determinar una expresión de opinión; además, limitan la selección a aquellas opiniones que se formaron a partir de oraciones relacionadas con la calidad de los productos.

Como puede verse en la Tabla 32, la medida F obtenida después de excluir las palabras de polaridad apenas disminuyó con respecto a la medida F obtenida con los corpora originales. Las palabras en sí mismas podrían ser solo un indicador de la polaridad del texto, pero no un factor determinante para identificar engaño. Por lo tanto, estos resultados muestran que la clasificación no depende de las palabras positivas y negativas como características importantes.

El hecho de que las palabras positivas y negativas no afecten los resultados indica que agregar polaridad como características a los vectores no es la mejor manera de mejorar la detección de engaño. Por lo tanto, basado en el hecho de que existe alguna evidencia que las características del engaño pueden variar según la



naturaleza del texto [20], esperamos mejorar la detección del engaño realizando una clasificación de polaridad antes de capturar las características.

**Tabla 32. Resultados de la clasificación excluyendo palabras de polaridad**

Corpus	P	R	Medida F	Medida F, corpus sin cambios
Aborto	0.871	0.870	0.870	0.875
Mejor amigo	0.870	0.870	0.870	0.870
Pena de muerte	0.819	0.815	0.814	0.800
OpSpam	0.865	0.865	0.865	0.867
DeRev	0.945	0.945	0.945	0.949

### 5.2.1 Clasificación de polaridad para mejorar la detección de engaño

Los resultados de la Sección 5.2 podrían sugerir que el clasificador no toma en cuenta las características relacionadas con la polaridad. Sin embargo, como se ha dicho, hay alguna evidencia que indica que la polaridad puede ser útil para identificar el engaño en el texto. Por lo tanto, aunque las palabras de polaridad no influyen directamente en la clasificación del texto engañoso, tales palabras pueden ayudar indirectamente al clasificar previamente las opiniones positivas y negativas (como un paso de preprocesamiento) y luego clasificar las opiniones como veraces o engañosas.

En este punto, las características generadas serían específicas para cada clase de polaridad. Por estas razones, se realizó una clasificación parcial en lugar de una clasificación general. Es decir, las opiniones se separan primero según su polaridad para mejorar los resultados de clasificación.

Como primera evaluación de esta estrategia, optamos por utilizar el corpus OpSpam dado que los autores ya lo han dividido en opiniones positivas y negativas [28].

La clasificación basada en la polaridad del corpus OpSpam, que se muestra en la Tabla 33, consistió en clasificar opiniones positivas, opiniones negativas y la

combinación de éstas. En este punto, hay dos métodos para dividir el corpus de OpSpam por polaridad: la clasificación de los autores OpSpam (OAP) [28] y la clasificación realizada en este trabajo (PTW). La primera, al menos parte de opinión engañosa, se considera como un estándar de oro. Los datos de OAP muestran los resultados de la clasificación teniendo en cuenta la división de los autores de OpSpam, mientras que los datos de PTW muestran los resultados con el corpus dividido por nuestro clasificador de polaridad.

**Tabla 33. Clasificación del corpus OpSpam, etiquetado por los autores (OAP) y el etiquetado realizado en este estudio (PTW), basado en la polaridad del texto.**

Polarity	# de docs	P	R	F
Positivos <sub>OAP</sub>	800	0.912	0.909	0.910
Negativos <sub>OAP</sub>	800	0.844	0.844	0.844
Positivos <sub>PTW</sub>	620	0.828	0.826	0.826
Negativos <sub>PTW</sub>	980	0.843	0.843	0.843
Pos+Neg	1600	0.867	0.867	0.867

**Tabla 34. Clasificación de los conjuntos de datos agregando característica de polaridad**

Corpus	P	R	F
Aborto	0.871	0.870	0.870
Mejor amigo	0.865	0.865	0.865
Pena de muerte	0.815	0.810	0.809
DeRev	0.928	0.928	0.928
OpSpam	0.865	0.865	0.865

Los últimos resultados muestran que el clasificador de polaridad podría no haber obtenido una clasificación óptima ya que el promedio de la medida F de PTW es menor que OAP. Por lo tanto, si esta suposición es cierta, los resultados de las clasificaciones para los corpora restantes podrían mejorarse utilizando un clasificador de polaridad más preciso.

Para certificar si este efecto se mantiene en los conjuntos de datos restantes, se realizaron experimentos separando textos por polaridad antes de realizar la identificación de engaño. Para este propósito, fue necesario implementar un clasificador de polaridad debido al hecho de que sólo OpSpam está previamente separado por polaridad.

**Tabla 35. Comparación de la medida F obtenida antes (F-B) y después (Avg. F) de la clasificación de polaridad. Las medidas precision (P), Recall (R), medida F (F), y el número de opiniones por clase (# docs) son también mostrados**

Corpus	# docs	Positivas			# docs	Negativas			Avg-F	F-B
		P	R	F		P	R	F		
Aborto	49	0.931	0.918	0.919	151	0.895	0.894	0.894	0.906	0.875
Mejor amigo	121	0.868	0.868	0.868	79	0.863	0.861	0.861	0.864	0.870
Pena de muerte	87	0.885	0.885	0.885	113	0.868	0.867	0.867	0.867	0.800
DeRev	158	0.930	0.930	0.930	78	0.987	0.987	0.987	0.958	0.949
OpSpam	620	0.826	0.826	0.826	980	0.843	0.843	0.843	0.835	0.867

Con el objetivo de construir un clasificador supervisado de opiniones positivas y negativas, seguimos el trabajo de Pang et al. [30]; así se obtuvieron unigramas de un corpus que contiene reseñas sobre películas. Luego, se ordenaron por frecuencia para hacer un diccionario de 16,165 elementos. Para formar el conjunto de entrenamiento, se leyó cada opinión de las películas para hacer un vector que consiste en frecuencias de palabras, es decir, si la palabra existe, se aumenta el conteo de esa característica. Finalmente, se usó un clasificador Bayesiano para procesar vectores. Este método basado en el aprendizaje automático solo incluía información léxica; sin embargo, ha demostrado obtener resultados aceptables. Explorar con otros enfoques para la clasificación de la polaridad se considerará como trabajo futuro.

Aplicamos este clasificador a los conjuntos de datos restantes, de modo que ahora cada uno de ellos está dividido en opiniones positivas y negativas (y cada clase tiene

sus opiniones verdaderas y engañosas, respectivamente). La Tabla 35 muestra la cantidad de opiniones positivas y negativas para cada corpus. Con este experimento se pone a prueba el método propuesto y se comprueba que existen ciertas mejoras en la clasificación.

Para comparar la efectividad entre hacer una clasificación previa de polaridad y simplemente agregar características de polaridad a los vectores, se realizó un segundo experimento donde se agrega solamente información de polaridad. Las características de polaridad se incluyeron en los patrones antes de proceder a clasificar corpora por subconjuntos. La diferencia radica en el hecho de que, en este caso, todo el corpus se toma en cuenta y se agrega una característica más que indica si la opinión es positiva o negativa. Para determinar la polaridad, el clasificador genera un valor real entre 0 y 1; por lo tanto, el valor de la característica se establecerá en 1 si el valor real está por encima de 0,5, de lo contrario se establecerá en 0. Como se puede ver en la Tabla 34, la medida F apenas cambia, en el caso de temas controvertidos (aborto, mejor amigo y pena de muerte) con respecto a los valores que se muestran en la Sección 5.1.1. Por otro lado, la clasificación DeRev y OpSpam mostró que la medida F se redujo en mayor medida. Estos resultados indican que incluir información de polaridad en los patrones (como característica adicional) no es útil e incluso es perjudicial para la detección de engaño.

Todo lo anterior demuestra que obtener características específicas al dividir el texto (en positivo y negativo) antes de buscar señales de engaño (enfoque propuesto para este trabajo, consulte la Tabla 35) arrojó mejores resultados que simplemente agregar información de polaridad como característica adicional (consulte la Tabla 34).

## Conclusiones y trabajo futuro

### *Contenido*

En este capítulo se muestran las conclusiones.



Se ha mostrado en el capítulo de resultados que las características generadas mediante un modelo de espacio semántico continuo han complementado en mayor o menor grado aquellas características generadas mediante otros enfoques. Se experimentó con tres casos de estudio, en donde los mejores resultados se obtuvieron clasificando documentos de un mismo dominio o tema; y también en este caso se obtuvieron los mejores resultados usando LDA. En los casos restantes (dominios mezclados y dominios cruzados) los métodos de generación de características con los mejores resultados fueron variados.

Aunque se ha probado que cuando escribimos un texto para engañar a otra persona, en la mayoría de los casos, tendemos a cambiar la forma en que seleccionamos las palabras, se mostró también que las palabras representantes de cada clase (texto veraz y texto engañoso) no serán estrictamente definitivas de una clase en particular en todos los casos de estudio. Esto podría sugerir que es mejor optar por un método que capture las características representativas de los documentos, antes que un método que busque palabras universales para detectar engaño, o en cambio, usar los algoritmos de aprendizaje automático, aplicarlos a un conjunto de datos enorme, y encontrar un conjunto de palabras o fenómenos lingüísticos que discriminen fuertemente una u otra clase. Sin embargo, etiquetar el texto engañoso y veraz es una tarea retórica, por lo que los estudios, por el momento, se limitan a estudios en un dominio específico.

El presente estudio puede tener un seguimiento enfocado a experimentos más específicos, por ejemplo, se tiene una idea en la cual se le quitan a los conjuntos de datos todas aquellas palabras positivas y negativas.

Usualmente, la mayoría de los trabajos en detección de engaño usan y combinan diferentes métodos para generar características para mejorar la detección de engaño; sin embargo, no tienen en cuenta el hecho de que las características pueden cambiar según la naturaleza del texto. En este trabajo, se llevó a cabo un estudio sobre el efecto de la polaridad sobre el conjunto de características. El método general consistió en implementar un clasificador de polaridad para generar subconjuntos de opiniones positivas y negativas. A continuación, se utilizó un método

semántico (LDA) y léxico (WSM) en los subconjuntos para generar características y construir vectores de entrenamiento. Luego, se implementó la selección de atributos y se formó un clasificador bayesiano con los vectores resultantes. Nuestros experimentos con cinco corpora diferentes muestran que la detección de engaño puede ser favorecida mediante el entrenamiento por separado de subconjuntos positivos y negativos de cada corpus.

Se ha explorado el impacto de la polaridad del texto en la detección del engaño, y se descubrió que efectivamente existe una relación entre ellos. Como trabajo futuro, este estudio abre el camino para realizar nuevos experimentos, no solo considerando textos positivos y negativos sino también explorando otros estados emocionales como la ira, la tristeza, la felicidad, etc. Entonces, una pregunta específica sería en qué medida el texto escrito por una persona que está en un estado emocional podría cambiar las señales de engaño; esto a su vez conduce a abordar el objetivo general de capturar características que podrían convertirse en un conjunto universal de señales de engaño.

La detección de engaño es una tarea compleja debido a que el engaño está relacionado con procesos cognitivos. Por esta razón, algunos enfoques intentan capturar características para mejorar los resultados; sin embargo, estas características parecen cambiar a través de diferentes dominios. Este estudio muestra cómo los conjuntos de datos particionados pueden mejorar los resultados de clasificación. Además, se ha demostrado parcialmente que el enfoque de esta investigación, basado en LDA + WSM, es capaz de encontrar características que son específicas para la detección de engaño, ya que excluir palabras de polaridad no tiene un impacto negativo en la detección del engaño.

## Referencias

- [1] Almela, Á., Valencia-García, R., & Cantos, P. (2012, April). Seeing through deception: A computational approach to deceit detection in written communication.

- In *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 15-22). Association for Computational Linguistics.
- [2] Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). MIT Press.
  - [3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
  - [4] Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3), 113-124.
  - [5] De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC* (Vol. 6, No. 2006, pp. 449-454).
  - [6] DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin*, 129(1), 74.
  - [7] Ekman, P. (1989). Why lies fail and what behaviors betray a lie. In *Credibility assessment* (pp. 71-81). Springer Netherlands.
  - [8] Ekman, P. (2009). *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company.
  - [9] Feng, S., Banerjee, R., & Choi, Y. (2012, July). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 171-175). Association for Computational Linguistics.
  - [10] Fornaciari, T., & Poesio, M. (2012, April). On the use of homogenous sets of subjects in deceptive language analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 39-47). Association for Computational Linguistics.
  - [11] Fornaciari, T., & Poesio, M. (2014, April). Identifying fake Amazon reviews as learning from crowds. In *EACL* (pp. 279-287).
  - [12] Fornaciari, T., & Poesio, M. (2014, April). Identifying fake Amazon reviews as learning from crowds. In *EACL* (pp. 279-287).
  - [13] Fusilier, D. H., Montes-y-Gómez, M., Rosso, P., & Cabrera, R. G. (2015, April). Detection of opinion spam with character n-grams. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 285-294). Springer International Publishing.



- [14] Gokhman, S., Hancock, J., Prabhu, P., Ott, M., & Cardie, C. (2012, April). In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 23-30). Association for Computational Linguistics.
- [15] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.
- [16] H. Mohtasseb and A. Ahmed. 2009. More blogging features for author identification. In: The International Conference on Knowledge Discovery (ICKD'09), Manila.
- [17] Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato).
- [18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10-18.
- [19] Hauch, V., Masip, J., Blandon-Gitlin, I., & Sporer, S. L. (2012, April). Linguistic cues to deception assessed by computer programs: a meta-analysis. In *Proceedings of the workshop on computational approaches to deception detection* (pp. 1-4). Association for Computational Linguistics.
- [20] Hernández-Castañeda, Á., & Calvo, H. (2017). Deceptive text detection using continuous semantic space models. *Intelligent Data Analysis*, 21(3), 679-695.
- [21] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.
- [22] Indurkha, N., & Damerau, F. J. (Eds.). (2010). *Handbook of natural language processing* (Vol. 2). CRC Press.
- [23] Kamp, H. (1981). A theory of truth and semantic representation. *Formal semantics-the essential readings*, 189-222.
- [24] Keila, P. S., & Skillicorn, D. B. (2005, October). Detecting unusual email communication. In *Proceedings of the 2005 conference of the Centre for Advanced Studies on Collaborative research* (pp. 117-125). IBM Press.
- [25] Masip, J., Bethencourt, M., Lucas, G., SEGUNDO, M. S. S., & Herrero, C. (2012). Deception detection from written accounts. *Scandinavian journal of psychology*, 53(2), 103-111.
- [26] Mihalcea, R., & Strapparava, C. (2009, August). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP*

- 2009 Conference Short Papers (pp. 309-312). Association for Computational Linguistics.
- [27] Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5), 665-675.
  - [28] Ott, M., Cardie, C., & Hancock, J. T. (2013, June). Negative Deceptive Opinion Spam. In *HLT-NAACL* (pp. 497-501).
  - [29] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 309-319). Association for Computational Linguistics.
  - [30] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.
  - [31] Pennebaker, J. W., Chung, C. K., Ireland, M. E., Gonzales, A. L., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net
  - [32] Pérez-Rosas, V., & Mihalcea, R. (2014). Cross-cultural Deception Detection. In *ACL* (2) (pp. 440-445).
  - [33] Pérez-Rosas, V., & Mihalcea, R. (2014, November). Gender Differences in Deceivers Writing Style. In *Mexican International Conference on Artificial Intelligence* (pp. 163-174). Springer International Publishing.
  - [34] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C. J., & Burzo, M. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2336-2346).
  - [35] Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4), 409-441.
  - [36] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(Apr), 1297-1322.
  - [37] Rayson, P., Wilson, A., & Leech, G. (2001). Grammatical word class variation within the British National Corpus sampler. *Language and Computers*, 36(1), 295-306.

- [38] Schelleman-Offermans, K., & Merckelbach, H. (2010). Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling*, 7(3), 247-260.
- [39] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3), 853-860.
- [40] Toma, C. L., & Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62(1), 78-97.
- [41] Twitchell, D. P., Nunamaker Jr, J. F., & Burgoon, J. K. (2004, June). Using speech act profiling for deception detection. In *International Conference on Intelligence and Security Informatics* (pp. 403-410). Springer Berlin Heidelberg.
- [42] Williams, S. M., Talwar, V., Lindsay, R. C. L., Bala, N., & Lee, K. (2014). Is the truth in your words? Distinguishing children's deceptive and truthful statements. *Journal of Criminology*, 2014.
- [43] Xu, Q., & Zhao, H. (2012, December). Using Deep Linguistic Features for Finding Deceptive Opinion Spam. In *COLING (Posters)* (pp. 1341-1350).