ISSN 1870-4069

Usando representaciones distribuidas para el análisis de técnicas propagandísticas en noticias falsas

Jennifer Pérez-Santiago, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Laboratorio en Tecnologías del Lenguaje,
México

{jpsantiago, villasen, mmontesg}@inaoep.mx

Resumen. La propaganda es un mecanismo que busca influir o persuadir al lector manipulando su juicio para fomentar una agenda predeterminada. Este tipo de mecanismos, como las falacias lógicas o apelando a las emociones de la audiencia, son comúnmente usados en las noticias falsas. En este artículo se realiza un análisis sobre la identificación automática de las técnicas propagandísticas más comunes usando diferentes métodos de minería de texto. En específico se aplicaron diferentes representaciones de los textos: *n-gramas* de palabras, *n-gramas* de caracteres, así como representaciones distribucionales (i. e., *embeddings*) de sentencias. La evaluación experimental realizada indica que los *embeddings* de sentencias son una representación adecuada para esta tarea. A través del análisis realizado se distinguen las técnicas más complicadas a identificar y se presentan algunas consideraciones específicas sobre el conjunto de datos utilizados.

Palabras claves: Técnicas de propaganda, *embeddings* de sentencias, clasificación de texto.

Using Distributed Representations for the Analysis of Propaganda Techniques in Fake News

Abstract: Propaganda is a mechanism that seeks to influence or persuade the reader manipulating its judgment to foster a predetermined agenda. These types of mechanisms, such as logical fallacies or appealing to the emotions of the audience, are commonly used in fake news. This article analyzes the automatic identification of the most common propaganda techniques using different text mining methods. Specifically, different representations of texts were applied: n-grams of words, n-grams of characters, as well as distributional representations (eg. embeddings) of sentences. The experimental evaluation carried out indicates that the sentence embeddings are an adequate representation for this task. Through the analysis carried out, the most complicated techniques to identify are distinguished and some specific considerations on the data set used are presented.

Keywords: Propaganda techniques, sentence embeddings, text classification.

1. Introducción

La propaganda es una forma de comunicación, un mecanismo que intenta lograr una respuesta que promueva la intención deseada del propagandista. Este mecanismo es usado, en la mayoría de los casos, en noticias extremadamente sesgadas y/o falsas.

Las noticias en los medios de comunicación pueden ser supuestamente neutrales hasta claramente sesgadas. Un artículo de noticia refleja el sesgo no solo del autor sino también del medio donde se publica. El autor puede no estar consciente de este sesgo o podría ser que el artículo es parte de la propia agenda del autor para persuadir sobre algo o un tema en específico [1]. Esta última situación refleja el uso de la propaganda.

Según el "*Institute for Propaganda Analysis*": la propaganda es un método de comunicación cuyo objetivo es dar a conocer una información con la intención de influir en el público para que actúe de una manera determinada o utilice un determinado servicio o producto. Utiliza técnicas psicológicas y retóricas para alcanzar su propósito. Hace uso del lenguaje emocional para inducir a la audiencia a estar de acuerdo con el hablante solo sobre la base del vínculo emocional que se está creando.

En este trabajo se realiza un análisis para determinar las dificultades en la identificación del uso de técnicas propagandísticas en artículos de noticias. Para ello se considera un conjunto de noticias donde se etiquetan fragmentos de texto identificando el tipo de técnica de propaganda usada. En particular, en este conjunto se distinguen 14 técnicas de propaganda. El análisis se realizó al aplicar técnicas tradicionales de clasificación de textos basadas en representaciones como *n-gramas* de palabras y caracteres, así como *embeddings* de sentencias. El análisis también considera el tamaño del contexto usado: (i) considerando exclusivamente el fragmento propagandístico a clasificar, (ii) considerando el contexto vecino al fragmento; y (iii) considerando la oración completa donde se encuentra el fragmento.

Cabe señalar, que estas técnicas tradicionales nos proporcionan modelos interpretables y la posibilidad de explicar lo que se aprende, para así poder identificar los principales problemas, y posibles alternativas de solución. De ahí, que se dejó fuera de este trabajo las técnicas de aprendizaje profundo actualmente en boga.

El resto del documento está organizado de la siguiente forma. Sección 2 es una revisión de los principales trabajos relacionados, la sección 3 introduce el método propuesto y describe el conjunto de datos utilizado para el desarrollo del experimento, el método de evaluación, los diferentes enfoques de procesamiento del texto y la medida de evaluación utilizada. La sección 4 expone los principales resultados, la sección 5 realiza un análisis de los resultados obtenidos y por último la sección 6 presenta las conclusiones y posibles trabajos futuros.

2. Trabajos relacionados

Debido al desarrollo de los medios de comunicación actuales, como por ejemplo la World Wide Web, el uso de las técnicas de propaganda han experimentado un gran auge.

Es por ello la necesidad de investigación encaminada a la identificación y detección automática de estas técnicas propagandísticas en los diferentes artículos de noticias que se publican a diario en la red.

Barrón-Cedeno et al. (2019)[1] presentaron PROPPY, como un sistema de detección de propaganda para noticias en línea y demostraron que los n-gramas de caracteres y otras características de estilo superan a las alternativas existentes para identificar propaganda basada en n-gramas de palabras.

Por su parte, Li, Ye, and Xiao (2019) [2] desarrollaron una herramienta basada en la regresión logística que clasifica automáticamente si una oración es propagandística o no. Para ello utilizaron una combinación de características lingüísticas y semánticas que arrojaron mejores resultados que varios resultados de referencia. A su vez, Da San Martino et al. (2019) propuso una técnica basada en BERT [3] para identificar problemas de propaganda en los artículos de noticias, considerando múltiples niveles de representaciones semánticas.

Como parte de la tarea compartida de Detección de Propaganda de grano fino: NLP4IF 2019 [4] exiten varios trabajos que abordaron la tarea de detectar y clasificar segmentos textuales que corresponden a una de las 18 técnicas de propaganda indicadas en un conjunto de artículos de noticias. Por ejemplo, Alhindi et al. (2019), propusieron una arquitectura que combina embeddings a nivel de caracteres con diferentes tipos de embeddings de palabras como entrada a un modelo BiLSTM-CRF, agregaron además un total de 30 características conceptuales asociados con palabras específicas, como ofensivo, vulgar, grosero o insulto a las representaciones de e*mbeddings* utilizadas. El mejor modelo obtenido fue un BiLSTM-CRF con G*love embeddings* y características codificadas manualmente, con el que obtuvieron un resultado de F1 = 0.13. De sus experimentos concluyeron que existía ruido en el conjunto de datos y propusieron investigar formas de limpiar aún más los datos. Por su parte, el equipo ganador, [6], propusieron realizar una clasificación a nivel de palabra utilizando BERT, obteniendo un desempeño de F1 = 0.2488. Sin embargo no se analizaron las causas de por qué tan baja puntuación, algo que pretendemos realizar en nuestro estudio.

Cabe señalar que la mayoría de los trabajos que se revisaron, particulamente los de la tarea compartida NLP4IF 2019 [4], abordaron las soluciones propuestas mediante el uso de redes neuronales. Las técnicas de aprendizaje profundo han mostrado altos desempeños en varias tareas de procesamiento de textos, sin embargo estás técnicas son débiles en proporcionar modelos interpretables [7]. Es por ello que nos apoyamos en métodos tradicionales de minería de texto como los *n-gramas* de palabras y caracteres y los *embeddings* para explorar y analizar dificultades de la tarea e identificar posibles problemas sobre el conjunto de datos presentado.

3. Diseño experimental

Los experimentos realizados siguen el enfoque de clasificación supervisada, es decir, dado un fragmento de texto identificado como propaganda y el contexto de su documento, el objetivo es identificar la técnica de propaganda usada de entre las 14 técnicas posibles (detalles en sección 3.1). Posteriormente, a partir de los resultados alcanzados, se analizaron los errores de clasificación para concluir sobre las dificultades de la tarea o las inconsistencias en el conjunto de datos dado.

Las representaciones de los textos usadas son: *n-gramas* de palabras con $n = 1,2,3$ y caracteres con $n = 3,4,5$ respectivamente, así como *embeddings* de sentencias pre-entrenados con USE (*Universal Sentence Encoder*) [8] y LASER (*Language-Agnostic Sentence Representations*) [9], ambos, entrenados y optimizados para texto de más de una longitud de palabra, como oraciones, frases o párrafos cortos. Para USE la entrada es un texto en inglés de longitud variable y la salida es un vector de 512 dimensiones, sin embargo con LASER la salida es un vector de tamaño fijo de 1.024 dimensiones para representar la oración de entrada.

Los algoritmos de clasificación empleados son *Support Vector Machines* (SVM), *Linear Support Vector Machines* (LSVM), y Regresión Logística; para los tres clasificadores utilizamos las versiones implementadas en sklearn con los parámetros establecidos por defecto.

3.1 Conjunto de datos

El conjunto de datos utilizado es proporcionado por la tarea compartida SEMEVAL 2020-TASK 11 "Detección de Técnicas de Propaganda en Artículos de Noticias". Los archivos de entrada para la tarea consisten en artículos de noticias en formato de texto libre recopilados de medios de comunicación propagandísticos y no propagandísticos.

El título está en la primera fila, seguido de una fila vacía y el contenido del artículo comienza desde la tercera fila, una oración por línea. Los fragmentos de propaganda fueron detectados y etiquetados manualmente por expertos.

La distribución del conjunto de datos en entrenamiento, validación y prueba se detallan en la tabla 1.

La Fig. 1 muestra la distribución por cada una de las técnicas de propaganda etiquetadas del número total de instancias, para los conjuntos de entrenamiento y validación.

Tabla 1. Distribución de los artículos y fragmentos etiquetados en el conjunto de datos.

Distribución	**Artículos**	**Fragmentos etiquetados (instancias)**
Entrenamiento	371	6129
Validación	75	1063
Prueba	90	1790
Total	**536**	**8982**

Aunque los datos han sido anotados en un principio con 18 técnicas de propaganda, dada la frecuencia relativamente baja de algunas de ellas, los responsables de la tarea decidieron fusionar técnicas similares en una sola superclase. Convirtiendo la tarea en un problema de clasificación de 14 clases.

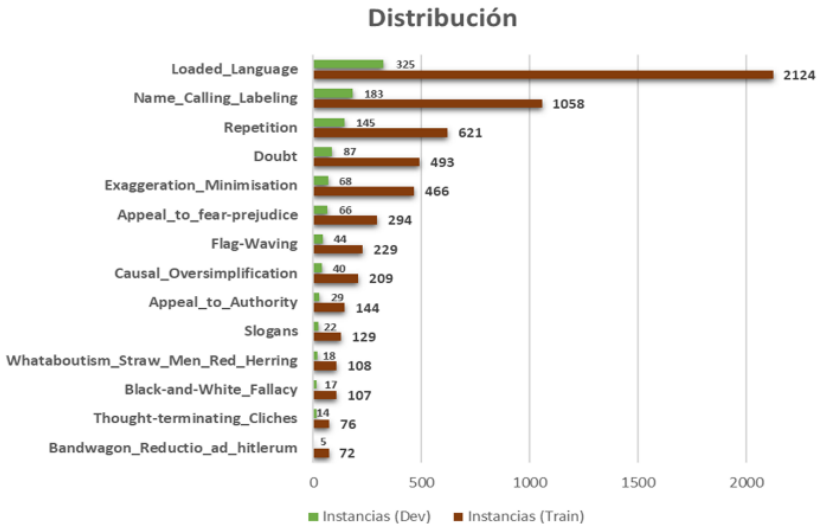

Fig. 1. Estadística sobre el número de instancias por técnica de propaganda para los conjuntos de entrenamiento y validación.

3.2 Técnicas de propaganda

Existen muchos tipos de técnicas de propagandas consideradas en la literatura, en esta sección describimos las 14 técnicas definidas para el desarrollo de esta tarea. La lista solo incluye técnicas que pueden encontrarse en artículos periodísticos y pueden juzgarse intrínsecamente, sin la necesidad de recuperar información de apoyo de recursos externos [3]:

1. **Appeal_to_Authority:** Afirmar que un reclamo es verdadero simplemente porque una autoridad válida o un experto en el tema dijo que era cierto, sin ninguna otra evidencia de respaldo ofrecida. Particularmente poderoso cuando se utilizan referencias religiosas [10]. Ej.: "*Pope Francis explains this again in AL 305 when he says "a small step, in the midst of great human limitations, can be more pleasing to God than a life which appears outwardly in order*"

2. **Appeal_to_fear-prejudice:** Tratar de generar apoyo para una idea, inculcando ansiedad y/o pánico en la población hacia una alternativa, posiblemente basada en juicios preconcebidos. Ej.: "*Ebola victims can be a risk, if mourners have direct contact with the body of the deceased* ".

3. **Bandwagon_Reductio_ad_hitlerum:** Técnica de propaganda que fue fusionada por los responsables de la tarea por dos técnicas diferentes: **Bandwagon:** Intenta persuadir al público objetivo para que se una y tome el curso de acción porque "todos los demás están tomando la misma acción". Está muy relacionada sobre todo con temas políticos [11]. Ej.: "*¿Would you vote for Clinton as president*? *57% say yes*". **Reductio_ad_hitlerum:** Persuadir a una audiencia para que desapruebe una acción o idea sugiriendo que la idea es popular entre grupos odiados por el público objetivo. Puede referirse a cualquier persona o concepto con una connotación negativa [12]. Ej.: "*Only one type of person can think this way: a communist"*. "*Adolf Hitler as a very great man*".

4. **Black-and-White_Fallacy:** Presentar dos opciones, alternativas como las únicas posibilidades, cuando en realidad existen más posibilidades. Como caso extremo, decirle a la audiencia exactamente qué acciones tomar, eliminando cualquier otra opción posible (dictadura). Ej.: "*America will face a choice between accepting a nuclear-armed Iran or acting to destroy as much of this capability as possible the only solution*".

5. **Causal_Oversimplification:** Asume una causa cuando hay múltiples causas detrás de un problema. También incluimos el chivo expiatorio: la transferencia de la culpa a una persona o grupo de personas sin investigar las complejidades de un problema. Ej.: "*the delay signaled the White House was having second thoughts about the nomination"*.

6. **Doubt:** Cuestionando la credibilidad de alguien o algo. Ej.: "*This is evidently of no consequence to a President who cares about nothing about the country and everything about his narrow self-interest*", "*why weren't the police rushing to the scene immediately?*".

7. **Exaggeration_Minimisation:** Representando algo de manera excesiva: haciendo las cosas más grandes, mejores, peores (por ejemplo, "lo mejor de lo mejor", "calidad garantizada") o haciendo que algo parezca menos importante o más pequeño de lo que realmente es (Jowett y O'Donnell, 2012, p. 303). Ej.: " *great big important point*", "*entirely insufficient*".

8. **Flag-Waving:** Jugar con un fuerte sentimiento nacional (o con respecto a un grupo, por ejemplo, raza, género, preferencia política) para justificar o promover una acción o idea. Un truco en el que el propagandista levanta un símbolo, como una bandera, que reconocemos y respetamos [11]. Ej.: "*American people*", "*America First*", "*our democracy*", "*our future*", "*our Republic*", "*our country*".

9. **Loaded_Language:** Usar palabras o frases con fuertes implicaciones emocionales (positivas o negativas) para influir en una audiencia [13]. Ej.: "*hypocritical*", "*filth and degenerate behavior*", "*the nefarious nature of this fascist theocracy*".

10. **Name_Calling_Labeling:** Etiquetar el objeto de la campaña de propaganda como algo que el público objetivo teme, odia, encuentra indeseable o ama o alaba. Apelar al odio o al miedo de una persona para tratar de crear un juicio sobre otra sin evidencia [10]. Truco para hacernos aceptar una conclusión sin tener en cuenta los hechos esenciales del caso [11]. Ej.: "*wanna be statist dictator and naive do-gooder*", "*Muslim Killer*".

11. **Repetition:** Repetir el mismo mensaje una y otra vez, para que la audiencia eventualmente lo acepte. Utilizado por Hitler como una de las principales técnicas fundamentales, si las personas escuchan un mensaje con suficiente frecuencia lo creerán [10]. Ej.: "*autonomy*", se repite alrededor de 25 veces dentro del mismo documento.

12. **Slogans:** Una frase breve y llamativa que puede incluir el etiquetado y los estereotipos. Los lemas tienden a actuar como apelaciones emocionales. Ej .: "*SMASH THE STATE*", "*America First*".

13. **Thought-terminating_Cliches:** Palabras o frases que desalientan el pensamiento crítico y la discusión significativa sobre un tema determinado. Suelen ser oraciones cortas y genéricas que ofrecen respuestas aparentemente simples a preguntas complejas o que distraen la atención de otras líneas de pensamiento [14]. Ej.: "*We all have choices to make*", "*Slowly but surely*", "*too good to be true*".

14. **Whataboutism_Straw_Men_Red_Herring:** Técnica de propaganda que fue fusionada por los responsables de la tarea por tres técnicas diferentes:

    Whataboutism: Desacreditar la posición de un oponente acusándolo de hipocresía sin refutar directamente su argumento. Empleada originalmente por la Unión Soviética y posteriormente continuada por la Rusia post-soviética que busca refutar cualquier crítica

occidental de Rusia por hipocresía [15]. **Straw_Men:** Cuando la propuesta de un oponente se sustituye por una similar que luego se refuta en lugar de la original [16] . Especifica las características de la propuesta sustituida: "caricaturizar una opinión opuesta para que sea fácil de refutar" [13]. **Red_Herring:** Introducir material irrelevante sobre el tema que se está discutiendo, para que la atención de todos se desvíe de los puntos planteados [13].

Los sujetos a un argumento de arenque rojo se alejan del tema que había sido el foco de la discusión y se les insta a seguir una observación o reclamo que pueda estar asociado con el reclamo original, pero que no es muy relevante para el tema en disputa. Ej.: "*don't focus on me, focus on the destructive Radical Islamic Terrorism that is taking place within the United Kingdom*", "*President Trump then reminded everyone of the Obama administration's failures in dealing with Russian meddling in the election*".

3.3 Enfoques de representación de la información de contexto

Para la representación y clasificación de los fragmentos propagandísticos consideramos tres niveles diferentes de información contextual:

- **Enfoque 1, Fragmento:** Solo se usa el fragmento propagandístico definido en los archivos de entrada, sin información de su contexto (i.e. palabras vecinas antes y después del fragmento).
- **Enfoque 2, Contexto:** Se considera el propio fragmento propagandístico y el contexto inmediato a su alrededor conformado desde 1 hasta 3 palabras antes y después de éste.
- **Enfoque 3, Oración:** Se toma en cuenta toda la oración donde se encuentra el fragmento propagandístico a clasificar.

4. Resultados

En las siguientes tablas mostramos los mejores resultados obtenidos para cada uno de los modelos propuestos para los diferentes enfoques de representación del texto propagandístico. El baseline propuesto por la tarea compartida "Detección de Técnicas de Propaganda en Artículos de Noticias, SEMEVAL 2020-TASK 11" usa un clasificador de regresión logística, solo en una característica: la longitud de la oración, obteniendo una medida ***F1 de 0.26529*** para el conjunto de desarrollo.

Para el uso de *n-gramas* de palabras, los mejores resultados fueron representados por los unigramas de palabras, en el caso de los *n-gramas* de caracteres los mejores resultados se obtuvieron para los trigramas de caracteres. (ver Tabla 2). La Tabla 3 ilustra los resultados obtenidos utilizando embeddings de sentencias.

El mejor resultado para todas las evaluaciones fue obtenido mediante embeddings de sentencias (con el modelo USE) usando solo el fragmento propagandístico, y usando Máquinas de Soporte Vectorial Lineal (LSVM). La Tabla 4 muestra las medidas F1 obtenidas para cada una de las técnicas de propagandas etiquetadas con el modelo obtenido con el mejor resultado.

Tabla 2. Resultados con unigramas de palabras y trigramas de caracteres. Se muestran los resultados usando la medida de evaluación micro-F1.

	Clasificador	Oración	Contexto de 3	Contexto de 2	Contexto de 1	Fragmento
Unigrams-palabras	**SVM**	0.309	0.353	0.377	0.406	**0.428**
	LSVM	0.282	0.378	0.407	0.421	**0.457**
	Regresión Logística	0.318	0.381	0.399	0.419	**0.456**
Trigramas-caracteres	**SVM**	0.308	0.375	0.406	0.435	**0.464**
	LSVM	0.276	0.380	0.392	0.435	**0.463**
	Regresión Logística	0.308	0.393	0.415	0.445	**0.459**

Tabla 3. *Embeddings* de sentencias con USE y LASER. Medida de evaluación micro-F1.

	Clasificador	Oración	Contexto de 3	Contexto de 2	Contexto de 1	Fragmento
USE	**SVM**	0.434	0.411	0.434	0.479	**0.517**
	LSVM	0.301	0.423	0.423	0.475	**0.528**
	Regresión Logística	0.296	0.408	0.435	0.475	**0.517**
LASER	**SVM**	0.325	0.458	0.476	0.485	**0.499**
	LSVM	0.326	0.458	0.476	0.492	**0.508**
	Regresión Logística	0.314	0.416	0.435	0.446	**0.475**

Tabla 4. Medidas F1 por clase utilizando *embeddings* de sentencias con USE a nivel de fragmento con LSVM.

Técnicas de Propaganda	F1-LSVM-USE-Fragmento	Datos Entrenamiento	Datos Prueba
Whataboutism_Straw_Men_Red_Herring	0.000	108	18
Black-and-White_Fallacy	0.311	107	17
Flag-Waving	0.000	229	44
Name_Calling_Labeling	0.087	1058	183
Slogans	0.270	129	22
Doubt	0.503	493	87
Thought-terminating_Cliches	0.403	76	14
Appeal_to_Authority	**0.667**	144	29
Exaggeration_Minimisation	**0.673**	466	68
Repetition	**0.574**	621	145
Appeal_to_fear-prejudice	0.284	294	66
Causal_Oversimplification	0.392	209	40
Bandwagon_Reductio_ad_hitlerum	0.000	72	5
Loaded_Language	0.222	2124	325

5. Análisis de resultados

A continuación, se presentan las principales observaciones sobre los distintos aspectos de los modeos elaborados, extraídas a partir del análisis de los errores de clasificación.

Sobre el contexto. Del análisis de los resultados obtenidos lo primero que pudimos apreciar es que para todos los experimentos realizados los mejores valores de F1 se obtuvieron con el enfoque de fragmento, que consiste en observar solo el fragmento propagandístico sin tomar en cuenta su contexto, esto inclusive para fragmentos de una sola palabra. Por otra parte, el uso de toda oración donde se encontraba el fragmento no obtuvo buenos resultados; creemos se debe a que en una misma oración pueden ocurrir más de un fragmento propagandístico para diferentes clases de propaganda.

Sobre la caracterización. Los resultados obtenidos con representaciones tradicionales, n-gramas de palabras y caracteres, fueron en siempre inferiores a los resultados generados por los embeddings de sentencias. De estos últimos USE brindó siempre los mejores resultados, alcanzando un F1 = 0.527 cuando se usó el enfoque de fragmento y el clasificador LVSM.

Sobre la discriminación entre clases. Según los resultados para cada una de las 14 técnicas de propagandas reportadas en la Tabla 4 se puede observar que el modelo tiene problemas distinguiendo prácticamente todas las clases de propaganda. Sin embargo, puede distinguir razonablemente bien clases como: Doubt, Exaggeration and Minimisation, Appeal to Authority, y Repetition. Estas cuatro clases son de las más repetidas o utilizadas en el corpus, así que el tener muchas instancias para el entrenamiento fue un aspecto decisivo.

Sobre la complejidad de las clases. Entre las clases con medida F1 = 0 tenemos el caso de Flag Waving, Whataboutism Straw Men Red Herring y Bandwagon Reductio ad Hitlerum. Para el caso particular de Flag_Waving sucede que su medida F1 es igual a cero para todos los enfoques considerando o no el contexto. Esta técnica de propaganda en particular juega con un fuerte sentimiento nacionalista o de grupo (raza, género, preferencia política) para justificar o promover una acción o idea. Tiende a confundirse con varias clases, pero sobre todo con: Slogans y Doubt.

Slogans, no son más que lemas, frases breves y llamativas que pueden incluir estereotipos y *Doubt*, por su parte es utilizada para cuestionar la credibilidad de alguien o de algo.

Flag_Waving utiliza fragmentos cortos como por ejemplo: *the American People*, *America first*, *our civilization*, *our country*, *our traditions*, our *people*, etc., que tienden a confundirse con fragmentos de *Slogans* como: *America first*, *we will protect our people*, *we will defend our civilization*, *the future is ours and not yours*, *the American people have a* (…).

Para el caso de Doubt, tienden a confundirse con fragmentos como: *the American people are clearly* (…), (…) *prevent it from reaching the American people*, *have the American people been had* (…), (…) *our country and our families*.

Sobre el conjunto de datos. *Whataboutism_Straw_Men_Red_Herring*, es una de las técnicas que es el resultado, como explicamos en la sección 3.1 de la unión de otras

varias técnicas de propagandas, nos referimos a: *Whataboutismo*, que es una técnica que intenta desacreditar la posición de un oponente y constituye la forma más común de evasión de la responsabilidad; *Straw_Men*, es una técnica de propaganda que se propone tergiversar la posición de alguien; y *Red_Herring*, que presenta datos irrelevantes sobre el tema que se está discutiendo para desviar la atención del público y llevar a falsas conclusiones.

Por otra parte, cuando analizamos las instancias del conjunto de entrenamiento correspondiente a esta categoría fusionada observamos fragmentos extremadamente largos que en ocasiones suelen ocupar toda la oración, no siendo así cuando analizamos el conjunto de validación donde los fragmentos son mucho más pequeños, con un tamaño relativo entre 1 a 4 palabras, es por ello que tiende a confundirse con Slogans y Exaggeration_Minimisation.

Bandwagon_Reductio_ad_Hitlerum, constituye el mismo caso de fusión de técnicas de propagandas; *Bandwagon* que consiste en presuadir al público para que se una a una acción o fin común y *Reductio_ad_Hitlerum* que intenta persuadir para que desapruebe una acción o idea, sugiriendo que esta idea es propuesta por grupos, personas o conceptos odiados por el público. La mezcla de ambas técnicas es la de menos instancias tanto para entrenar como para validar en el conjunto de datos, los temas principales de sus fragmentos propagandísticos están encaminados sobre todo a comparaciones o alusiones a la época del fascismo y a la personalidad de Adolfo Hitler. Algunos ejemplos de *Bandwagon_Reductio_ad_Hitlerum* son: *Adolf Hitler as a very great man, to smear Donald Trump as the new Hitler, Hitler exterminating six million Jews, fascist theocracy y Stalinist sense*.

Tiende a confundirse sobre todo con *Slogans*, ya que muchos de sus lemas hacen la misma referencia a esta dictadura y a su principal personaje. Algunos ejemplos de *Slogans*: *long live Hitler*, *to death the Jews*, *Jews are my enemy*.

Loaded_Language, es otra de las técnicas con la que tiende también a confundirse el clasificador en ejemplos como: *Adolf Hitler as a "very great man"*, *the nefarious nature of this "fascist theocracy"*, (…) *most harrowing Stalinist sense*.

Como podemos ver, esta última técnica, *Loaded_Language*, junto con otras como *Slogans* y *Name_Calling_Labeling*, utilizan mucho el contenido enfático, es decir, muchas frases en mayúsculas o entre comillas, y esta es una característica sobre la que no se trabajó y pudiera ser de mucha ayuda en el proceso de clasificación.

6. Conclusiones y trabajo futuro

Para el problema de identificación automática de técnicas de propaganda a nivel de fragmento dado el contexto del documento, experimentamos con varios modelos. Se consideraron diferentes contextos y diferentes caracterizaciones. El modelo de mejor rendimiento obtenido contempla el uso de los embeddings de sentencias de USE a nivel de solo el fragmento propagandístico y con Máquinas de Soporte Vectorial Lineal, logrando un puntaje F1 de 0.528 que supera la línea base propuesta que fue de 0.2653 utilizando Regresión Logística.

A partir del análisis realizado es importante recomponer el corpus equilibrando el número de instancias por clases, o al menos tener una mejor distribución de las 18 técnicas de propagandas con cantidad de instancias suficientes.

Además, es necesario evitar unir clases para evitar ambigüedades. Es necesario evitar oraciones que reúnen varios fragmentos de diferentes técnicas propagandísticas, al menos, para una primera fase del modelo de clasificación. También se descubrió que el tamaño de los fragmentos está en relación a la técnica utilizada y que algunas utilizan marcadores enfáticos, por tanto sería de interés futuro estudiar otras características como: longitud (sentencia, fragmento), uso de contenido enfático (frases en mayúsculas o entre "comillas") y la incorporación del uso de emociones que se pudieran generar de las técnicas usadas.

Agradecimientos. El primer autor agradece el apoyo otorgado por el CONACYT a través de la beca No. 1017548. Los autores también agradecen el apoyo recibido por el CONACYT a través del proyecto de CB-2015-01-257383 para la realización de esta investigación.

Referencias

1. Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., Nakov, P.: Proppy: organizing the news based on their propagandistic content. Inf. Process. Manag., 56(5), pp. 1849–1864 (2019)
2. Li, J., Ye, Z., Xiao, L.: Detection of propaganda using logistic regression. In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pp. 119–124 (2019)
3. da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P.: Fine-Grained Analysis of Propaganda in News Articles. In: (EMNLP-IJCNLP) Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, pp. 5636–5646 (2020)
4. da San Martino, G., Barrón-Cedeño, A., Nakov, P.: Findings of the (NLP4IF) shared task on fine-grained propaganda detection. In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pp. 162–170 (2019)
5. Alhindi, T., Pfeiffer, J., Muresan, S.: Fine-tuned neural models for propaganda detection at the sentence and fragment levels. pp. 98–102 (2019)
6. Yoosuf, S., Yang, Y.: Fine-grained propaganda detection with fine-tuned BERT. In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pp. 87–91 (2019)
7. Zhou, X., Zafarani, R.: Fake news: a survey of research, detection methods, and opportunities. http://arxiv.org/abs/1812.00315 (2018)
8. Cer, D. et al.: Universal sentence encoder for English. In: EMNLP, Conf. Empir. Methods Nat. Lang. Process. Syst. Demonstr. Proc., pp. 169–174 (2018)
9. Schwenk, H., Douze, M.: Learning joint multilingual sentence representations with neural machine translation. pp. 157–167 (2017)
10. Torok, R.: Symbiotic radicalisation strategies: propaganda tools and neuro linguistic programming. pp. 58–65 (2015)
11. Hobbs, R.: Teaching about propaganda: an examination of the historical roots of media literacy. J. Media Lit. Educ., 6(2), pp. 56–67 (2016)
12. Teninbaum, G.H.: Reductio ad hitlerum: trumping the judicial nazi card. Suffolk Univ. Law Sch. Res. Pap., 09-37(3), pp. 541–78 (2009)
13. Anderson, S.L.: A rulebook for arguments. 10(2) Hackett Publishing (1987)
14. Hunter, J.: Brainwashing in a large group awareness training ? The classical conditioning hypothesis of brainwashing. University of KwaZulu-Natal, Pietermaritzburg (2015)

15. Richter, M.L.: The Kremlin's platform for 'useful idiots' in the west: an overview of RT's editorial strategy and evidence of impact. Kremlin Watch, pp. 53 (2017)
16. Macagno, F., Walton, D., Pragmatics, T. : Interpreting Straw Man. 14, Springer (2017)