Interpretabilidad categórica de clasificadores automáticos sobre contenido relacionado a la percepción de la seguridad

Andrés Julián Bermúdez García

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Matemáticas
Bogotá, Colombia
2022

Interpretabilidad categórica de clasificadores automáticos sobre contenido relacionado a la percepción de la seguridad

Andrés Julián Bermúdez García

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magíster en Ciencias - Matemática Aplicada

Director(a):
Francisco Albeiro Gómez Jaramillo
Ph.D. en Ingeniería - Sistemas y Computación

Línea de Investigación:
Interpretabilidad en aprendizaje automático
Grupo de Investigación:
Computational Modeling of Biological Systems Research Group - COMBIOS

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Matemáticas
Bogotá, Colombia
2022

A mi hermano Deybid (Q.E.P.D). Lamento no poder ayudarte a tiempo, ahora es mi turno de alejar a los monstruos.

Agradecimientos

En primer lugar quisiera agradecer a la profesora Luisa Fernanda Chaparro quien direccionó en gran medida la línea de investigación realizada, cuyos avances en la misma permitieron la ejecución de esta investigación además de su aporte en la idea base de la medida de contribución planteada. A su vez al director de este trabajo, el profesor Francisco Gómez por su colaboración en el análisis, interpretación, delimitación del alcance de este proyecto y financiación de esta maestría .

A mi colega y amigo MSc. Cristian Pulido por su infinita ayuda en la ejecución de este proyecto y sus grandes aportes en el planteamiento y definición de la medida de interpretabilidad y el modelo de comparación base. También agradezco a mi colega y amigo MSc Juan Leal por incentivarme y apoyarme en el transcurso de esta maestría, sin el apoyo de ambos no hubiese sido posible.

Finalmente a mi esposa, mi padre, mi madre y mis hermanos, por ser mi motivación.

Resumen

Título en español: Interpretabilidad categórica de clasificadores automáticos sobre contenido relacionado a la percepción de la seguridad

La percepción de la seguridad está relacionada con los sentimientos de los ciudadanos ante el riesgo asociado a los sucesos de seguridad y la magnitud de sus consecuencias. Debido a esta naturaleza subjetiva, es un tema complejo de cuantificar. Por ello, las redes sociales surgieron como una alternativa para cuantificar estas opiniones. Recientemente, los métodos de aprendizaje automático supervisado multiclase cuantificaron distintos niveles de percepción de la seguridad, sin embargo, estos métodos carecen de interpretabilidad sobre por qué un grupo de tweets clasifica en el mismo nivel de percepción de seguridad. Este trabajo propone una novedosa estrategia de interpretabilidad categórica y selección agnóstica al modelo para un grupo de predicciones relacionadas con el mismo nivel de percepción de la seguridad. Los resultados sugieren que el modelo propuesto presenta altos niveles de interpretabilidad para las diferentes categorías de PoS. Adicionalmente, las métricas de interpretabilidad introducidas mejoran el proceso de selección de los modelos.

Palabras clave: Percepción de Seguridad (PoS), Interpretabilidad Local y Categórica, Procesamiento de Lenguaje Natural (NLP), LIME.

Abstract

Título en inglés: Categorical interpretability of automatic classifiers on content related to the perception of security

The perception of security relates to citizens' feelings in the face of risk associated with security events and the magnitude of its consequences. Because of this subjective nature, it is a complex subject to quantify. Therefore, social networks emerged as an alternative to quantifying these opinions. Recently, multiclass supervised machine learning methods quantified different levels of security perception. However, these methods lack interpretability about why a group of tweets classifies in the same level of perception of security. This work proposes a novel strategy of categorical interpretability and model-agnostic selection for a group of predictions related to the same level of perception of security. The results suggest that the proposed model presents high levels of interpretability for the different PoS categories. Additionally, the introduced interpretability metrics improve the model selection process.

Keywords: Perception of Security (PoS), Local and Categorical interpretability, Natural Language Processing (NPL), LIME

Contenido

Lista de Figuras

1. Introducción

1.1. Percepción de Seguridad

La percepción de seguridad (PoS- del inglés *Perception of Security*) se relaciona a la medición de la respuesta subjetiva de los ciudadanos frente al riesgo relacionado con los eventos de seguridad y la magnitud de sus consecuencias [Rundmo and Moen, 2006]. La percepción de seguridad está altamente relacionada con el tipo de sentimiento descrito como miedo al crimen [Brown et al., 2021, Pulido et al., 2019], esto es, la respuesta emocional a la que se enfrentan las personas cuando son víctimas de un delito o cuando estas notan un cambio negativo en la noción de seguridad de su propio entorno trayendo consigo consecuencias graves a la sociedad. La percepción de la seguridad cambia en el tiempo y en el espacio [Yadav and Sheoran, 2018], dependiendo de las circunstancias individuales y las experiencias que sufren las personas. Dada su naturaleza individual y subjetiva, cuantificar tal percepción propone un gran reto. Motivo por el cual, las encuestas sobre la opinión de los ciudadanos representan la alternativa más utilizada para cuantificar este sentimiento [Brooker and Schaefer, 2015].

En particular, para la ciudad de Bogotá D.C en Colombia, la Secretaría de Seguridad, Convivencia y Justicia posee mecanismos que proveen esta percepción (desde el año 2014) tales como la *Encuesta de Percepción y Victimización de la Cámara De Comercio De Bo-*

gotá [Cámara de Comercio de Bogotá, 2022].

Los resultados, recomendaciones y conclusiones de este instrumento sirven como un insumo para que la Administración Distrital defina y ajuste sus políticas públicas en torno a la seguridad en la cuidad [Cámara de Comercio de Bogotá, 2022]. Sin embargo, estas encuestas no se adaptan bien a la naturaleza dinámica de la percepción de seguridad y suelen centrarse únicamente en la cuantificación de los niveles de victimización [Brooker and Schaefer, 2015].

Para resolver tal limitante, recientemente las redes sociales se han propuesto como una alternativa para cuantificar estas opiniones [Schultz-Jones, 2009]. Estas redes permiten transmitir en tiempo real acontecimientos y noticias relacionadas con diferentes ámbitos, incluyendo las opiniones de los ciudadanos sobre la seguridad [Brown et al., 2021]. Además, las redes sociales difunden rápidamente estos contenidos [Java et al., 2007], lo que las convierte en fuentes valiosas para observar datos dinámicos que pueden ayudar a entender cómo cambia el PoS de las personas a lo largo del tiempo. Basándose en esta observación, diferentes trabajos han utilizado el contenido de Twitter para caracterizar cuantitativamente la PoS [Schultz-Jones, 2009, Chaparro et al., 2021a, Chaparro et al., 2020, Victorino et al., 2020].

Con el propósito de estimar y cuantificar la percepción de seguridad, se ha propuesto el uso de clasificadores basados en aprendizaje supervisado que toman como entrada el texto contenido de Twitter y cuantifican el nivel de percepción asociada a estos textos. Como complemento a estas propuestas, este trabajo se centrará en el contenido de Twitter. Esta red social no solo provee contenidos de microblogging sino que también genera una gran cantidad de datos debido que cuenta con más de 150 millones de usuarios activos diarios [Java et al., 2007]. Los contenidos de Twitter podrían explicar el tipo de respuesta que una comunidad frente a determinado acontecimiento y el cómo las personas reciben, procesan e interpretan la información procedente de los contenidos de los medios de comunicación. Los tweets son activos valiosos para entender las percepciones de la gente , en particular, para

caracterizar la percepción ciudadana de la seguridad [Brown et al., 2021].

Al evaluar el rendimiento de los clasificadores de aprendizaje automático supervisado para la clasificación de diferentes niveles de sentimiento en textos, surge la necesidad de complementar las métricas habituales debido a su similitud en los resultados reportados [Chaparro et al., 2020, Chaparro et al., 2021b]. Para comprender las razones y características relevantes en la clasificación de conjuntos de tweets, se han utilizado técnicas de interpretación local, como LIME (Local Interpretable Model-Agnostic Explanations) [Ribeiro et al., 2016a, Chaparro et al., 2021a]. Sin embargo, aún no se ha logrado una interpretación satisfactoria que explique *por qué un grupo de tweets se clasifica en el mismo nivel de percepción de la seguridad.*

En este trabajo se propone una nueva estrategia de interpretabilidad categórica independiente al modelo para un conjunto de predicciones relacionadas con el mismo nivel de percepción de la seguridad. Para ello, en primer lugar, se construye una explicación de las clasificaciones individuales basada en LIME. Con esto se extiende la noción de interpretabilidad local a partir de una representación que permita establecer la contribución global de las características presentes en las diferentes categorías de PoS. Adicionalmente se propone una medida de interpretabilidad independiente del modelo. Finalmente, para cada categoría de PoS, las contribuciones permiten cuantificar el rol de ciertas palabras tales categorías.

1.2. Análisis de sentimiento en la percepción de seguridad

Actualmente, el análisis procedente de las redes sociales y el impacto que los contenidos tienen en su difusión están altamente relacionados con el tono en que se generaron los tweets. Por ello, es necesario analizar el tono del contenido, lo que es denominado por los expertos como análisis de sentimiento [Prathap and Ramesha, 2018].

Los métodos de aprendizaje automático supervisado multiclase cuantifican los niveles de percepción de la seguridad basándose en las publicaciones de Twitter [Chaparro et al., 2021b]. Estos métodos pretenden explicar los diferentes niveles de sentimiento (1: muy negativo, 2: negativo, 3 neutro, 4: positivo y 5: muy positivo) relacionados con a la percepción de subyacente en las publicaciones de Twitter.

1.3. Interpretabilidad local en la percepción de seguridad

El desarrollo continuo de los algoritmos de aprendizaje automático ha conllevado recientemente al desarrollo de técnicas que, a nivel humano, permitan distinguir las decisiones y motivaciones por las cuales estos algoritmos realizan las tareas para los que fueron diseñados [Ribeiro et al., 2016a, Molnar, 2022]. Aunque su definición no es matemáticamente precisa, podemos entender la interpretabilidad como el grado en que un ser humano puede predecir consistentemente el resultado del modelo [Kim et al., 2016]. Un modelo se interpreta mejor que otro si sus decisiones son más fáciles de comprender para un ser humano que las decisiones del otro modelo [Miller, 2019]. Cuanto mayor sea la interpretabilidad de un modelo de aprendizaje automático, más fácil será para alguien comprender por qué se han tomado ciertas decisiones o predicciones [Molnar, 2022].

Entre las técnicas desarrolladas para interpretar modelos de aprendizaje automático dos de estas destacan, los modelos sustitutos locales conocidos como LIME y *SHapley Additive exPlanations*(SHAP)[Messalas et al., 2019], ambas operan sobre predicciones individuales y realizan procesos de selección y relevancia de características [Molnar, 2022, Messalas et al., 2019].

Estas técnicas de interpretabilidad local han explorado las características de mayor relevancia en las categorías de PoS a nivel de predicción individual [Chaparro et al., 2021a], es decir, de un sólo tweet. En concreto, una explicación agnóstica del modelo local interpreta-

ble, como LIME, proporcionó interpretaciones de mensajes individuales para clasificadores de aprendizaje automático. Sin embargo, no existe una interpretación del *¿Por qué un grupo de tweets se clasifica en el mismo nivel de percepción de la seguridad?*, es decir, *no existe una interpretabilidad categórica.*

1.4. Planteamiento del problema

Definido el contexto de este trabajo, se plantea el problema de investigación relacionado a la extensión de las nociones de interpretabilidad local agnóstica a nivel de categorías de percepción de seguridad PoS. Asociado a esta problemática se plantean las siguientes preguntas de investigación: ¿Cuál métrica de desempeño permitiría establecer la elección del mejor modelo de clasificación que aporte a la mejoría de la interpretabilidad de los elementos presentes en el conjunto de datos? y ¿es posible establecer una interpretación global de las características presentes en el conjunto de tweets, que representen los factores relevantes de percepción de seguridad?

1.5. Objetivos

Con el objetivo de abordar las preguntas de investigación planteadas, se plantean los siguientes objetivos:

1.5.1. Objetivo general

Establecer una estrategia computacional de interpretabilidad categórica agnóstica del modelo de aprendizaje automático para la cuantificación de niveles de percepción de seguridad.

1.5.2. Objetivos específicos

- Formular un modelo matemático para la estrategia computacional de la interpretabilidad categórica agnóstica sobre datos de texto provenientes de redes sociales.
- Implementar el modelo matemático propuesto basado en información de interpretaciones locales.
- Evaluar el modelo compuesto utilizando datos provenientes de diferentes niveles de percepción de seguridad procedentes de la red social Twitter.

1.6. Contribuciones

Como resultado de este trabajo se realizaron las siguientes contribuciones:

- Una nueva estrategia de interpretabilidad categórica que considera una diferente estructura de caracterización de los datos agnóstica a los modelos de clasificación utilizados.
- Participación en calidad de Ponente en la Segunda Conferencia Colombiana de Matemáticas Aplicadas e Industriales (MAPI 2), que se llevó a cabo en Medellín–Colombia de la ponencia titulada *Interpretability on categories of perception of security.*

2. Antecedentes

2.1. Minería de Texto y Procesamiento de Lenguaje Natural

La minería de texto es un conjunto de técnicas que se utilizan para procesar y explorar grandes cantidades de datos. También podría describirse como el descubrimiento de nueva información, que antes no se conocía, mediante la extracción automática de información de recursos escritos [Zhang et al., 2015]. Actualmente el área considera técnicas que abarcan el aprendizaje automático, la recuperación de la información, el procesamiento del lenguaje natural (NLP-del inglés Natural Language Processing), entre otros [Ghosh et al., 2012].

Se estima que cerca del 80 % de la información se almacena por medio de texto, por lo cual la minería de texto tiene un gran potencial para generar valor agregado para el análisis de información [Ghosh et al., 2012].

Las técnicas de NLP se han utilizado en diferentes problemas, incluyendo, traducciones automáticas, reconocimiento de enfermedades, reconocimiento de discurso, generación de texto y análisis de sentimientos, entre otros [Hirschberg and Manning, 2015, Luque et al., 2019].

Una rama de NLP es la extracción de información, ésta ha sido desarrollada para aislar información de fuentes de texto no estructurado, no obstante, la mayoría de sus aplicaciones

son para idioma inglés [Hirschberg and Manning, 2015].

2.1.1. Preprocesamiento

El aumento del tamaño de las colecciones de datos e información en los últimos años ha hecho necesario el desarrollo de herramientas que faciliten el acceso a la información. La recuperación de información se centra en facilitar a los usuarios el acceso a la información que necesitan [Balakrishnan and Lloyd-Yemoh, 2014]. Aquí no sólo busca la información correcta, sino que la represente de forma fácilmente comprensible para los usuarios. [Chowdhury and Chowdhury, 2003].

Debido a la naturaleza de los datos, textos en lenguaje natural no estructurados, es necesario realizar un conjunto de tareas que faciliten una representación apropiada [Chowdhary, 2020]. Entre las tareas más usuales en el preprocesamiento se encuentran:

Estandarización de Formato

La estandarización de formato es una labor que agrupa acciones como dejar todo el texto en minúsculas, retirar caracteres especiales (por ejemplo # @) y marcas diacríticas [Bokinsky et al., 2013]. De esta manera palabras iguales pero con acentos diferentes se identifican con la misma cadena de caracteres con el formato homogeneizado.

Eliminación de signos de puntuación y palabras vacías:

Mediante esta tarea se remueven las palabras frecuentes que no aportan información para caracterizar los textos, éstas se denotan como palabras vacías (o de parada) y se eliminan los signos de puntuación, los cuales en vez de enriquecer la representación pueden generar ruido innecesario [Ladani and Desai, 2020]. Las palabras vacías corresponden a palabras altamente usadas en el idioma trabajado, usualmente incluye conjunciones, artículos, pronombres, preposiciones, entre otras. [Kaur and Buttar, 2018].

Lematización y Stemming

El stemming es un método para reducir todas las formas flexionadas[1] [Agirre et al., 1992] de palabras a su «raíz» o «tallo» (stem, en inglés), cuando estas comparten una misma raíz. Por ejemplo, las palabras limpieza, limpiar y limpio tienen todas la misma raíz: «limp». Las palabras raíz que se obtienen al aplicar stemming no necesariamente existen por sí solas como palabra. Aplicar stemming a textos puede simplificarlos, al unificar palabras que comparten la misma raíz, y evitando así tener un vocabulario más grande de lo necesario. A diferencia de la lematización, en donde cada lema es una palabra que existe en el vocabulario del lenguaje correspondiente y se realiza un proceso por medio del cual se extrae la forma canónica «lema» de las formas flexionadas presentes en el texto individualmente. La lematización debe considerar la intención con la que se utiliza la palabra y asignar un representante por convención.

2.1.2. Representación

Posterior al preprocesamiento, en NLP generalmente se recurren a herramientas de representación numérica del texto. A continuación, se describen algunas de estas.

Bolsa de Palabras-Bag of Words

Bag of Words (BoW) es un método clásico para la representación numérica del texto, es una herramienta que se utiliza regularmente en aplicaciones de NLP y en particular en aplicaciones relacionadas con clasificación de texto. Permite generar una versión numérica en un vector de longitud fija [Soumya George and Joseph, 2014].

Inicialmente se individualiza cada elemento del texto correspondiente, conocidos como «tokens», fragmentando al mismo en cada uno de los espacios entre palabras. Luego, se es-

[1] La flexión es la alteración que experimentan las palabras mediante morfemas constituyentes según el significado gramatical o categórico para expresar sus distintas funciones dentro de la oración y sus relaciones de dependencia o de concordancia con otras palabras o elementos oracionales.

tructura un vector cuya longitud es igual a la cantidad de «tokens» presentes en todo el conjunto de datos. Cada una de las entradas debe corresponder a un token de manera fija y para la representación del texto las entradas deben almacenar la frecuencia del token en el texto correspondiente. En otras palabras, este método consiste en transformar textos en vectores que almacenan la frecuencia de las palabras. Se construye un diccionario con todas las palabras utilizadas emparejadas con su frecuencia de ocurrencia en el texto que se quiere representar [Le and Mikolov, 2014]. Para la representación de múltiples documentos se realiza la representación en vectores que mantienen el mismo orden de correspondencia entre entradas y palabras, y se construyen matrices donde cada fila corresponde a un documento. Siguiendo el anterior proceso es posible representar el conjunto de datos como una matriz donde cada fila corresponde a una descripción de un documento analizado y cada columna esta asociada con un token. Esta técnica no es muy eficiente si se aplica al texto original, debido a que procesa las conjugaciones y errores ortográficos como diferentes palabras, adicionalmente le puede dar un gran peso a palabras que no son relevantes para la representación del texto (las denominadas «palabras vacías») debido que estas suelen ser usadas continuamente en los escritos [Soumya George and Joseph, 2014, Le and Mikolov, 2014].

TF-IDF

Term frequency-inverse document frequency (TF-IDF) es una variante de BoW en la que las entradas de los vectores están asociadas con la relevancia de las palabras en los textos. La relevancia de la palabra t en el documento d se cuantifica como $tf - idf(t, d)$ y es la frecuencia de los términos multiplicada por el inverso de la frecuencia de los documentos.

$$tf - idf(t, d) = tf(t, d) \cdot idf(t, d) \tag{2-1}$$

Donde

$$idf(t, d) = log\left(\frac{1 + n}{1 + df(t)} + 1\right) \tag{2-2}$$

Siendo n la cantidad de documentos en todo el conjunto y $df(t)$ la cantidad de documentos que contienen la palabra t. De esta manera las entradas con mayor peso corresponden a los tokens que aparecen más veces en el documento y menos veces en los otros registros, es decir que diferencian al texto del resto de documentos del conjunto de datos. Adicionalmente, se divide cada vector por su norma.

Con el propósito de mitigar aquellas complicaciones que puedan presentarse en las vectorizaciones mencionado es necesario llevar a cabo actividades de preprocesamiento del texto, tales como lematización y stemming antes mencionados, corrección ortográfica semiautomática, remoción de palabras vacías, puntuación y caracteres especiales.

Con el fin de mitigar las complicaciones asociadas a las vectorizaciones mencionadas, es necesario realizar actividades de preprocesamiento del texto, como la lematización, el stemming, la corrección ortográfica semiautomática, la eliminación de palabras vacías y la eliminación de puntuación y caracteres especiales.

La representación obtenida mediante el método TF-IDF se caracteriza por tener un subconjunto pequeño de palabras en los textos, lo que resulta en vectores dispersos con la mayoría de las entradas igual a cero. Aunque este método es intuitivo, fácil de interpretar y puede obtener buenos resultados en algunas aplicaciones, presenta desventajas importantes, como la pérdida del orden de las palabras, la falta de consideración de la semántica y la generación de vectores de alta dimensionalidad [Le and Mikolov, 2014]. Además, requiere un preprocesamiento exhaustivo del texto y, si se implementa en datos que contienen tokens no presentes en el corpus inicial, es necesario volver a representar los datos.

2.2. Aprendizaje automático relacionado a Seguridad

El uso de aprendizaje automático y redes sociales para cuantificar los sentimientos de las sociedades a temáticas relacionadas con seguridad no es nuevo. Por ejemplo, Wang et al. [Wang et al., 2012] y Gerber [Gerber, 2014] utilizaron el contenido de Twitter para mejorar la capacidad de predicción de la delincuencia complementando los modelos de predicción con información espacial de los temas que surgen de los mensajes de Twitter. Para ello, utilizaron el análisis semántico latente para caracterizar los temas que aparecen en Twitter. Estos enfoques consideran el conjunto completo de Tweets, lo que aumenta el riesgo de incluir contenido que no está necesariamente relacionado con la delincuencia y la seguridad. Además, estos análisis no tienen en cuenta el sentimiento (positivo o negativo) implícito en cada de cada mensaje. Recientemente, Chen et al. [Chen et al., 2015] incluyeron una puntuación de análisis del sentimiento en un modelo de predicción de la delincuencia, lo que ha dado lugar a en una mejora de la capacidad de predicción. Este enfoque se basa en un modelo de lexicón de uso general, que puede no reflejar percepciones y polaridades para ámbitos concretos, como la seguridad [Yadav and Sarkar, 2018]. Además, este enfoque también se centra en el contenido de Twitter, lo que aumenta el riesgo de incluir contenido no relacionado con la seguridad.

Otros trabajos se centran en el estudio de la relación entre los sucesos delictivos y el contenido de Twitter. Por ejemplo, en [Bendler et al., 2014] estudiaron la conexión entre el número de Tweets publicados en diferentes zonas de la ciudad y la ocurrencia de diferentes tipos de delitos. En [Malleson and Andresen, 2015] utilizaron el contenido publicado en Twitter como estimador de la cantidad de población en riesgo de ser de ser víctima de un delito. Ambos enfoques se centraron de nuevo en el contenido general de contenido de Twitter y también ignoraron el contenido y la polaridad de los mensajes. También, en [Cvetojevic and Hochmair, 2018] estudiaron la difusión de Tweets en respuesta a ataques cri-

minales. Ellos utilizaron un conjunto de palabras clave para filtrar el contenido relacionado con el evento delictivo de interés. Sin embargo, este enfoque se centró en un solo suceso delictivo y no tuvo en cuenta la percepción de seguridad implícita en las publicaciones de Twitter. En trabajos relacionados, [Kounadi et al., 2015] estudiaron la percepción de la gente sobre los homicidios. Para ello, filtraron Tweets que enlazaban con páginas web relacionadas con homicidios. Este enfoque requería la identificación previa de estas páginas, pero no tuvo en cuenta la percepción de la seguridad. En resumen, la comprensión de cómo la percepción de seguridad en los tweets está todavía limitada [Gaisbauer et al., 2021, Rutjens and Brandt, 2019].

Evidencia reciente sugiere que los mensajes de Twitter reflejan el miedo de la delincuencia en los países latinoamericanos para una ventana de observación de unos dos meses [Prieto Curiel et al., 2020]. Sin embargo, la percepción de seguridad incluye otros factores subjetivos más allá del miedo a la delincuencia, incluida la confianza en la policía y los signos de trastornos sociales y físicos [Hummelsheim et al., 2011, Drakulich, 2015]. Además, las percepciones pueden tener también una naturaleza dinámica que varía en tiempo y el espacio [Curiel and Bishop, 2017]. Por lo tanto, la comprensión de las percepciones de la seguridad puede requerir desarrollos adicionales para tener en cuenta de estos factores subjetivos y la naturaleza dinámica de estas opiniones [Curiel and Bishop, 2017, Camargo et al., 2016].

Además, la posible relación entre las percepciones de seguridad procedentes de las redes sociales y la delincuencia real sigue siendo poco conocida [Hollis et al., 2017]. Una mejor comprensión de estas relaciones puede mejorar las políticas de seguridad de los ciudadanos, por ejemplo, mejorando la asignación de recursos de seguridad de los ciudadanos o mejorando las estrategias de asignación de recursos en vigilancia para mitigar delitos particulares o mejorando la relación policía-ciudadano [Kleck and Barnes, 2014, Latané, 1981].

Algunas investigaciones adicionales sobre la PoS, seguridad predictiva y el uso de aprendizaje automático han explorado el contexto de Bogotá D.C. En un análisis [Reyes et al., 2020]

reciente, se examinó la propagación del miedo al crimen en las comunidades de Bogotá D.C. El estudio buscó comprender cómo las interacciones preferenciales entre individuos influencian la difusión del miedo al crimen y si existía un efecto de aislamiento del miedo al crimen en diferentes comunidades. Los resultados revelaron la presencia de un efecto de aislamiento del miedo al crimen en diferentes comunidades, especialmente en aquellas con baja susceptibilidad al crimen, lo que sugiere que los grupos menos vulnerables experimentan una mayor sensación de seguridad en sus interacciones comunitarias.

Otra investigación se enfocó en la predicción de homicidios utilizando modelos de aprendizaje automático y datos espaciotemporales [Villegas et al., 2022]. Los modelos evaluados incluyeron enfoques como el recuento estático, el modelo de Kernel Warping y el modelo de Graph Laplace of Gaussian. Los resultados resaltaron la superioridad de los modelos de aprendizaje automático, especialmente el modelo de Kernel Warping, en comparación con el modelo de recuento simple. Estos hallazgos destacan la importancia de implementar estrategias espaciales de alta resolución para optimizar la asignación de recursos policiales escasos.

En otro estudio, se compararon diferentes modelos de predicción del crimen en Bogotá [Barreras et al., 2016]. Se describieron cuatro métodos de predicción del crimen, incluyendo el modelo de puntos, el modelo de elipses espaciales, la estimación de densidad del núcleo y la dimensión temporal. Los resultados mostraron que el modelo de densidad del núcleo fue el más adecuado para predecir el crimen en la ciudad. Además, se observó que la dimensión temporal desempeñó un papel crucial en la precisión de la predicción, y la inclusión de características socioeconómicas mejoró la eficacia de los modelos. Los autores también desarrollaron un software con una interfaz general y funcionalidades de mapeo para la predicción del crimen.

En otro estudio, se analizó cómo las interacciones sociales en las comunidades afectan el

miedo al crimen [Pulido et al., 2019]. El estudio utilizó un modelo matemático para simular el miedo al crimen en un conjunto de individuos pertenecientes a una región particular. Se examinó la difusión del miedo al crimen a través de interacciones secuenciales y aleatorias entre individuos. Los resultados mostraron que el miedo al crimen se ve influenciado por la percepción de riesgo personal, la experiencia previa de victimización y la información transmitida socialmente. Además, se identificó que individuos con mayor influencia social pueden tener un impacto significativo en la propagación del miedo al crimen en la comunidad.

En otro artículo analizado, se presentó una metodología de aprendizaje semisupervisado basada en el aprendizaje de variedades para predecir homicidios en la ciudad de Bogotá [Pabón et al., 2020]. Su propósito fue desarrollar modelos predictivos que ayudaran a los agentes encargados de hacer cumplir la ley en su trabajo. Los resultados mostraron que la metodología propuesta pudo predecir con éxito los homicidios en Bogotá con una precisión del 80 %. Además, se encontró que las peleas callejeras eran un indicador importante para predecir los homicidios, lo que sugiere que las políticas públicas deberían haberse centrado en reducir las peleas callejeras como medida preventiva contra los homicidios.

Finalmente, en la investigación de [Ordoñez-Eraso et al., 2020] se enfocaron en la detección de tendencias de homicidios en Colombia utilizando aprendizaje automático. Presentaron un modelo de aprendizaje automático para analizar los datos y predecir la tendencia de aumento o disminución de la tasa de homicidios en Colombia y sus principales ciudades. Los resultados obtenidos fueron alentadores y muestran que los homicidios en Colombia tienden a disminuir.

Estas investigaciones resaltan la importancia de abordar la problemática del crimen y la seguridad en Colombia y en particular para Bogotá D.C., y ofrecen herramientas y enfoques novedosos para comprender, predecir y mitigar el crimen en la ciudad. Los hallazgos de estos estudios pueden respaldar la toma de decisiones por parte de los departamentos policiales y

contribuir a la implementación de políticas públicas efectivas en la lucha contra el crimen.

2.3. Interpretabilidad en aprendizaje automático

El aprendizaje automático interpretable significa que los humanos pueden capturar conocimiento relevante de un modelo en relación con las relaciones contenidas en los datos o aprendidas por el modelo [Doshi-Velez and Kim, 2017, Molnar, 2022]. Aunque este concepto se ha popularizado con el avance del aprendizaje de máquina. Hasta donde conocemos, no existe una definición formal para este concepto. En el área de interpretabilidad se espera que el modelo de aprendizaje automático además de obtener una predicción adecuada debe adicionalmente explicar (en un sentido humano) cómo llegó a la predicción (el ¿por qué?), ya que, para ciertos casos, una predicción correcta solo resuelve parcialmente el problema original establecido [Doshi-Velez and Kim, 2017, Molnar, 2022].

El artículo de [Cambria et al., 2022] analizó la naturaleza multidisciplinaria del análisis de sentimientos y su interacción con disciplinas como la lingüística, la psicología y la ciencia cognitiva. Se presentaron recomendaciones para mejorar su aplicación en el campo financiero, destacando la necesidad de enfoques interdisciplinarios y la adaptación a las teorías y estructuras de datos financieros. Se discutieron aplicaciones en la predicción de precios de acciones y la asignación de activos, así como desafíos en la captura precisa de emociones y la mitigación del p-hacking. Esto es, se resaltó la importancia de la colaboración multidisciplinaria para avanzar en el análisis de sentimientos y su aplicación en diversos campos, incluyendo las finanzas.

Para un tomador de decisiones, por ejemplo, la administración pública de una ciudad, es necesario entender las características principales de las opiniones de los ciudadanos relativas a su percepción de seguridad en el momento. Adicionalmente es útil comprender, tanto el nivel de PoS que los ciudadanos perciben como las motivaciones o causas detrás del mismo, ya sea,

por ejemplo, a nivel emocional (miedo, estrés, rabia, entre otros) como a nivel institucional (desconfianza en las instituciones, incertidumbre en la gobernabilidad, entre otros). En este sentido la interpretabilidad puede inclusive convertirse este en un aspecto de interés político cuando afecta el comportamiento de los ciudadanos y la imagen que tenga la ciudad hacia el público en general [Sánchez, 2008].

Algunos métodos de interpretación global (como los modelos sustitutos globales o de descomposición funcional, entre otros) han sido usados para establecer medidas de interpretabilidad global en modelos específicos [Zschech et al., 2022, Wang et al., 2021].Estos modelos pueden ser vistos como medidas de comportamiento promedio de los modelos a través de las categorías establecidas, a menudo expresados como valores esperados de la distribución de los datos [Molnar, 2022]. En particular, para el caso de los modelos sustitutos globales, resulta difícil plantear un medida de proximidad adecuada (R^2, Coeficientes de Correlación, entre otros) entre estos y el modelo de caja negra que impida el sobreajuste al conjunto de datos, sumado a que tales modelos no ven el resultado "real" del problema [Molnar, 2022] sino la aproximación brindada por el modelo a sustituir, esto es, el desempeño de la caja negra no es relevante al momento del entrenamiento del modelo sustituto; o como en el caso de la descomposición funcional que está limitada por la interacción de componentes de alta dimensionalidad entre dos o más características [Hooker, 2004], lo que lo convierte, en muchos casos, en un proceso costoso computacionalmente e inviable en muchos sentidos. Sumado a lo anterior, es importante observar que el enfoque de "abajo" hacia "arriba" (construir modelos de regresión) implica un proceso bastante manual e impone muchas restricciones en el modelo que pueden afectar el rendimiento predictivo [Hooker, 2004, Hooker, 2007]. Por estas razones, en este trabajo se opta por modelos de interpretabilidad local computacionalmente menos costosos y restrictivos, aunque estos dependan demasiado de las condiciones alrededor de una instancia en particular [Wang et al., 2021].

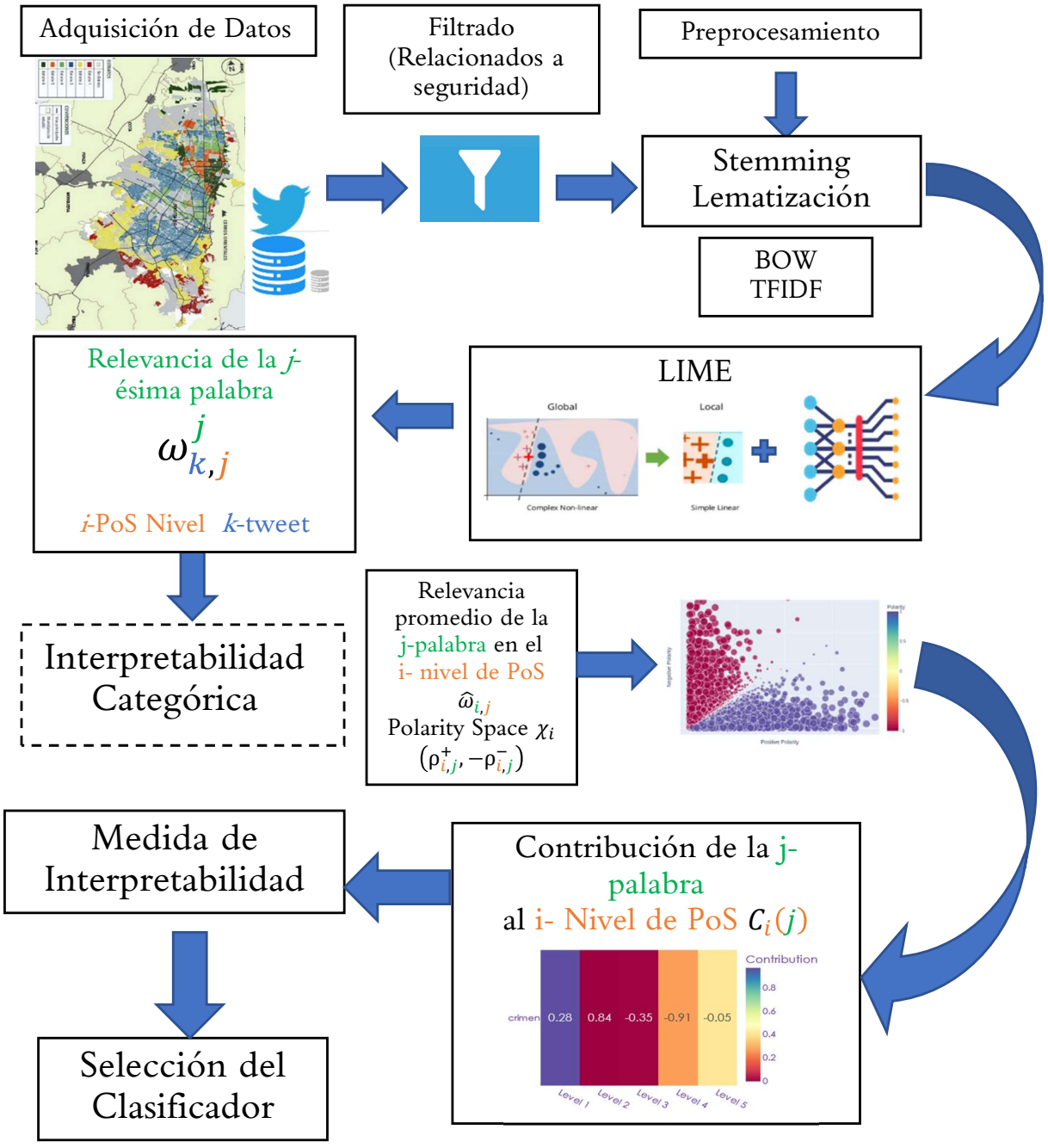
Recientemente, las técnicas de interpretabilidad local exploraron esta cuestión a nivel de predicción individual, esto es, un único tweet [Chaparro et al., 2021a]. Sin embargo, no existe una interpretación de *¿Cuáles son las características inherentes a un grupo de tweets que los clasifican en el mismo nivel de percepción de la seguridad?*.

3. Materiales y métodos

La Figura **3-1** ilustra el método propuesto. La estrategia comienza con la adquisición de datos, seguida de un filtro de datos relevantes a la temática de seguridad. Los datos pasan por un preprocesamiento y extracción de características (stemming y lematización) para luego crear la representación vectorial (diccionarios) en BoW y TF-IDF. De esta manera es posible implementar los diferentes métodos de clasificación de análisis de sentimientos (MNB, LR, BNB y SGD) para la distribución en las distintas categorías de PoS. Posteriormente, un esquema LIME basado en el operador de contracción y selección del mínimo absoluto (LASSO por sus siglas en inglés) se utilizó para proporcionar interpretabilidad a cada tweet. A continuación, se combinaron las contribuciones lineales de cada palabra proporcionadas por LIME para los tweets del mismo nivel de PoS con el fin de proporcionar interpretaciones para cada categoría. Finalmente, se establece una medida de interpretabilidad basada en el ML-SentiCon [Cruz et al., 2014, Chaparro et al., 2020] para seleccionar al mejor clasificador analizado.

3.1. Datos

Los contenidos analizados proceden de la red social Twitter. La base de datos contiene 26.255 tweets geolocalizados en la ciudad de Bogotá D.C, Colombia. Para obtenerlos, en primer lugar, los datos fueron recolectados utilizando la API de streaming de Twitter



Fuente: Elaboración Propia.

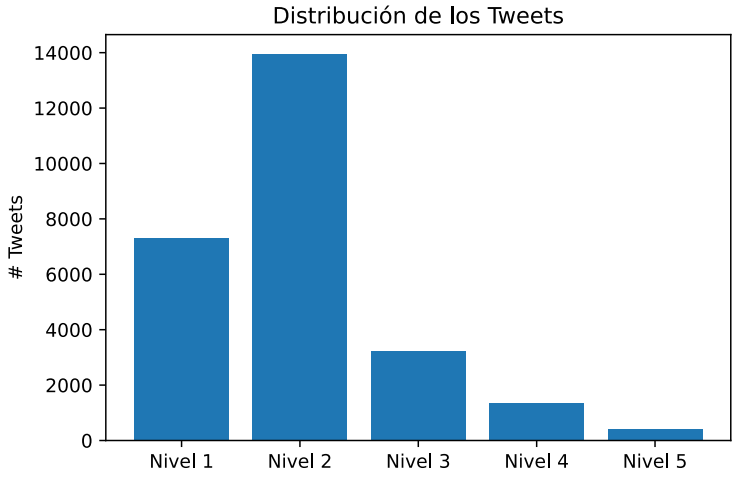
Figura 3-1.: *Representación esquemática de la interpretabilidad del análisis de sentimientos sobre los Tweets.*

entre el 18 de marzo de 2019, y el 28 de abril de 2020 (411 días). Se recogieron 1.701.668 tweets por el proyecto "Diseño y validación de modelos de analítica predictiva de fenómenos de seguridad y convivencia para la toma de decisiones en Bogotá", del Banco de Programas y Proyectos de Inversión Nacional, Departamento Nacional de Planeación, Gobierno de Colombia (BPIN: 2016000100036). Estos tweets se filtraron además utilizando una lista de palabras relacionadas con la seguridad para identificar el contenido relacionado. Aunque el primer filtro de tweets con palabras clave estaba relacionado con la seguridad, en algunos casos, los tweets resultantes hacían referencia a otros escenarios fuera de los deseados, como la seguridad bancaria, la seguridad sanitaria, entre otros [Chaparro et al., 2021b]. Además, este primer filtrado dio lugar a contenidos que descalificaban o insultaban a las personas. Debido a esto, con la ayuda de la Secretaría de Seguridad Convivencia y Justicia de Bogotá (SSCJ - Secretaría de Seguridad, Convivencia y Justicia de Bogotá), fue definido un nuevo conjunto de palabras clave [Chaparro et al., 2021b]. Estas palabras clave corresponden a pares o tríos de palabras centradas en la seguridad ciudadana que suelen aparecer en la misma frase y están relacionadas con la seguridad. Con estas nuevas palabras clave, se definió un filtro para considerar los mensajes que contienen estas palabras clave compuestas. Esta estrategia limitó el universo de búsqueda de una manera más específica y objetiva para los tweets relacionados con la seguridad en Bogotá. La lista final de términos clave contenía 1.758 términos relacionados con la seguridad. Es importante destacar que la estrategia de filtrado basada en palabras de dominios aquí utilizada se basa en un trabajo previo relacionado [Prieto Curiel and Bishop, 2016].

3.1.1. Datos relacionados a Seguridad

Se seleccionó aleatoriamente un subconjunto de 43.322 tweets con contenido vinculado a la seguridad y se anotaron como genuinamente relacionados o no con la seguridad. Estos tweets se distribuyeron en conjuntos no superpuestos a siete anotadores (cinco autores y

dos colaboradores del SCCJ). Estos expertos proporcionan un instrumento de filtrado sobre los términos relevantes en los textos de seguridad, debido a su alto grado de conocimiento sobre estos temas, ya que trabajan diariamente en este campo [Chaparro et al., 2021b]. La tarea de anotación consistió en etiquetar cada tweet como relacionado o no con la seguridad ciudadana basados en un proceso de anotación similar utilizado previamente en el análisis de datos relacionados con la delincuencia [Prieto Curiel et al., 2020]. Estos tweets se utilizaron posteriormente para entrenar una estrategia de clasificación automática para discriminar entre tweets relacionados o no con la seguridad [Pulido et al., 2021]. Luego de este proceso, se obtuvieron los 26.255 tweets usados en esta investigación.



Fuente: Elaboración propia.

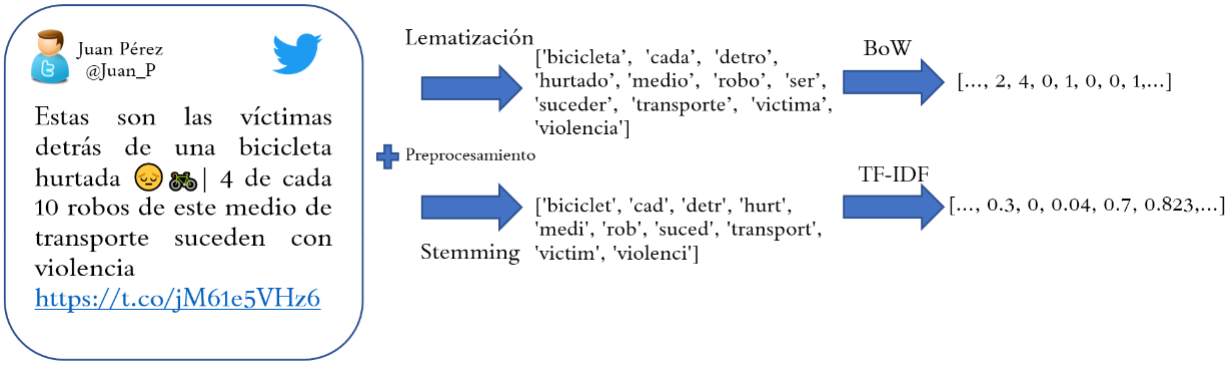
Figura 3-2.: *Distribución de la puntuación realizada por los expertos sobre la base de datos. Los expertos puntúan cada tweet, en función del sentimiento en que fueron escritos originalmente, en una escala de 1 a 5, donde 1 significa un sentimiento negativo (como ira o frustración) y 5 un sentimiento positivo (como optimismo o felicidad).*

3.2. Preprocesamiento y Representación

Con el conjunto de datos definidos se realiza el proceso de limpieza de los mismos eliminando hashtags, links, signos especiales y de puntuación así como una normalización a partir de una estandarización de las letras mayúsculas [Da Silva et al., 2014]. La Figura **3-3** muestra la distribución en nube de las palabras con mayor frecuencia en el texto. Luego, se eliminaron las palabras vacías usando el paquete de procesamiento de texto *Natural Language Toolkit: ntlk-stopwords-spanish* [Bird, 2006] previo a la creación de los diccionarios requeridos para la representación vectorial que se obtienen por los métodos de BoW y TF-IDF, el primero a través del algoritmo conocido como SnowballStemmer [Porter, 2001] y el segundo con el paquete de NLP conocido como Stanza [Qi et al., 2020], con el fin de minimizar el espacio de representación creado [Bhagvat, 2011, Chaparro et al., 2020]. Finalmente, se obtiene de esta forma una representación concisa y consistente, tal como lo muestra la Figura **3-4**, la cual será usada en los métodos de clasificación descritos en la siguiente sección.

Fuente: Elaboración Propia.

Figura 3-3.: *Distribución de las palabras más frecuentes en el conjunto de Tweets.*



Fuente: Elaboración Propia.

Figura 3-4.: *Ejemplo de representación vectorial de un tweet para los pares extractor de características + diccionario. Es necesario recalcar que ambos métodos ignoran la organización semántica del texto enfocándose en la ocurrencia de las palabras presentes en él.*

3.3. Métodos de Clasificación

En los métodos de clasificación supervisados, los documentos etiquetados se agrupan en clases predeterminadas. Esto significa que se puede construir un modelo según muestras existentes en función de las cuales se asignan los datos no etiquetados a sus respectivas categorías [Ayodele, 2010]. Para clasificar la PoS de los ciudadanos, en [Chaparro et al., 2021b] utilizaron distintos métodos de clasificación obteniendo en métricas de desempeño estándar resultados similares. Posteriormente en [Chaparro et al., 2020] se seleccionaron los mejores 4 métodos, los cuales serán usados y evaluados en esta investigación y descritos a continuación.

3.3.1. Multinomial Nave Bayes

El método supervisado conocido como *multinomial Nave Bayes o multinomial NB (MNB)* es un modelo de aprendizaje probabilístico [Christopher et al., 2008]. Se basa en la aplicación del teorema de Bayes para predecir la probabilidad de que un documento d pertenezca a una

clase c [Anguiano-Hernández, 2009], la cual es calculada como:

$$P(c|d) \propto P(c) \prod_{1 \leqslant k \leqslant n_d} P(t_k|c) \tag{3-1}$$

Donde $P(t_k|c)$ es la probabilidad condicional de que el término t_k ocurra en un documento de clase c y puede interpretarse como una medida de la contribución del término (o token) t_k para que c sea la clase correcta, $P(c)$ es la probabilidad a priori de que un documento pertenezca a la clase c y $n_d = |\{t_k\}|$ es la cantidad de términos en d. Debido a que las probabilidades $P(t_k|c)$ y $P(c)$ por defecto se desconocen estas son reemplazadas por sus respectivos estimadores de máxima verosimilitud, los cuales son los usados en las ecuaciones (3-4) y (3-5).

$$P(t_k|c) \propto \hat{P}(t_k|c) = \frac{T_{ct_k}}{\sum_{t_j \in V} T_{ct_j}} \tag{3-2a}$$

$$P(c) \propto \hat{P}(c) = \frac{N_c}{N} \tag{3-2b}$$

Donde $V = \bigcup_k t_k$ es el vocabulario o diccionario obtenido del conjunto de entrenamiento, T_{ctk} es el número de ocurrencias del término t_k en los documentos de entrenamiento para la clase c, N_c es el número de documentos en la clase c y N el número de documentos en total. Para evitar valores nulos en los MLE se suele realizar la denominada *suavización de Laplace*:

$$\hat{P}(t_k|c) = \frac{T_{ct_k} + 1}{\sum_{j \in V} T_{ct_j} + 1} \tag{3-3}$$

En clasificación de textos, el objetivo es encontrar la mejor clase para el documento [Christopher et al., 2008]. Esto es el estimador de máxima verosimilitud o *Maximum at posteriori* c_{MAP}:

$$c_{MAP} = \arg \max_{c \in \mathcal{C}} \hat{P}(c|d) = \arg \max_{c \in \mathcal{C}} \hat{P}(c) \prod_{1 \leqslant k \leqslant n_d} \hat{P}(t_k|c) \tag{3-4}$$

Aunque la ecuación (3-4) es más común en su forma logarítmica para evitar desbordamientos de punto flotante.

$$c_{MAP} = \arg \max_{c \in \mathcal{C}} \log \left(\hat{P}(c) + \sum_{1 \leqslant k \leqslant n_d} \hat{P}(t_k|c) \right) \tag{3-5}$$

La suma de las ponderaciones log a priori y por término es entonces una medida de cuánta evidencia hay de que el documento d pertenece a la clase c_{MAP}, y la ecuación (3-5) selecciona la clase para la cual se tiene mayor evidencia [Christopher et al., 2008].

3.3.2. Bernoulli Nave Bayes

Similar al anterior, el modelo de Bernoulli multivariante o BNB funciona de manera equivalente a un modelo de independencia binaria que genera para cada término t_k un indicador de la presencia o ausencia del término [Christopher et al., 2008]. Aquí, un documento d es una vector binario $d = \left[B_1, B_2, \cdots, B_{|V|}\right]$ sobre un espacio de palabras V donde cada dimensión de este espacio corresponde a la palabra t_k con $k = 1, \cdots, |V|$. Bajo la suposición de independencia propia de los modelos Naive Bayes se tiene:

$$P(d|c) = \prod_k B_k \cdot P(t_k|c) + (1 - B_k)(1 - P(t_k|c)) \tag{3-6}$$

Donde $P(w_t|c)$ se estima, utilizando la suavización de Laplace y MLE nuevamente, como:

$$P(t_k|c) \propto \hat{P}(t_k|c) = \frac{N_{ct_k}}{N} \tag{3-7}$$

Aquí N_{ct_k} son en número de documentos en la clase c que contienen al término t_k y N el número de documentos del conjunto de entrenamiento.

Ambos modelos Naive Bayes asumen que la características presentes, en este caso los términos, son independientes entre si. En general, la hipótesis de independencia condicional no es válida para los datos de texto dado que los términos dependen condicionalmente unos

de otros. Pero, los modelos Naive Bayes funcionan bien a pesar del supuesto de independencia condicional [Christopher et al., 2008].

3.3.3. Descenso de Gradiente Estocástico

El Descenso de Gradiente Estocástico es uno de los modelos de clasificación más usados en múltiples ámbitos [Bottou, 2010], dada la basta teoría que soporta la convergencia de este algoritmo [Ketkar, 2017]. Su funcionamiento es sencillo de formular para el contexto de clasificación de texto. Sean $d_1,\ d_2, \cdots,\ d_n$ los documentos y $x_1,\ x_2, \cdots,\ x_n, y_1,\ y_2, \cdots,\ y_n$ sus representaciones vectoriales y etiquetas respectivamente. Se desea encontrar una función $f(x|\omega)$, típicamente una función $f(x) = \omega \cdot x + b$, que prediga el valor y observado para un vector x. Para encontrar estos parámetros usualmente se plantea el problema de optimización:

$$\omega_n^* = \arg\min_{\omega} \frac{1}{n} \sum_{i=0}^{n} \mathbf{L}(y_i, f(x_i)) + \alpha \mathbf{R}(\omega) \tag{3-8}$$

Donde $\mathbf{L}(y_i, f(x_i)) = \log\left(1 + \exp\{-y_i f(x_i)\}\right)$ y $\mathbf{R}(\omega) = \|\omega\|_2^2$. Encontrar el parámetro optimización suele realizarse a través de un algoritmo de descenso [Nocedal and Wright, 2006], tal como sigue:

$$\omega \leftarrow \omega - \eta \left[\alpha \frac{\partial \mathbf{R}(\omega)}{\partial \omega} + \frac{\partial \mathbf{L}(y_i, f(x_i))}{\partial \omega} \right]$$

El Descenso de Gradiente Estocástico admite la clasificación multiclase combinando varios clasificadores binarios en un esquema "uno contra todos"(OVA por sus siglas en inglés). Para cada clase, se aprende un clasificador binario que discrimina entre esa clase y todas las demás. En el momento de la prueba, Se calcula la puntuación de confianza (es decir, las distancias con signo al hiperplano) de cada clasificador y se elige la clase con la confianza más alta [Pedregosa et al., 2011].

3.3.4. Regresión Logística

La regresión logística también se conoce en la literatura como regresión logit, clasificación de máxima entropía (MaxEnt) o clasificador log-lineal. En este método, las probabilidades que describen los posibles resultados de un solo ensayo se modelan mediante una función logística [Pedregosa et al., 2011].

En el procesamiento del lenguaje natural, los clasificadores LR multiclase se usan comúnmente como una alternativa a los clasificadores ingenuos de Bayes porque no asumen la independencia estadística de las variables aleatorias (comúnmente conocidas como características) que sirven como predictores. Sin embargo, el aprendizaje en un modelo de este tipo es más lento que para un clasificador de Bayes ingenuo, por lo tanto, puede no ser apropiado dado un gran número de clases para aprender [Bishop and Nasrabadi, 2006].

Para su formulación, se define al igual que el modelo anterior $d_1, d_2, \cdots, d_n$ los documentos en el conjunto de entrenamiento y $x_1, x_2, \cdots, x_n$, $y_1, y_2, \cdots, y_n \in \{1, \cdots, K\}$ sus representaciones vectoriales y etiquetas respectivamente. En lugar de un vector de parámetros se tiene una matriz W donde su k-ésima fila W_k corresponde a los parámetros de la clase k. De esta forma la probabilidad deseada se estima como:

$$P(y_i = k|x_i) \propto \hat{p}_k(x_i) = \frac{\exp\left(x_i \cdot W_k + W_{0,k}\right)}{\sum_{j=1}^{K-1} x_i \cdot W_j + W_{0,j}} \tag{3-9}$$

Donde la función u objetivo de optimización se convierte en:

$$\min_{W} -C \sum_{i=1}^{n} \sum_{k=1}^{K-1} P_k \log\left(\hat{p}_k(x_i)\right) + \mathbf{R}(W) \tag{3-10}$$

Donde P_k es un indicador binario de $y_i = k$ y $\mathbf{R}(W) = \frac{1}{2}\|W\|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}|W_{i,j}|$

3.3.5. Métricas de Desempeño

Las métricas de desempeño son coeficientes adimensionales que permiten evaluar el rendimiento de un clasificador [Carvalho et al., 2019]. Entre las métricas más usadas en evaluación de modelos de aprendizaje automático son: Precisión, Exhaustividad (Recall), Valor-F (F1-score), Exactitud (Accuracy), Matrices de Confusión y las curvas ROC (Receiver Operating Characteristic) junto a la medida AUC(Áreas bajo la curva) de estas.

Previamente, en [Chaparro et al., 2021b] realizaron la evaluación de distintos métodos de aprendizaje automático para el ajuste de modelos en temas relacionados a la PoS con algunas de estas métricas, por lo cual una de estas (F_1 Score) será usada en esta investigación. A su vez diferentes evaluaciones realizadas para conjuntos desbalanceados [Jurman et al., 2012], mostraron que otra métrica conocida como el *Coeficiente de correlación de Matthews (MCC)* [Matthews, 1975] que presenta ciertas ventajas sobre las métricas de desempeño estandár en el momento de la evaluación de modelos binarios y evaluación de modelos desbalanceados multiclase [Chicco and Jurman, 2020],[Madasamy and Ramaswami, 2017]. Con esto las métricas a usar serán el F_1 Score y MCC descritas a continuación.

F_1- Score

Esta métrica corresponde a la media armónica entre los indicadores de precisión y exhaustividad [Naser and Alavi, 2020]. La precisión es la proporción de predicciones que fueron etiquetadas de una clase y fueron correctamente etiquetadas por el modelo, este refleja la confianza que tiene el modelo en asignar a un individuo la etiqueta de la clase correcta, y la exhaustividad corresponde a la probabilidad de clasificar correctamente la clase correcta. Debido a que la media armónica toma valores cercanos a uno cuando la precisión y la exhaustividad encuentran un equilibrio, el $f1$ score permite identficar el balance entre ambas cantidades. Esta medida se caracteriza por dar la misma importancia a las clases sin importar su participación. Para el caso de un modelo multiclase se define: TP_k los predi-

chos correctamente en la clase k, TN_k los predichos incorrectamente en la clase k, FP_k los predichos incorrectamente que debieron ser de la clase k [Naser and Alavi, 2020]. Con esto:

$$f_1(k) = 2\ \frac{precision_k \cdot recall_k}{precision_k + recall_k} = 2\ \frac{TP_k}{2TP_k + FP_k + FN_k} \tag{3-11}$$

Como cada puntaje anterior depende de la clase a predecir k se utiliza como medida en el modelo multiclase un ponderado de cada una. Es decir:

$$F_1\ Score = \frac{\sum_k |c_k| f_1(k)}{\sum_k |c_k|} \tag{3-12}$$

Donde $|c_k|$ representa la cantidad de elementos de la clase k en el conjunto de validación. Teniendo en cuenta que está medida fue inicialmente diseñada para clasificaciones binarias, en algunos casos la definición de la ecuación (3-12) es conocida como $F_1\ Score\ Macro$, aunque no es la única manera en la que ha sido definida [Grandini et al., 2020].

Coeficiente de Correlación de Matthews (MCC).

Inicialmente propuesto por Matthews en [Matthews, 1975] y conocido también como el estadístico $\mathbf{R}_k$ es comparable al coeficiente ϕ de Pearson o de Yule en su interpretación [Zhu, 2020].

Basado en los datos obtenidos en la matriz de confusión multiclase de la clasificación obtenida se calcula el MCC simplificando los términos usados como sigue :

- $t_k = \sum_{i=1}^{K} C_{ki}$ el número de elementos de la clase k
- $p_k = \sum_{i=1}^{K} C_{ik}$ el número de veces en la que la clase k fue predicha.

- $c = \sum_{k}^{K} C_{kk}$ el número de elementos clasificados correctamente.
- $s = \sum_{i}^{K} \sum_{j}^{K} C_{ij}$ el número total de elementos en el conjunto de validación.

Con esto:

$$MCC = \frac{c \cdot s - \sum_{i}^{K} p_k \cdot t_k}{\sqrt{\left(s^2 - \sum_{i}^{K} p_k^2\right) \cdot \left(s^2 - \sum_{i}^{K} t_k^2\right)}} \tag{3-13}$$

Es necesario notar que este coeficiente tiene en cuenta a los elementos no clasificados en una clase i y que estén correctamente clasificados en su clase j. En clasificación binaria, esto significa que los no clasificados en la clase "positiva" pertenezcan a la clase "negativa" [Chicco and Jurman, 2020]. Pero se debe también resaltar que aunque incluye otros factores relevantes para el análisis del desempeño del método de clasificación que se use, no es de lejos el mejor y no necesariamente puede establecerse como una medida estándar para modelos de clasificación con datos desbalanceados [Zhu, 2020].

Finalmente se utilizó un método de validación cruzada k-fold para estimar la rendimiento del enfoque de aprendizaje automático [Refaeilzadeh et al., 2009]. Esta estrategia usa un único parámetro llamado k que se refiere al número de conjuntos en los que se dividirán los datos. El procedimiento de validación cruzada aquí utilizado fue el siguiente. En primer lugar, los datos se mezclaron aleatoriamente y se dividieron en $k = 10$ grupos. Cada grupo se usó como un conjunto de datos de prueba, mientras que los otros grupos como entrenamiento en una proporción $90\,\% - 10\,\%$ dando como resultados los valores que se muestran en las Figuras **4-5** y **4-6**.

3.4. Interpretabilidad Local

El esquema de interpretabilidad usado en esta investigación sigue la estructura mostrada en [Chaparro et al., 2021a] para el clasificador MNB, la cual a su vez se basa en los modelos de interpretabilidad local agnóstica al modelo [Ribeiro et al., 2016b]. Este modelo sustituto de explicación funciona tomando un elemento del conjunto de datos ya vectorizado (usualmente llamado instancia) $x \in \mathbb{R}^d$ y su respectiva etiqueta y. A partir de x se crea una versión "interpretable", usualmente la versión binaria del vector x, definida como $x^{'} \in \mathbb{R}^{d'}$ y se crea un conjunto $Z' = \{z'_i\}_{i \in I} \subset \mathbb{R}^{d'}$ de elementos similares a $x^{'}$ con la única diferencia de que estos tienen alguna de sus entradas nulificadas. Esto es, existe $k \in \{1, \cdots, d'\}$ tal que $z'_i \cdot e_i \neq x'_i \cdot e_i$. Posteriormente, para cada z' se recupera la muestra en su representación original $z \in \mathbb{R}^d$. Seguido a esto, se usa el modelo de clasificación para obtener las etiquetas del conjunto $Z = \{z_i\}_{i \in I}$ y a partir de estos valores se define un modelo lineal substituto $g(x^{'}) \coloneqq \omega \cdot x^{'} + b$ de tal manera que $g(Z') = Y$. Para definir lo mejor posible a g se plantea el problema de optimización:

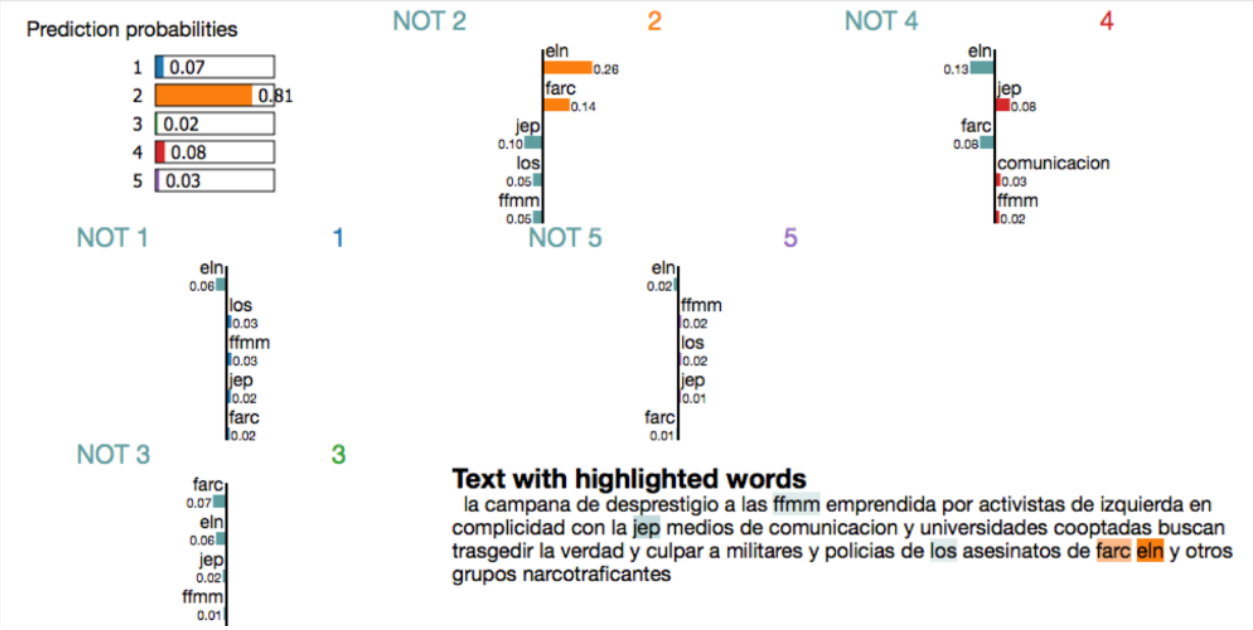
$$\arg \min_{g \in G} \mathbf{L}(f, g, \pi_x) + \Omega(g) \tag{3-14}$$

Donde $\mathbf{L}((f, g, \pi_x)) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z'))^2$ es la función de perdida de mínimos cuadrados con la contracción $\pi_x(z) = \exp\{-D(x, z)^2/\sigma^2\}$, donde $D(x, z)$ representa la "distancia" en clasificación de texto conocida como *similitud coseno* y por último $\Omega(g) = \|W\|$ la penalización de la regresión definido como el número de pesos no nulos.

El método de aproximación de estos pesos se conoce como el *operador de selección y contracción mínima absoluta (LASSO)* [Efron et al., 2004] con el cual se pretende anular gran parte de estos coeficientes. Una muestra de representación de estos valores se puede ver en la Figura **3-5** obtenida gracias al paquete de software[1]desarrollado por [Ribeiro et al., 2016b].

[1]Disponible en el repositorio de Github de [Ribeiro et al., 2016b] `https://github.com/marcotcr/lime`.

En esta figura es posible observar las probabilidades de clasificación en cada categoría, las cuales provienen del clasificador presente en el modelo. A su vez, al variar la categoría de la instancia observada se realizan las regresiones planteadas y se crean los diagramas de barras donde se plasman los valores de mayor magnitud sin importar el signo para cada una de las categorías. En otras palabras, representa las características de la instancia en particular que son relevantes (o no si el signo es negativo) para la clasificación.



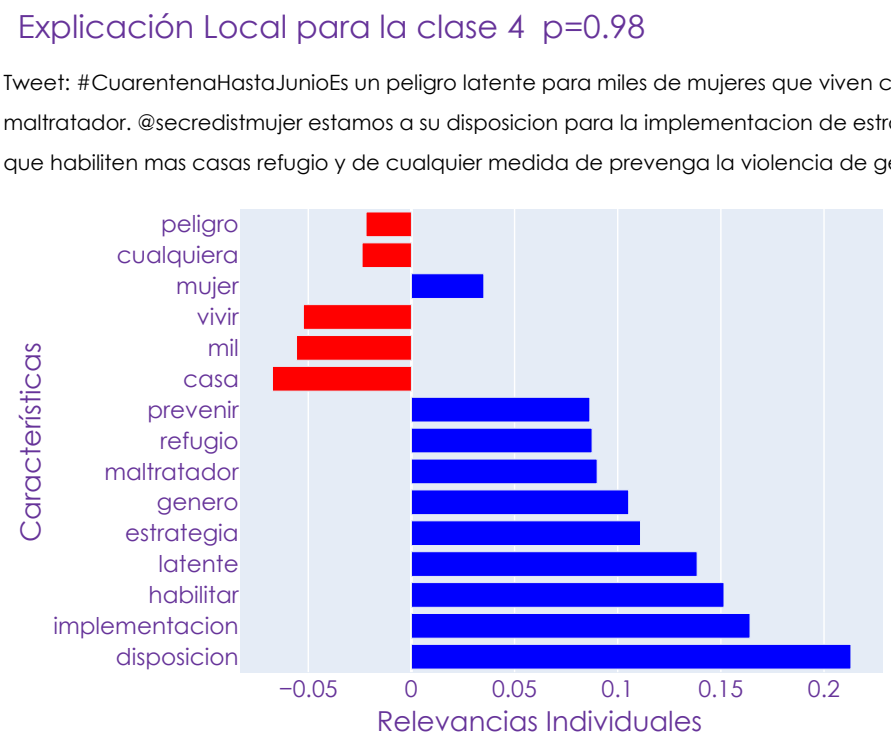
Fuente: Elaboración propia usando LIME [Ribeiro et al., 2016b].

Figura 3-5.: *Funcionamiento e interfaz de la explicación agnóstica al modelo LIME de un tweet.*

El proceso de extracción de pesos puede repetirse cambiando la etiqueta de la instancia utilizada por cualquiera de las 5 disponibles, los valores obtenidos para la j-ésima característica del k-ésimo tweet en la i categoría de percepción de seguridad se interpretan como la medida de relevancia de esa característica para que el tweet tenga determinada clasificación (o la no pertenencia en caso de ser negativa). Con esto, en lo sucesivo a este trabajo a tal coeficiente se le denominará como $\omega_{i,k}^{j}$: *Relevancia individual de la característica j en el k-ésimo tweet para la categoría PoS i .*

Las figuras **3-6** y **3-7** reportan relevancias individuales obtenidas para dos tweets con

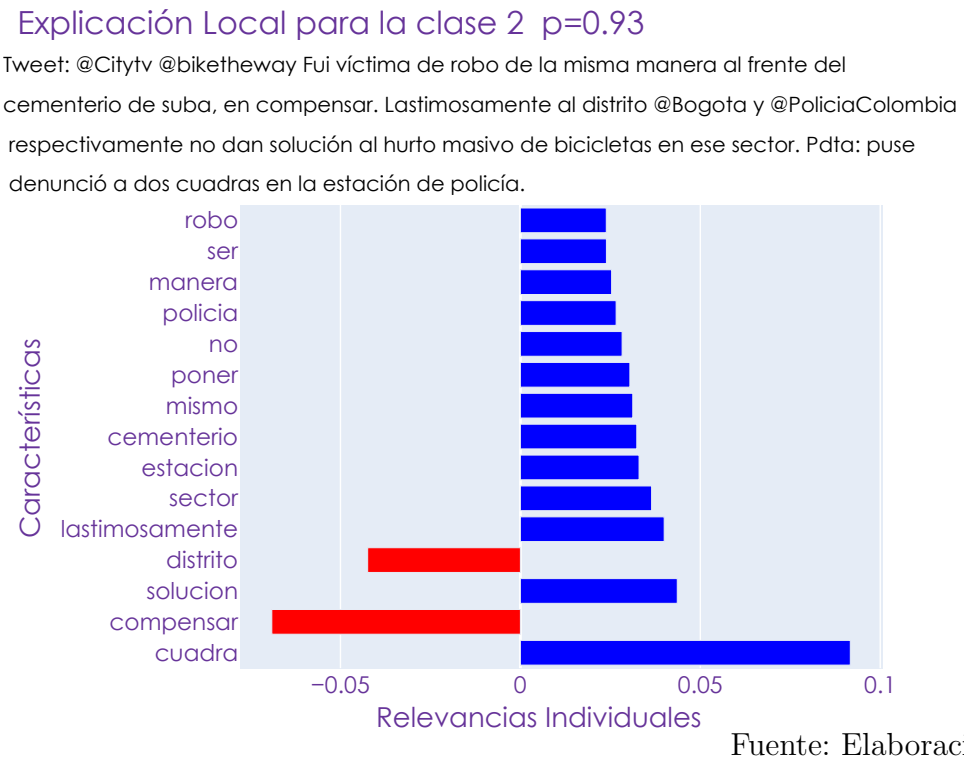
distintos niveles de PoS, niveles 4 y 2, respectivamente. Para estas dos instancias, que representan ejemplos de las clases "positivas" y "negativas", es importante observar la selección y polaridad atribuida por el método de interpretabilidad local a algunos de sus términos constituyentes. Por ejemplo, las palabras "refugio", "estrategia" y "prevenir" del tweet presente en la Figura **3-6**, corresponden a elementos de polaridad e interpretación positiva en el lenguaje común. Es decir, la presencia de estas palabras puede que el mensaje es positivo. A su vez, las palabras "peligro", "vivir" y "casa" tienen una relevancia negativa que podrían motivar la clasificación de este tweet como negativa. Dado que el mensaje invita la protección de un hecho negativo, podría pensarse que el clasificador automático discierne entre los distintos significados compuestos presentes en el mensaje.



Fuente: Elaboración propia.

Figura 3-6.: *Relevancias individuales del modelo seleccionado para la categoría 4.*

En la Figura **3-7** es posible observar el comportamiento de la transformación propuesta para las polaridades establecidas en el lexicón Ml-Senticon. Si se analiza el comportamiento de las palabras "robo", "lastimosamente", "cementerio" de connotación negativa, sus respectivos valores de relevancia son positivos para esta categoría de PoS. Por el contrario, palabras como "policía", "estación", "sector" y "cuadra" representan, en este ejemplo, características para la clasificación en la categoría negativa (2) aunque por si solas tengan polaridades positivas o neutras. Sin embargo, lo que estos dos ejemplos nos muestran es la imposibilidad de vislumbrar características de los grupos de tweet a los que estos pertenecen. No es posible a partir de su relevancias individuales dar una explicación interpretable de ciertas palabras ni tampoco es posible comprender el comportamiento general de cada una de las palabras presentes en ellos, ¿Por qué toma estos valores específicamente?, ¿Cómo se evidencia el ajuste del modelo, con elementos de entrenamiento de diferentes categorías PoS, al momento de crear las relevancias individuales?.



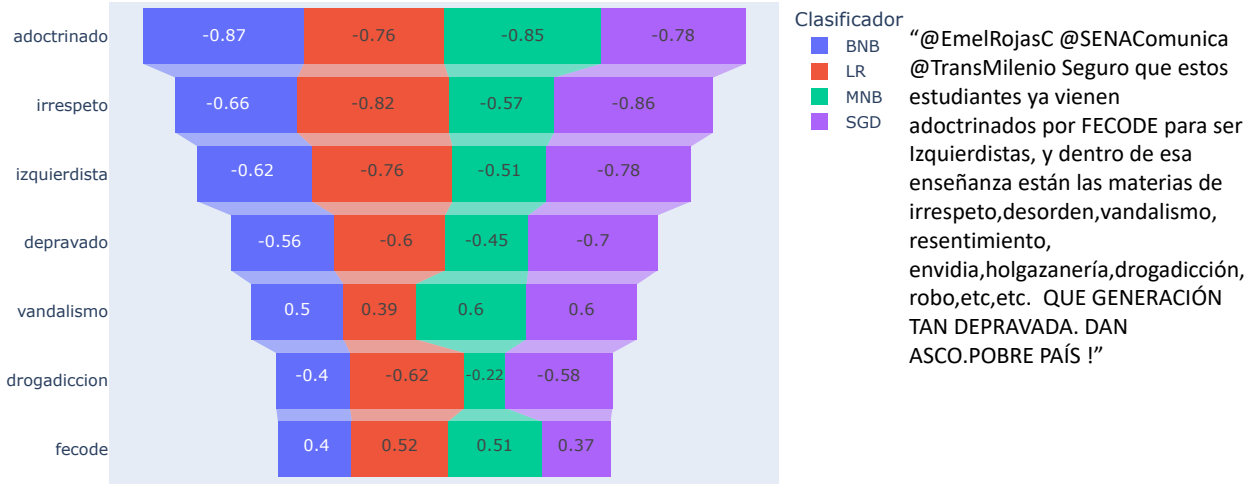
Fuente: Elaboración propia.

Figura 3-7.: *Relevancias individuales del modelo seleccionado para la categoría 2.*

3.5. Interpretabilidad Categórica

3.5.1. Comportamientos Globales Caracterizables

La limitación existente en las explicaciones locales dadas radican en la variabilidad de los valores obtenidos. Esta variabilidad resulta de las condiciones de regresión y convergencia planteadas en la ecuación (3-14) para cada palabra o característica analizada. Sin embargo, comportamientos evidenciados en la interpretabilidad local mostraron la posibilidad de establecer un parámetro similar a la relevancia planteada. Por ejemplo, al examinar los comportamientos en instancias de los distintos modelos fue posible observar cierta "independencia" al modelo ajustado en el sentido de obtener valores similares tanto en magnitud como en polaridad tal como se observa en la Figura **3-8**.



Fuente: Elaboración propia.

Figura 3-8.: *Relevancias individuales para los diferentes modelos estudiados. En este ejemplo el modelo esta compuesto por la extracción de características Lema y representación TF-IDF.*

A su vez, al compararse tweets pertenecientes a la misma categoría PoS y que comparten ciertas similitudes semánticas fue posible observar que tales similitudes se transferían a las relevancias establecidas. Como lo muestra la Figura **3-9**. En esta figura, ambos tweets

pertenecen a la categoría 1 de PoS, las relevancias positivas en ambos casos corresponden a palabras de significado peyorativo y/o polaridad negativa en el lenguaje español. Inclusive, la característica más relevante para ambas instancias es la misma palabra con la misma polaridad y magnitud similar.

Fuente: Elaboración propia.

Figura 3-9.: *Relevancias individuales para 2 tweets de categoría* 1 *de PoS bajo el modelo Lema-TFIDF-Regresión Logística.*

De esta manera, es razonable concluir la existencia de comportamientos globales potencialmente caracterizables. Esto significa que es posible introducir una metodología que solucione los inconvenientes locales con el fin de conocer el comportamiento general de la categoría PoS analizada, a este proceso se le denominará *Interpretabilidad Categórica o a nivel de Categorías.*

3.5.2. Relevancia promedio y polaridades

Para definir la interpretabilidad en las categorías PoS es necesario puntualizar algunos elementos a partir de la noción de *relevancia* $\omega_{i,k}^{j}$ de la sección anterior. En primer lugar

sea $W_i^j = \left\{\omega_{i,k}^j\right\}_{\substack{j\in k\\ k\in T}}$, la colección de todas las relevancias individuales de la característica j para la categoría de PoS i, donde T es el conjunto de datos correctamente clasificados por el modelo seleccionado. Seguido a esta definición, se construyen los términos de las ecuaciones (3-15a), (3-15b), (3-15b) denominados *relevancia promedio, polaridad positiva y negativa de la característica j para el nivel PoS (i)* respectivamente. Estos valores pretenden representar el "peso" general de la palabra y las tendencias tanto positivas como negativas de todas las instancias donde esta se encuentre para la categoría i de PoS.

$$\widehat{\omega}_{i,j} = \frac{\sum\limits_{\omega\in W_i^j} \omega}{\left|W_i^j\right|} \tag{3-15a}$$

$$\rho_{i,j}^{+} = \frac{1}{\left|I(W_i^j)\right|} \sum_{\substack{\omega\in W_i^j\\ \omega>0}} \omega \tag{3-15b}$$

$$\rho_{i,j}^{-} = \frac{1}{\left|I(-W_i^j)\right|} \sum_{\substack{\omega\in W_i^j\\ \omega<0}} \omega \tag{3-15c}$$

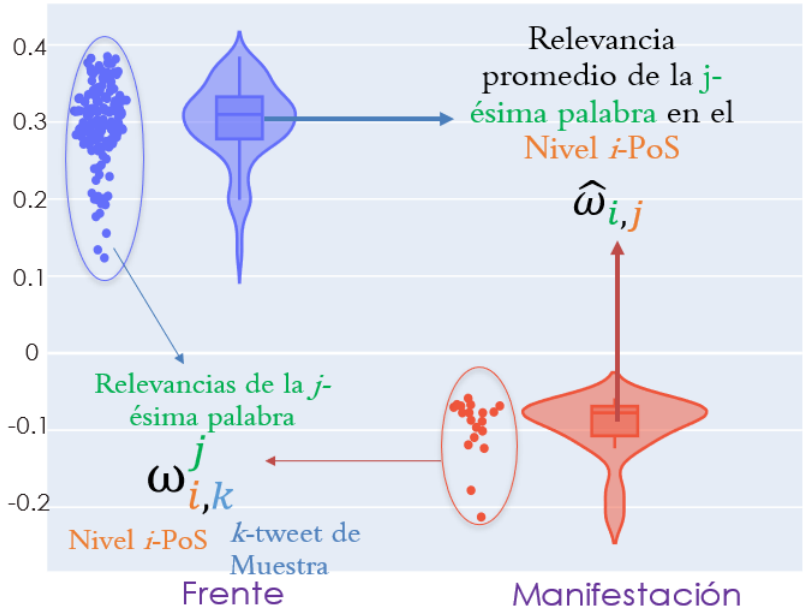
La Figura **3-10** muestra una representación de estos valores, para las características "Frente" y "Manifestación".

Con estas polaridades es posible definir una representación bidimensional (véase Figura **3-11**) que reduce las características y permite establecer una medida de contribución a las categorías de PoS de cada palabra que este presente en el conjunto de datos.

Esta representación se definirse como:

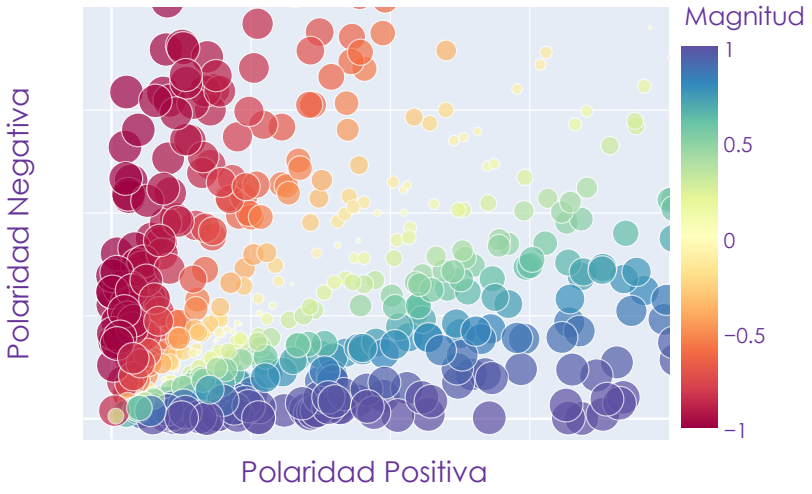
$$\chi_i = \left\{\left(\rho_{i,j}^{+}, -\rho_{i,j}^{-}\right)\ , \forall j \in diccionario(T)\right\} \tag{3-16}$$

Finalmente, dadas estos valores es posible definir la medida ***Contribución a la categoría***



Fuente: Elaboración propia.

Figura 3-10.: *Esquema de construcción de relevancias y polaridades para dos características "Frente" y "Manifestación"para la categoría de percepción de seguridad 2. Para estas palabras sus respectivas relevancias promedio coinciden con sus polaridades, más es una excepción no una norma que esto suceda.*



Fuente: Elaboración propia.

Figura 3-11.: Representación bidimensional del espacio de polaridades para la categoría PoS 2 (χ_2) bajo el modelo LEMA-BOW-MNB.

***de PoS** i **de la característica** j como sigue:

$$C_i(j) = |S(\widehat{\omega_{i,j}}, \kappa)| \cdot \cos\big(2 \cdot \theta(\rho_{i,j}^{+}, -\rho_{i,j}^{-})\big) \cdot \left|W_i^j\right| \tag{3-17}$$

$$= \left|\left(\frac{1 - e^{-\kappa \cdot \widehat{\omega}_{i,j}}}{1 + e^{-\kappa \cdot \widehat{\omega}_{i,j}}}\right)\right| \cdot \left(\frac{(\rho_{i,j}^{+})^2 - (\rho_{i,j}^{-})^2}{(\rho_{i,j}^{+})^2 + (\rho_{i,j}^{-})^2}\right) \cdot \left|W_j^i\right| \tag{3-18}$$

$$NC_i(j) \coloneqq S(C_i(j), \kappa) \tag{3-19}$$

Con $S(X, \kappa) = \dfrac{2}{1 + e^{-\kappa X}} - 1$ una transformación de la función sigmoidea y $\kappa = 8$ un parámetro de holgura.

La interpretación que se brinda de esta contribución debe analizarse en sus 3 componentes. En primer lugar, la transformación sigmoidea en valor absoluto pretende normalizar las relevancias promedio y establecerlas en el intervalo $[-1, 1]$, el valor absoluto reflejaría el nivel de relevancia de una característica o palabra para determinado nivel de PoS. El segundo término basado en el ángulo obtenido en la representación bidimensional pretende transformar la polaridad existente en un cociente que refleje las mismas propiedades, esto es un nuevo coeficiente de polaridad del término analizado. Por último el parámetro de frecuencia del termino reflejaría la importancia en el conjunto de datos de la expresión usada. Luego, la versión normalizada de esta contribución realiza un "control" de estos términos comprimiéndolos al intervalo $[-1, 1]$.

3.5.3. Medida de Interpretabilidad

Con la introducción de la noción de *contribución de la $i-$ésima palabra a las categorias PoS* $C_i(j)$ dado un modelo es posible asociar un conjunto de contribuciones de las palabras a las categorías para cada uno de estos. Es decir, es posible cuantificar la importancia de las características en cada una de las categorías dependiendo del modelo de clasificación utilizado. En esta sección, utilizando estas contribuciones se introducirá una nueva medida

que cuantifique el nivel de interpretabilidad del modelo. Para ello, se tomará como base el conjunto de lexicón de polaridades semánticas a nivel de lemas en el idioma español conocido como *ML-Senticon* obtenido de [Cruz et al., 2014]. Con esto, es posible obtener un diccionario de valores (llamados polaridades) para un buen número de palabras en el idioma español, estos valores definen si la palabra puede clasificarse como *positiva, negativa o neutra* y presenta un valor cuantitativo de tal categorización. Ahora bien, con estos elementos, se traducirán tales polaridades a contribuciones para las categorías PoS de las palabras. Esto es, se establecerá que el comportamiento de las palabras negativas, positivas y neutras del idioma español tengan un valor similar para las diferentes categorías PoS, asumiendo que Tweets con palabras negativas serían clasificados como negativos para la PoS, Tweet con palabras neutras representarían categorías neutras y tweets con características positivas serían positivamente clasificados. Tal conversión se explica en el Algoritmo 1.

Algoritmo 1 Polaridades a Contribuciones

Entradas: Conjunto de Tweets preprocesados, vectorizados y clasificados T, ML-Senticon.
Salidas: Un conjunto de vectores $\Omega := \{\omega_t : t \in T\}$ los cuales representan la transformación de las polaridades de las palabras presentes en los tweets a contribuciones para las categorías de PoS.

$\Omega \leftarrow \{\}$
for $t \in T$ **do**
  $l_t \leftarrow$ Lematización(t) ▷ $l_t \in \mathbb{R}^p$
  $\omega_t = \mathbf{0}_{1\times p}$
  $\omega_t \leftarrow ML - Senticon(l_t)$ ▷ Cuando las palabras $l_t \in dict(ML - Senticon)$
  $s_t \leftarrow$ Categoría PoS(t)
  if $s_t \in \{1, 2\}$ **then**
    $\omega_t \leftarrow -\omega_t$
  else if $s_t \in \{4, 5\}$ **then**
    $\omega_t \leftarrow \omega_t$
  else
    $\omega_t \leftarrow 2e^{-\frac{\omega_t}{\rho}}$ ▷ Con $\rho = \frac{0.3^2}{\ln 2}$
  end if
  $\Omega \leftarrow \Omega \cup \{\omega_t\}$
end for
return Ω

Definido el conjunto de vectores conformados por las transformaciones de las polaridades para cada característica presente en el conjunto de tweets[2], se comparará para cada modelo el vector de contribuciones obtenido por cada uno de los tweets en su categoría PoS correspondiente y se evaluará su similaridad. De esta manera, definimos la *medida de interpretabilidad del modelo* como se establece en el Algoritmo 2 cuyos valores obtenidos se pueden observar en la tabla **A-1**.

Algoritmo 2 Medida de Interpretabilidad (IM).

Entradas: Conjunto de Tweets preprocesados, vectorizados y clasificados T, ML-Senticon, modelo (m).
Salidas: ρ coeficiente que representa la medida de interpretabilidad del modelo.

$\rho = 0$
Ω Como en el Algoritmo 1
for $t \in T$ **do**
  $s_t \leftarrow$ Categoría PoS(t)
  $\omega_t \leftarrow \Omega(t)$
  $C_t \leftarrow NC_{s_t}(t)|_m$ ▷ Este vector depende del modelo.
  $\rho_t \leftarrow 0.5 \cdot Similaridad\ Coseno(\omega_t, C_t) + 0.5$
  $\rho \leftarrow \rho + \rho_t$
end for
$\rho \leftarrow \frac{\rho}{\|T\|}$
return ρ

A su vez, comparando las métricas de desempeño (F1-Score y MCC) con la medida de interpretabilidad, se obtiene una representación bidimensional de cada modelo, para la selección del mismo se procura obtener el modelo que cuyos valores en ambos ejes sean similares, esto es, que interprete tan bien como se ajuste a los datos, para ello se definen las selecciones:

$$\text{Selección } F1\ (m) := \left(1 - \left|\frac{b\overline{F1}(m) + aIM(m) + c}{\sqrt{a^2 + b^2}}\right|\right)\sqrt{\left(\overline{F1}(m)^2 + IM(m)^2\right)}$$

[2]Debido a que el Lexícon no posee una polaridad para todas las palabras del idioma español y algunas palabras de la jerga en común pueden estar presentes se tomará como 0 el valor de dicha palabra en dado caso.

$$\text{Selección } MCC(m) := \left(1 - \left|\frac{b\overline{MCC(m)} + aIM(m) + c}{\sqrt{a^2 + b^2}}\right|\right) \sqrt{\left(\overline{MCC}(m)^2 + IM(m)^2\right)}$$

Finalizadas las selecciones se establece un promedio de ambos valores y de esta manera elegir al modelo más adecuado.

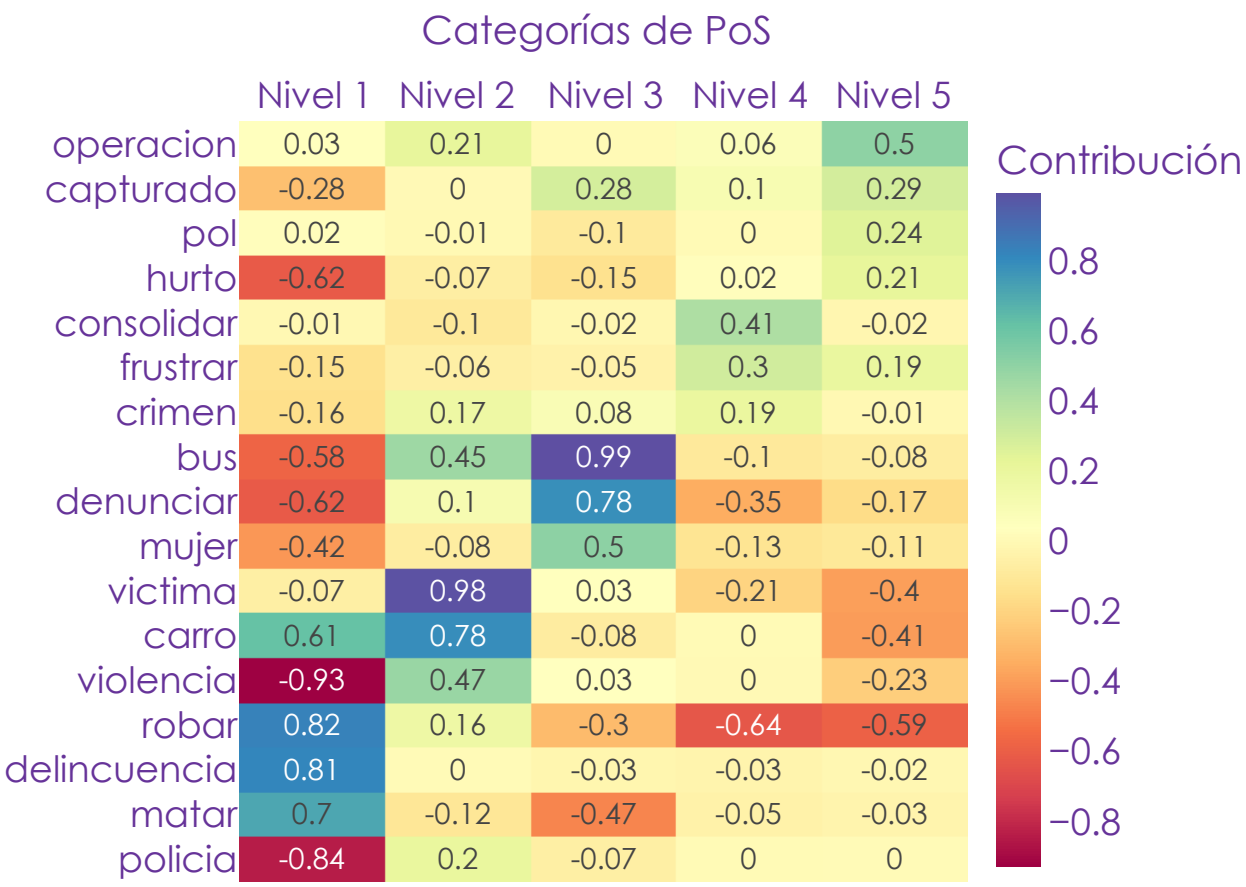
4. Resultados y discusión

Los resultados de este trabajo se organizan de la siguiente forma. En primer lugar, se reportan resultados de interpretabilidad categórica para las categorías PoS del conjunto de tweets estudiados para el modelo de clasificación individual seleccionado. A continuación, se reportan los resultados relacionados con la estrategia de comparación de modelo utilizada para seleccionar el modelo más interpretable y con altos desempeños. Y finalmente, se reportan las métricas de desempeño utilizadas para los métodos de clasificación analizados.

4.1. Interpretabilidad Categórica

La Figura **4-1** reporta las contribuciones (normalizadas) para algunas palabras del lenguaje común, consideradas como relevantes por juicio experto, para los diferentes niveles de percepción de seguridad, por el modelo de clasificación seleccionado (LEMA BOW MNB ver Sección 4.2). Como puede observarse, la figura sugiere una asociación directa entre los sentimientos asociados al nivel de PoS de estas palabras con la medida de contribución establecida. Por ejemplo, las palabras “operación”, “consolidar” y “capturado” comúnmente utilizadas por organismos estatales para la divulgación de actividades en contra de la delincuencia están presentes en la categoría 5 de PoS. Estas palabras aparecen en los tweets de la categoría 5 (percepción positiva), y adicionalmente son relevantes para la clasificación en dicha categoría. En constaste, las palabras “víctima”, “delincuencia”, “robar” y “matar” se

asocian a mensajes con categorías bajas de percepción de PoS (1 y 2), y también contribuyen para la clasificación de los mensajes en esa categoría.



Fuente: Elaboración propia.

Figura 4-1.: *Contribuciones normalizadas de palabras relevantes para la PoS a través de las categorías de PoS.*

Es interesante observar como el comportamiento de sinónimos puede ser interpretado de formas distintas, como en el caso de las palabras "hurto"-"robar". Como puede observarse en la Figura **4-1**, la palabra "hurto" contribuye más negativamente a la categoría 1, mientras la palabra robar contribuye más negativamente a la categoría 4. Es decir, palabras con significados individuales similares pueden contribuir a categorías de PoS opuestas. Esto indica que, aunque estas palabras posiblemente representen lo mismo aisladas, para los ciudadanos pueden asociarse sentimientos diferentes.

Igualmente, es interesante observar el comportamiento de palabras de relevancia neutra como “mujer”, “bus” o “crimen” en la Figura **4-1** los cuales muestran contribuciones, principalmente negativas, para varias categorías de forma simultánea. Es posible que estos resultados se deban a elementos adicionales en los respectivos tweets donde estas palabras aparecen, o a los emisores del mensaje, ya que algunos elementos analizados provienen de las cuentas institucionales de los organismos de administración pública en Bogotá D.C.

La Figura **4-2** reporta las palabras que poseen una mayor relevancia (tanto positiva como negativa) en las categorías PoS estudiadas. Es decir, aquellos términos que contribuyen más para la clasificación de los tweets en estas categorías. Por ejemplo, el término operación en el nivel 5. Mientras que valores negativos indican que el termino contribuye a que el tweet no sea clasificado en esa categoría. Por ejemplo, el termino “violencia” contribuye a que los tweets no sean clasificados en el nivel 2.

Fuente: Elaboración propia.

Figura 4-2.: *Conjunto de palabras con mayor relevancia normalizada tanto positiva como negativamente para las categorías de PoS. Palabras con un alto nivel de contribución (positiva y negativa) aportar a la clasificación del tweet en la categoría correspondiente.*

Como puede observarse, en las categorías negativas (1-2) se observan términos despectivos y lenguaje soez, característicos de este tipo de mensajes. En la categoría neutra (3) aparecen términos como "denunciar" o actores políticos como "alcalde", indicando una posible bipolaridad en el uso de estas palabras. En los mensajes de las categorías positivas (4-5) se utiliza un lenguaje posiblemente asociado a operaciones policiales y administrativas, que posiblemente reflejan el accionar de estos actores hacía los eventos que las personas perciben como negativos.

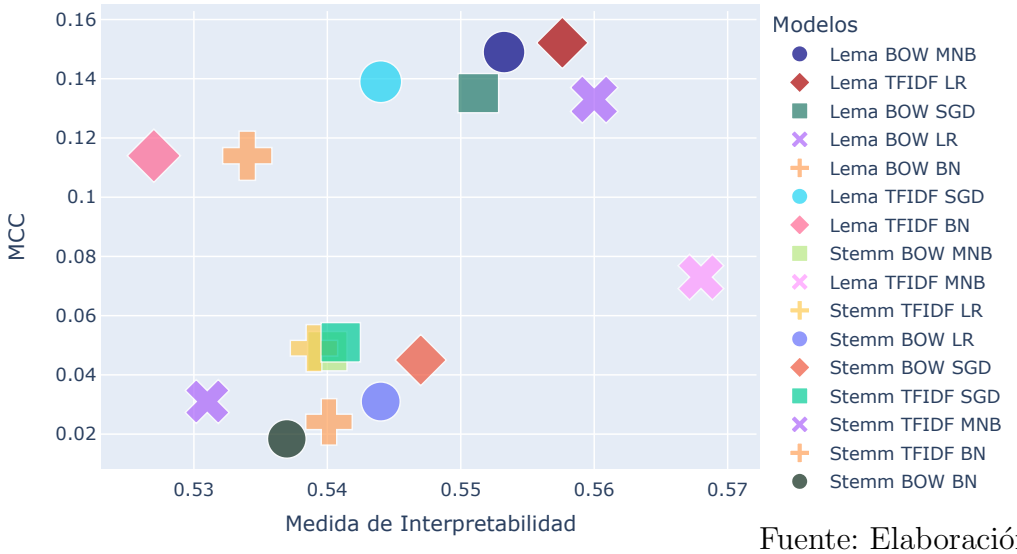
4.2. Selección del modelo

La figuras **4-3** y **4-4** reportan el comportamiento de los modelos de aprendizaje explorados respecto a las métricas de desempeño escogidas y la medida de interpretabilidad diseñada. En particular, MCC vs Medida de Interpretabilidad (ver Figura **4-4**) y F1 Score vs Medida de Interpretabilidad (ver Figura **4-4**. Es importante recordar que para la selección del modelo se consideró un criterio multi-objetivo, debido a que la medida de desempeño F1 Score indica el balance entre sensibilidad y especificidad, pero no considera la desvarianza de clase. En contraste, MCC tiene en cuenta la clasificación en sus respectivas categorias y el imbalance de clase, pero no considera la capacidad predictiva del modelo.

Como puede observarse los modelos conforman dos grupos de desempeño. En la parte inferior de figuras se observan los modelos basados en stemming (salvo la excepción del modelo Lema-TFIDF-MNB). Mientras que en la parte superior se encuentran modelos basados en lematización. Esto indica que para este conjunto de datos la lematización en la representación de los datos textuales en el idioma español puede contribuye al desempeño del modelo.

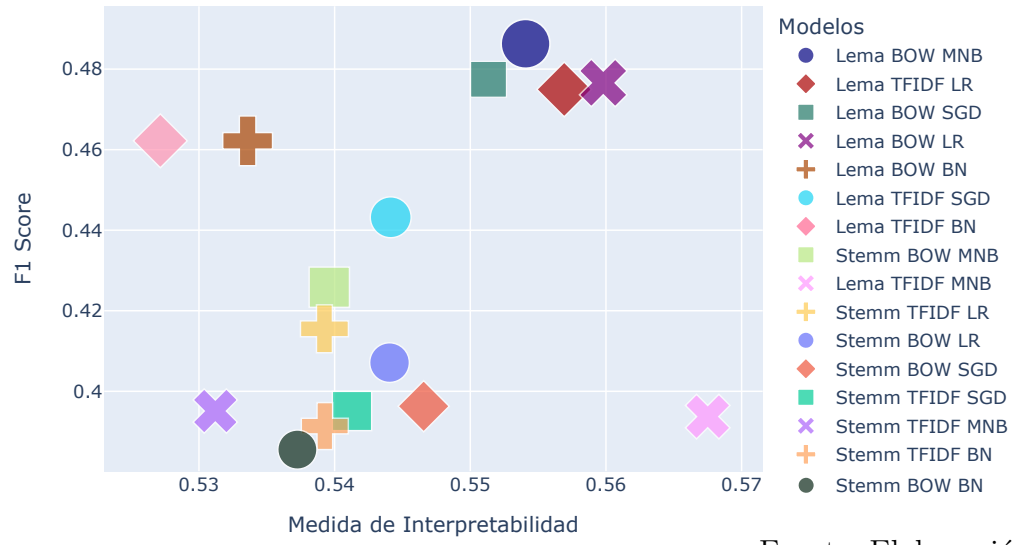
Teniendo en cuenta los dos esquemas de selección propuestos el modelo con el mejor desempeño fue LEMA BOW MNB. Sin embargo, otro modelo con desempeños destacables es

la Regresión Logística. Finalmente, los modelos que incluyen la clasificación NB Bernoulli muestran desempeños pobres, independientemente de la representación o extracción utilizada. Los resultados obtenidos de la comparación de modelos se reportan de forma detallada en el Anexo A.



Fuente: Elaboración propia.

Figura 4-3.: *Comparación de modelos respecto al MCC y su nivel de interpretabilidad.*

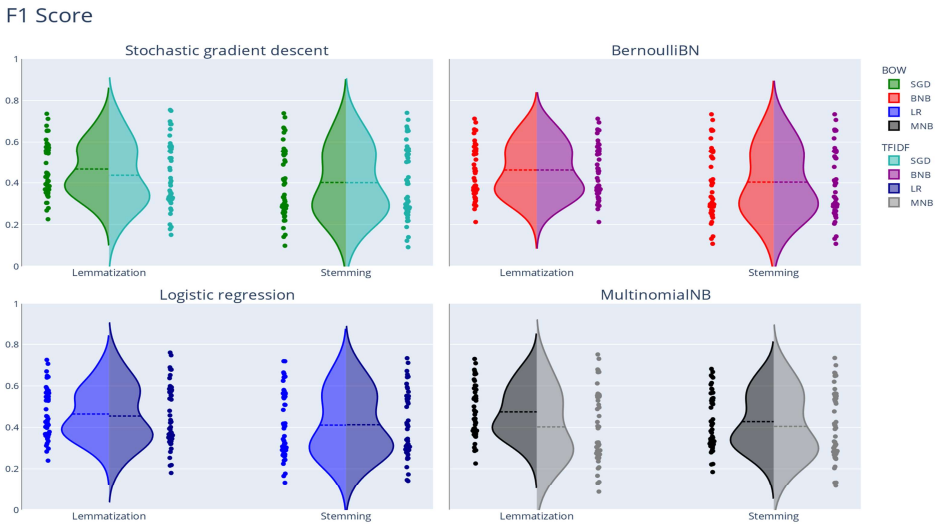


Fuente: Elaboración propia.

Figura 4-4.: *Comparación de modelos respecto al F1 Score y su nivel de interpretabilidad.*

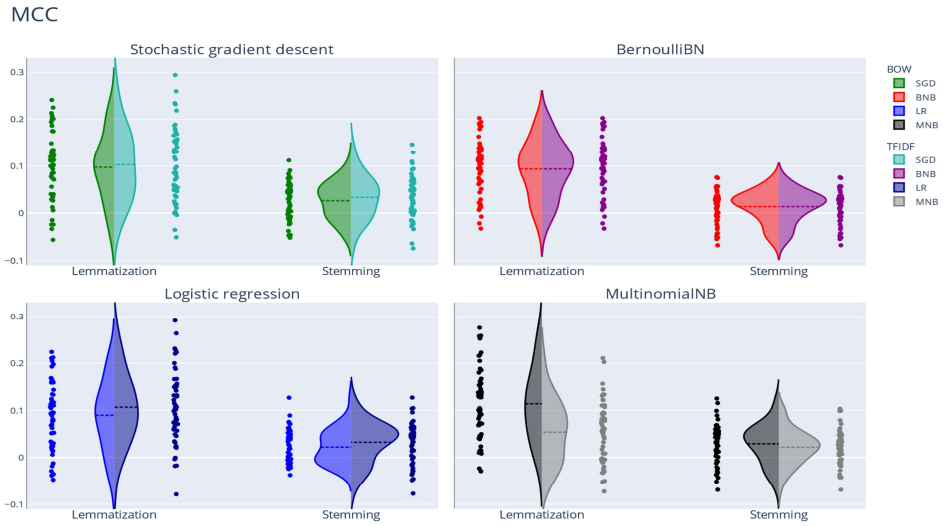
4.3. Métricas de Desempeño

La introducción de la estrategia de extracción de características conocida como lematización fue capaz de mejorar las métricas de desempeño para todos los clasificadores obtenidas en [Chaparro et al., 2021b, Chaparro et al., 2020] tal como se observa en las figuras **4-5** y **4-6**. Sin embargo, es útil mencionar que los parámetros dados en este trabajo difieren de los presentados en el artículo citado tanto en el número de particiones en el algoritmo de validación $k - fold$ de Scikit-Learn [Pedregosa et al., 2011] variando de 50 a 10 particiones e iterando 5 veces sobre diferentes parámetros de semillas los parámetros de semilla estableciéndolo con el valor $s = 1, 34, 22, 56, 7$ y shuffle como la variable booleana de mezcla $True$. Estos cambios se realizan con el fin de minimizar los posibles sobreajustes que puedan tenerse en el momento del entrenamiento y establecer la reproducibilidad de los resultados hallados.



Fuente: Elaboración propia.

Figura 4-5.: *Métrica de desempeño $F1-Score$ para los modelos estudiados.*



Fuente: Elaboración propia.

Figura 4-6.: *Métrica de desempeño MCC para los modelos estudiados.*

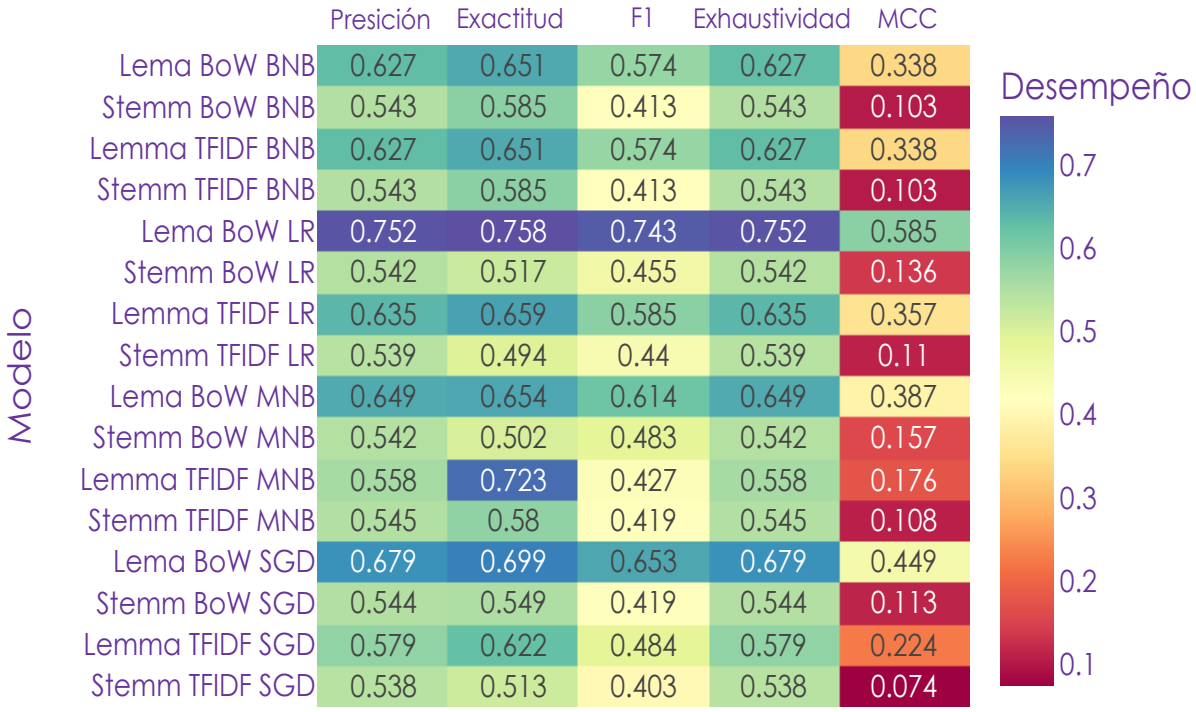
5. Conclusiones y recomendaciones

En el presente trabajo se introdujeron dos estrategias para mejorar comprensión de la interpretabilidad sobre los niveles de PoS: una medida de interpretabilidad categórica independiente del modelo y una medida de contribución normalizada de las categorías PoS. Los resultados obtenidos en las métricas de desempeño sugieren que el uso de estrategias de interpretabilidad puede mejorar el ajuste de los modelos y brindar interpretaciones de los contenidos de texto a representaciones vectoriales de una forma clara y más interpretable, que los métodos de interpretabilidad local y global establecidos. La medida de interpretabilidad propuesta permite definir un criterio de selección de modelos complementario a las métricas usuales, y hace posible evaluar cuantitativamente el comportamiento intrínseco al modelo respecto a su nivel de interpretabilidad. Los modelos analizados y el modelo seleccionado bajo esta estrategia permiten explicar características inherentes al sentimiento general de los ciudadanos hacia la percepción de seguridad.

En el transcurso de esta investigación se identificó la necesidad de una transformación del conjunto de datos diferente a la mostrada en [Chaparro et al., 2020]. Sin embargo, es posible utilizar más tipos de representaciones basadas en lematización, tales como Fasttext y Doc2vec entre otras. Sumado a esto el paquete de preprocesamiento de *ntlk* contiene un conjunto limitado de palabras vacías que podría complementarse con el análisis semántico de las partes de la oración o incluyendo de forma manual términos que no brinden información alguna

sobre la PoS sumado a un proceso de corrección semiautomática del conjunto de datos. A su vez, en [Chaparro et al., 2021b] se utilizaron muchos más tipos de clasificadores que podrían presentar mejores desempeños en este nuevo esquema y que en posibles nuevas investigaciones podrían abordarse de nuevo. Particularmente, el segundo modelo analizado *Lema-TFIDF-Regresión Logística* presentó desempeños ligeramente inferior al modelo seleccionado. Sin embargo, las metodología $k - fold$ de evaluación de Scikit-Learn presento valores donde ambos modelos eran sumamente similares -véase las figuras **4-5** y **4-6**- pero en condiciones diferentes de ajuste y validación del modelo en proporciones $80\,\% - 20\,\%$ y diferentes semillas de aleatoriedad este modelo logró resultados en ambas métricas que no se lograron en la validación cruzada, tal como se observa en la Figura **5-1** en uno de estos experimento. En este sentido, puede que ser necesario una modificación o reestructuración de la metodología de evaluación de desempeño.

Las medidas de interpretabilidad propuestas están sujetas al modelo base de comparación basado en el lexicón de polaridades ML-Senticon. En este diccionario de polaridades algunas características del conjunto de datos no reportaban polaridad, por tanto, la creación del vector de comparación posiblemente sea dispersa y tenga desviaciones a ajustar en próximas evaluaciones. Adicionalmente, la definición de la selección por modelo puede no ser lo suficientemente robusta para realizar tal elección, en investigaciones futuras es posible abordar alternativas complementarias.



Fuente: Elaboración propia.

Figura 5-1.: *Métricas de desempeño de los modelos analizados bajo el esquema de entrenamiento y validación en proporción (80 %- 20 %) con el paquete train_test_split de Scikit-Learn y semilla*= 1.

A. Anexo: Tabla de Resultados

La tabla **A-1** muestra los valores obtenidos para las diferentes métricas de desempeño (en este caso se presenta las medias obtenidas), medida de interpretabilidad y selecciones definidas en la sección 3.5.3.

Resultados Experimentales						
Modelos	$F1$-Score	MCC	M.I.[1]	Selección F1	Selección MCC	Selección Final
LEMA BOW MNB	0,48	0,14	0,55	0,69	0,40	0,554
LEMA TFIDF LR	0,47	0,14	0,55	0,69	0,40	0,549
LEMA BOW SGD	0,47	0,13	0,55	0,69	0,40	0,54
LEMA BOW LR	0,47	0,13	0,56	0,69	0,40	0,54
LEMA BOW BNB	0,46	0,11	0,53	0,67	0,38	0,52
LEMA TFIDF SGD	0,44	0,13	0,54	0,65	0,4	0,52
LEMA TFIDF BNB	0,46	0,11	0,52	0,66	0,38	0,52
STEMM BOW MNB	0,42	0,04	0,5	0,63	0,35	0,49
LEMA TFIDF MNB	0,39	0,07	0,56	0,60	0,37	0,48
STEMM TFIDF LR	0,41	0,04	0,53	0,62	0,35	0,48
STEMM LR BOW	0,40	0,03	0,54	0,61	0,34	0,48
STEMM SGD BOW	0,39	0,04	0,54	0,60	0,35	0,47
STEMM SGD TFIDF	0,39	0,05	0,54	0,60	0,35	0,47
STEMM TFIDF MNB	0,39	0,03	0,53	0,59	0,34	0,47
STEMM TFIDF BNB	0,39	0,02	0,53	0,59	0,34	0,46
STEMM BOW BNB	0,39	0,02	0,53	0,59	0,34	0,46

Tabla A-1.: Resultados de los modelos en todas las métrica evaluadas.

B. Anexo: Artículo de Ponencia.

A continuación se presenta el articulo anexo a la ponencia presentada en la Segunda Conferencia Colombiana de Matemáticas Aplicadas e Industriales (MAPI 2), que se llevó a cabo en Medellín–Colombia de la ponencia titulada *Interpretability on categories of perception of security* donde se presentaron algunos resultados preliminares de este trabajo.

Interpretability of categories of perception of security

Andrés Bermúdez
Departamento de Matemáticas
Universidad Nacional de Colombia
Bogotá, Colombia
anbermudezg@unal.edu.co

Francisco Gómez
Departamento de Matemáticas
Universidad Nacional de Colombia
Bogotá, Colombia
fagomezj@unal.edu.co

Luisa Chaparro
Departamento de Física
Instituto Tecnologico y
de Estudios Superiores de Monterrey
Monterrey, México
lchaparr@tec.mx

***Abstract*—The perception of security relates to citizens' feelings in the face of risk associated with security events and the magnitude of its consequences. Because of this subjective nature, it is a complex subject to quantify. Therefore, social networks emerged as an alternative to quantifying these opinions. Recently, multiclass supervised machine learning methods quantified different levels of security perception. However, these methods lack interpretability about why a group of tweets classifies in the same level of perception of security. This work proposes a novel strategy of interpretability for a group of predictions related to the same level of perception of security.**

***Index Terms*—Perception of Security (PoS), local and Categorial interpretability, Natural Language Processing (NPL), LIME.**

I. Introduction

The perception of security relates to citizens' feelings in the face of risk related to security events and the magnitude of its consequences [12] . This feeling is associated with the fear of crime [3], which is the emotional response people face when they are crime victims and may have severe negative societal consequences. The perception of security changes in time and space [17] and depends on individual circumstances, and experiences people suffer. Because of this subjective nature, it is a complex subject to quantify. Surveys on citizens' opinions represent the most widely used alternative to quantifying this feeling [2]. However, they are not well adapted to the dynamic nature of the perception of security and commonly focus only on the victimization levels.

Social networks emerged as an alternative to quantifying these opinions [13]. Social networks allow transmitting events and news related to different fields in real-time, including citizens' views about security [3]. Moreover, social networks' rapidly spread this content [7], making them valuable sources for observing dynamic data that may help understand how people's PoS changes over time. Based on this observation, different works used Twitter content to characterize PoS quantitatively [4]-[6]-[13]-[16].

Recently, multiclass supervised machine learning methods quantified different levels of security perception based on Twitter posts [5]. These methods aim to explain the different levels of sentiment (1: very negative, 2: negative, 3: neutral, 4: positive, 5: very positive) related to security underlying twitter posts. However, because complex machine learning classifiers perform these quantifications, there is no human interpretation of *why particular perceptions of security were predicted out of the content?*. Recently, techniques of local interpretability explored this question at the individual prediction level, i.e., a single tweet. In particular, a local interpretable model-agnostic explanation (LIME) provided individual posts' interpretations for machine learning classifiers. However, there is no interpretation of *why a group of tweets classifies in the same level of perception of security?*.

This work proposes a novel strategy of interpretability for a group of predictions related to the same level of perception of security. For this, first, we construct an explanation of individual classifications based on LIME [11]. Then, for each sentiment group, these explanations are aggregated into 1) important explainable words for each category of perception and 2) relevance of categories of perception across words.

II. Materials and methods

Figure 1 illustrates the proposed method. First, a database of labeled tweets is constructed. Later, different supervised classification methods are trained to predict the level of PoS. These models were trained using two different vectorizations for representing the textual data. A local interpretability scheme (LIME) based on linear regression (LASSO) was used to provide interpretability for each tweet. Linear contributions of each word provided by LIME were then combined for tweets in the same PoS level to provide interpretations for each category.

A. Data

The analyzed content comes from the social network Twitter. The database contains 26.255 tweets geolocated in the city of Bogotá D.C, Colombia. This database was filtered to obtain those tweets related to security issues and subsequently unbalanced classified into five levels of perception (or feeling) by experts belonging to the District Secretary of Security and the University National of Colombia [4].

B. Preprocessing

For pre-processing and due to the nature of tweets, links, mentions, and hashtags were removed per [14]. In addition, the texts were normalized by eliminating capital letters and

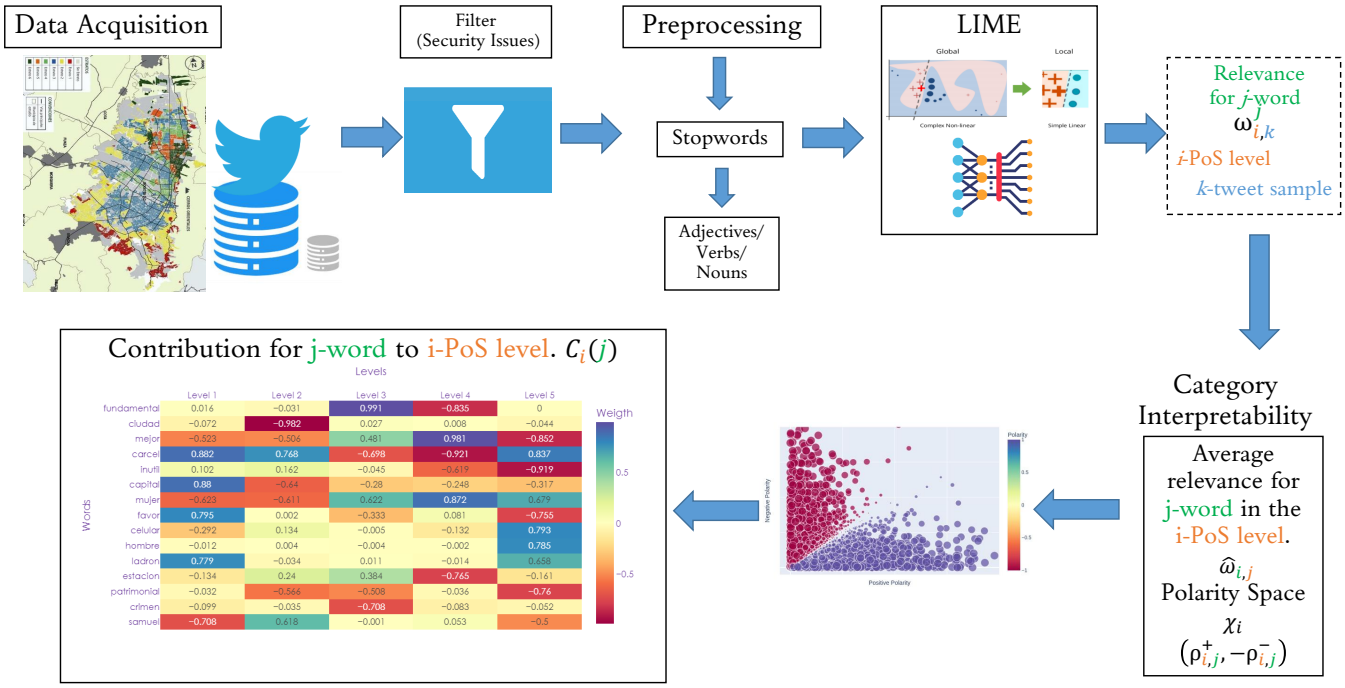


Fig. 1. Schematic Representation of the Interpretability of Sentiment Analysis over Tweets. Start with data acquisition followed by a filter of relevant data. The data pass over a preprocessing step to implement the different sentiment Analysis approaches. As a result, a final classification: positive, negative, or neutral for each Tweet. In order to understand the perception of Security, a Local Interpretability was made over the classifier.Finally, we define the average relevance, the polarity (with its respective representation space) and thus establish the contribution of the word to the PoS level.

punctuation marks (both initial and final), considering that they are texts in Spanish. Additionally, the stopwords [1] have been eliminated. A database without noisy words resulted from this process suitable for tokenization and vectorization to extract features and subsequent sentiment analysis using machine learning.

C. Supervised Learning

1) Features Extraction: For the supervised machine learning methods, two models were used to determine the dictionaries: Bag of Words (BoW) [15] and Term frequency – Inverse document frequency (TF-IDF) [10]. A similar approach was previously used in [6].

2) Classification Methods: After obtaining the dictionaries, different classification methods were tested, including Multinomial Bayes Naives (MNB), Logistic Regression (LR), Naive Bayes classifier for multivariate Bernoulli models (BNB), and stochastic gradient descent (SGD). Finally, a (Ruled Based) Lexicon was used as baseline [6].

A K-fold cross-validation scheme was used for comparison [8]. The F1-score was used to determine the performance of each model+dictionary pair.

D. Local Interpretability

Understanding the reasons behind the predictions of all implemented classifiers is essential for assessing confidence and guiding how to choose one classifier over another. This work uses an explanation technique that describes the predictions of any classifier in an interpretable way by learning a locally interpretable model around the prediction. An approach is known as LIME (Local Interpretable Model-Agnostic Explanations) [11].

LIME focuses on training local surrogate models to explain individual predictions. The goal is to understand why the machine learning model made a particular prediction. LIME checks what happens to the predictions when data variations are fed into the machine learning model [9].

For each prediction to explain, LIME permutes the observation n times. It then predicts the outcome of all the permuted observations and computes the distance of all permutations from the original observation, converting the distance to a similarity score. After LIME selects m features, those features best describe the result of the complex model of the permuted data. It then fits a simple model to the permuted data, explaining the model output with the m features of the permuted data weighted by their similarity score. Finally, LIME extracts the weights of the features from the simple model and uses these features as explanations for the local behavior models.

E. Interpretability for categories

To define interpretability at the category level (PoS level), let $\omega^{j}_{i,k}$ the *relevance* of the j-th word at the i-th PoS level for

the k-th tweet as long as the word belongs to the tweet. The value of $\omega_{i,k}^{j}$ will be provided by LIME used in each tweet. Therefore, we can define:

$$W_i^j = \left\{\omega_{i,k}^{j}\right\}_{\substack{j\in k \\ k\in Tweets}}$$

as the set of all coefficients of the j-th word for the PoS level i. Furthermore, we can define the *average relevance* for the j-th word in the PoS i as the average as:

$$\widehat{\omega}_{i,j} = \frac{\sum\limits_{\omega\in W_i^j} \omega}{\left|W_i^j\right|}$$

It is possible to define both the positive and negative *polarity* of each of the words in the corresponding PoS levels. For this, lets define the *positive polarity* ($\rho_{i,j}^{+}$) and the *negative polarity* ($\rho_{i,j}^{-}$) as:

$$\rho_{i,j}^{+} = \frac{1}{\left|I(W_i^j)\right|}\sum_{\substack{\omega\in W_i^j \\ \omega>0}} \omega, \qquad \rho_{i,j}^{-} = \frac{1}{\left|I(-W_i^j)\right|}\sum_{\substack{\omega\in W_i^j \\ \omega<0}} \omega$$

Note that the expression $|I(X)|$ corresponds to the norm of the indicator function and represents the number of elements of the set X that are positive.

Using polarities a two-dimensional representation space can be established -see figure 2- for each PoS level (χ_i), also called the *polarity space*, as follows:

$$\chi_i = \left\{\left(\rho_{i,j}^{+}, -\rho_{i,j}^{-}\right) | \forall j \in Words\right\}$$

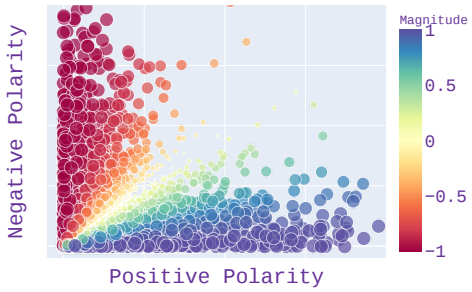


Fig. 2. Polarity space for the PoS level 1 under for the multinomial BOW model.

Finally, it is possible to define a *contribution level* for each word present in the PoS i, taking into account its frequency, polarity and average relevance, as follows:

$$\begin{aligned} C_i(j) &= \left|\left(\frac{2}{1+e^{-k\cdot\widehat{r}_{i,j}}} - 1\right)\right| \cdot \cos(2\cdot\theta(\rho_{i,j}^{+}, -\rho_{i,j}^{-})) \cdot \left|W_i^j\right| \\ &= \left|\left(\frac{1-e^{-k\cdot\widehat{r}_{i,j}}}{1+e^{-k\cdot\widehat{r}_{i,j}}}\right)\right| \cdot \left(\frac{(\rho_{i,j}^{+})^2 - (\rho_{i,j}^{-})^2}{(\rho_{i,j}^{+})^2 + (\rho_{i,j}^{-})^2}\right) \cdot |W_j^i| \end{aligned}$$

Note that the use of the sigmoid transformation given to the average relevance is carried out to control the ranges in which these values are presented, making an injection to the interval $(-1, 1)$ that preserves the signs of each of these.

III. Results

Figure 3 reports the F1 performances for the different classification methods used to quantify PoS at different levels. The multinomial NB provided the highest performance in the classification tasks.

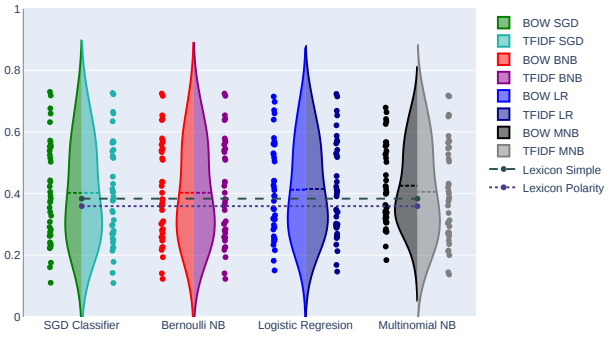


Fig. 3. F1 score for all the pairs model+dictionary compared to the results of Lexicon.

Figure 4 shows the contribution value of a set of different words for different levels of PoS for the model with the highest performance. These words were selected because of their relationship with security. As observed, the proposed interpretability model based on polarity reflects the association between sentiment and the level of PoS. For instance, the word *ladron* positively contributed to the lowest PoS level (level 1). Interestingly, words as *verdad* contribute to the highest and lowest levels of PoS.

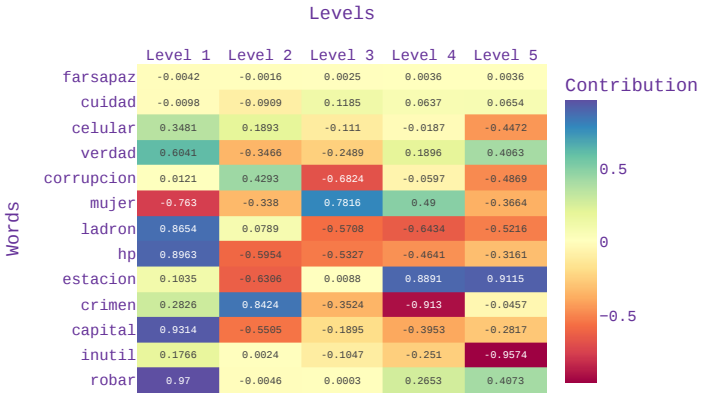


Fig. 4. Contribution to PoS of a set of words relevant in the security context.

Figure 5 shows the words with the highest and lowest contributions for each PoS level. The words that define different PoS differ across different PoS levels. The level of PoS is

defined by terms contributing positively and negatively to the category.

Fig. 5. Most important words in the prediction of five categories of perception using the multinomial Bayes classifier. Positive (blue) and negative (red) contributions to each category are indicated with the color scale.

IV. Conclusions

This work introduces a strategy of interpretability for a group of predictions related to the same level of perception of security. The proposed strategy provides positive and negative contributions of terms for different levels of PoS.

However, the behavior of interpretability on classifiers with similar behavior (both local and categorical) in the performance metrics that were studied in previous works still needs to be explored. Interpretability models open a door of transparency to the black-boxes of the classifiers, allowing better confidence in the prediction models.

References

[1] Surya Bhagvat. "Clustering of twitter technology tweets and the impact of stopwords on clusters". In: (2011).

[2] Russell G Brooker and Todd Schaefer. "Methods of measuring public opinion". In: *Public opinion in the 21st century. https://www. uky. edu/AS/PoliSci/Peffley/pdf/473Measuring% 20Public% 20Opinion. pdf* (2015).

[3] Mary Ellen Brown, Patricia A Dustman, and Juan J Barthelemy. "Twitter impact on a community trauma: An examination of who, what, and why it radiated". In: *Journal of community psychology* 49.3 (2021), pp. 838–853.

[4] Luisa Fernanda Chaparro et al. "Interpretability Of The Perception Of Security Based On Tweets Content". In: *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. IEEE. 2021, pp. 1–6.

[5] Luisa Fernanda Chaparro et al. "Quantifying Perception of Security Through Social Media and Its Relationship With Crime". In: *IEEE Access* 9 (2021), pp. 139201–139213.

[6] Luisa Fernanda Chaparro et al. "Sentiment analysis of social network content to characterize the perception of security". In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2020, pp. 685–691.

[7] Akshay Java et al. "Why we twitter: An analysis of a microblogging community". In: *International Workshop on Social Network Mining and Analysis*. Springer. 2007, pp. 118–138.

[8] Ling Liu and M Tamer Özsu. *Encyclopedia of database systems*. Vol. 6. Springer, 2009.

[9] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book.

[10] Shahzad Qaiser and Ramsha Ali. "Text mining: use of TF-IDF to examine the relevance of words to documents". In: *International Journal of Computer Applications* 181.1 (2018), pp. 25–29.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[12] Torbjørn Rundmo and Bjørg-Elin Moen. "Risk perception and demand for risk mitigation in transport: A comparison of lay people, politicians and experts". In: *Journal of Risk research* 9.6 (2006), pp. 623–640.

[13] Barbara Schultz-Jones. "Examining information behavior through social networks: An interdisciplinary review". In: *Journal of Documentation* (2009).

[14] NFF Silvaa, ER Hruschkaa, and ER Hruschka. "Tweet sentiment analysis with classifier ensembles". In: *Decision Support Systems* 66 (2014), pp. 170–179.

[15] K Soumya George and Shibily Joseph. "Text classification by augmenting bag of words (BOW) representation with co-occurrence feature". In: *IOSR Journal of Computer Engineering* 16.1 (2014), pp. 34–38.

[16] Jorge Victorino et al. "Spatial-temporal patterns of aggressive behaviors. A case study Bogotá, Colombia". In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2020, pp. 667–672.

[17] Romika Yadav and Savita Kumari Sheoran. "Crime prediction using auto regression techniques for time series data". In: *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*. IEEE. 2018, pp. 1–5.

Bibliografía

[Agirre et al., 1992] Agirre, E., Alegria, I., Arregi, X., Artola, X., de Ilarraza, A. D., Maritxalar, M., Sarasola, K., and Urkia, M. (1992). Xuxen: A spelling checker/corrector for basque based on two-level morphology. In *Third Conference on Applied Natural Language Processing*, pages 119–125.

[Anguiano-Hernández, 2009] Anguiano-Hernández, E. (2009). Naive bayes multinomial para clasificación de texto usando un esquema de pesado por clases.

[Ayodele, 2010] Ayodele, T. O. (2010). Types of machine learning algorithms in new advances in machine learning. croatia, rijeka.

[Balakrishnan and Lloyd-Yemoh, 2014] Balakrishnan, V. and Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances.

[Barreras et al., 2016] Barreras, F., Diaz, C., Riascos, A., and Ribero, M. (2016). Comparison of different crime prediction models in bogotá. 2016.

[Bendler et al., 2014] Bendler, J., Brandt, T., Wagner, S., and Neumann, D. (2014). Investigating crime-to-twitter relationships in urban environments-facilitating a virtual neighborhood watch. *Association for Information Systems (AIS) eLibrary*.

[Bhagvat, 2011] Bhagvat, S. (2011). Clustering of twitter technology tweets and the impact of stopwords on clusters.

[Bird, 2006] Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

[Bishop and Nasrabadi, 2006] Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

[Bokinsky et al., 2013] Bokinsky, H., McKenzie, A., Bayoumi, A., McCaslin, R., Patterson, A., Matthews, M., Schmidley, J., and Eisner, L. (2013). Application of natural language processing techniques to marine v-22 maintenance data for populating a cbm-oriented database. In *AHS Airworthiness, CBM, and HUMS Specialists' Meeting, Huntsville, AL*.

[Bottou, 2010] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

[Brooker and Schaefer, 2015] Brooker, R. G. and Schaefer, T. (2015). Methods of measuring public opinion. *Public opinion in the 21st century. https://www. uky. edu/AS/PoliSci/Peffley/pdf/473Measuring % 20Public % 20Opinion. pdf.*

[Brown et al., 2021] Brown, M. E., Dustman, P. A., and Barthelemy, J. J. (2021). Twitter impact on a community trauma: An examination of who, what, and why it radiated. *Journal of community psychology*, 49(3):838–853.

[Cámara de Comercio de Bogotá, 2022] Cámara de Comercio de Bogotá, h. (2022). Encuesta de percepción y victimización de bogotá-2021.

[Camargo et al., 2016] Camargo, J. E., Torres, C. A., Martínez, O. H., and Gómez, F. A. (2016). A big data analytics system to analyze citizens' perception of security. In *2016 IEEE International Smart Cities Conference (ISC2)*, pages 1–5. IEEE.

[Cambria et al., 2022] Cambria, E., Xing, F., Thelwall, M., and Welsch, R. (2022). Sentiment analysis as a multidisciplinary research area. *IEEE Transactions on Artificial Intelligence*, 3(2):1–3.

[Carvalho et al., 2019] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

[Chaparro et al., 2020] Chaparro, L. F., Pulido, C., Rudas, J., Reyes, A., Victorino, J., Narváez, L. a., Gómez, F., and Martinez, D. (2020). Sentiment analysis of social network content to characterize the perception of security. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 685–691. IEEE.

[Chaparro et al., 2021a] Chaparro, L. F., Pulido, C., Rudas, J., Reyes, A. M., Victorino, J., Narváez, L., Martinez, D., and Gómez, F. (2021a). Interpretability of the perception of security based on tweets content. In *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–6. IEEE.

[Chaparro et al., 2021b] Chaparro, L. F., Pulido, C., Rudas, J., Victorino, J., Reyes, A. M., Estrada, C., Narvaez, L. A., and Gómez, F. (2021b). Quantifying perception of security through social media and its relationship with crime. *IEEE Access*, 9:139201–139213.

[Chen et al., 2015] Chen, X., Cho, Y., and Jang, S. Y. (2015). Crime prediction using twitter sentiment and weather. In *2015 systems and information engineering design symposium*, pages 63–68. IEEE.

[Chicco and Jurman, 2020] Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

[Chowdhary, 2020] Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.

[Chowdhury and Chowdhury, 2003] Chowdhury, G. G. and Chowdhury, S. (2003). *Introduction to digital libraries*. Facet publishing.

[Christopher et al., 2008] Christopher, D. M., Prabhakar, R., and Hinrich, S. (2008). Introduction to information retrieval.

[Cruz et al., 2014] Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Ml-senticon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, 53:113–120.

[Curiel and Bishop, 2017] Curiel, R. and Bishop, S. (2017). Modelling the fear of crime. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 473:20170156.

[Cvetojevic and Hochmair, 2018] Cvetojevic, S. and Hochmair, H. H. (2018). Analyzing the spread of tweets in response to paris attacks. *Computers, Environment and Urban Systems*, 71:14–26.

[Da Silva et al., 2014] Da Silva, N. F., Hruschka, E. R., and Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision support systems*, 66:170–179.

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[Drakulich, 2015] Drakulich, K. M. (2015). Social capital, information, and perceived safety from crime: The differential effects of reassuring social connections and vicarious victimization. *Social Science Quarterly*, 96(1):176–190.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

[Gaisbauer et al., 2021] Gaisbauer, F., Pournaki, A., Banisch, S., and Olbrich, E. (2021). Ideological differences in engagement in public debate on twitter. *Plos one*, 16(3):e0249241.

[Gerber, 2014] Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

[Ghosh et al., 2012] Ghosh, S., Roy, S., and Bandyopadhyay, S. K. (2012). A tutorial review on text mining algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 1(4):7.

[Grandini et al., 2020] Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

[Hirschberg and Manning, 2015] Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

[Hollis et al., 2017] Hollis, M. E., Downey, S., Del Carmen, A., and Dobbs, R. R. (2017). The relationship between media portrayals and crime: perceptions of fear of crime among citizens. *Crime prevention and community safety*, 19(1):46–60.

[Hooker, 2004] Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580.

[Hooker, 2007] Hooker, G. (2007). Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732.

[Hummelsheim et al., 2011] Hummelsheim, D., Hirtenlehner, H., Jackson, J., and Oberwittler, D. (2011). Social insecurities and fear of crime: A cross-national study on the impact of welfare state policies on crime-related anxieties. *European sociological review*, 27(3):327–345.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: An analysis of a microblogging community. In *International Workshop on Social Network Mining and Analysis*, pages 118–138. Springer.

[Jurman et al., 2012] Jurman, G., Riccadonna, S., and Furlanello, C. (2012). A comparison

of mcc and cen error measures in multi-class prediction.

[Kaur and Buttar, 2018] Kaur, J. and Buttar, P. K. (2018). A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4):207–210.

[Ketkar, 2017] Ketkar, N. (2017). Stochastic gradient descent. In *Deep learning with Python*, pages 113–132. Springer.

[Kim et al., 2016] Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

[Kleck and Barnes, 2014] Kleck, G. and Barnes, J. C. (2014). Do more police lead to more crime deterrence? *Crime & Delinquency*, 60(5):716–738.

[Kounadi et al., 2015] Kounadi, O., Lampoltshammer, T. J., Groff, E., Sitko, I., and Leitner, M. (2015). Exploring twitter to analyze the public's reaction patterns to recently reported homicides in london. *PloS one*, 10(3):e0121848.

[Ladani and Desai, 2020] Ladani, D. J. and Desai, N. P. (2020). Stopword identification and removal techniques on tc and ir applications: A survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 466–472. IEEE.

[Latané, 1981] Latané, B. (1981). The psychology of social impact. *American psychologist*, 36(4).

[Le and Mikolov, 2014] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

[Luque et al., 2019] Luque, C., Luna, J. M., Luque, M., and Ventura, S. (2019). An advanced

review on text mining in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1302.

[Madasamy and Ramaswami, 2017] Madasamy, K. and Ramaswami, M. (2017). Data imbalance and classifiers: Impact and solutions from a big data perspective. *International Journal of Computational Intelligence Research*, 13(9):2267–2281.

[Malleson and Andresen, 2015] Malleson, N. and Andresen, M. A. (2015). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2):112–121.

[Matthews, 1975] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

[Messalas et al., 2019] Messalas, A., Kanellopoulos, Y., and Makris, C. (2019). Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–7. IEEE.

[Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

[Molnar, 2022] Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

[Naser and Alavi, 2020] Naser, M. and Alavi, A. (2020). Insights into performance fitness and error metrics for machine learning. *arXiv preprint arXiv:2006.00887*.

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. J. (2006). Numerical optimization 2nd edition.

[Ordoñez-Eraso et al., 2020] Ordoñez-Eraso, H.-A., Pardo-Calvache, C.-J., and Cobos-Lozada, C.-A. (2020). Detection of homicide trends in colombia using machine learning. *Learning*, 29(54):e11740.

[Pabón et al., 2020] Pabón, J. S. M., Rubio, M. D., Castaño, Y., Riascos, A. J., and Díaz, P. R. (2020). A manifold learning data enrichment methodology for homicide prediction. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pages 1–4. IEEE.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Porter, 2001] Porter, M. F. (2001). Snowball: A language for stemming algorithms.

[Prathap and Ramesha, 2018] Prathap, B. R. and Ramesha, K. (2018). Twitter sentiment for analysing different types of crimes. In *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pages 483–488. IEEE.

[Prieto Curiel and Bishop, 2016] Prieto Curiel, R. and Bishop, S. R. (2016). A metric of the difference between perception of security and victimisation rates. *Crime Science*, 5(1):1–15.

[Prieto Curiel et al., 2020] Prieto Curiel, R., Cresci, S., Muntean, C. I., and Bishop, S. R. (2020). Crime and its fear in social media. *Palgrave Communications*, 6(1):1–12.

[Pulido et al., 2021] Pulido, C., Chaparro, L. F., Rudas, J., Reyes, A. M., Victorino, J., Narváez, L. Á., Martínez, D., and Gómez, F. (2021). Data filtering and classification for the identification of texts related to security in bogotá colombia. 7th International Conference on Computational Social Science IC2S2 2021.

[Pulido et al., 2019] Pulido, C., Prieto, J., and Gómez, F. (2019). How the social interactions in communities affect the fear of crime. *Systems Research and Behavioral Science*, 36(6):789–798.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

[Refaeilzadeh et al., 2009] Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, 5:532–538.

[Reyes et al., 2020] Reyes, A. M., Rudas, J., Pulido, C., Victorino, J., Martínez, D., Narváez, L. Á., and Gómez, F. (2020). Characterization of temporal patterns in the occurrence of aggressive behaviors in bogotá (colombia). In *2020 7th International conference on behavioural and social computing (BESC)*, pages 1–4. IEEE.

[Ribeiro et al., 2016a] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Ribeiro et al., 2016b] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

[Rundmo and Moen, 2006] Rundmo, T. r. and Moen, B. r.-E. (2006). Risk perception and demand for risk mitigation in transport: A comparison of lay people, politicians and experts. *Journal of Risk research*, 9(6):623–640.

[Rutjens and Brandt, 2019] Rutjens, B. T. and Brandt, M. J. (2019). *Belief systems and the perception of reality*. Routledge London, UK.

[Sánchez, 2008] Sánchez, M. (2008). La percepción de seguridad y la realidad social. *Cua-*

dernos de seguridad, 219.

[Schultz-Jones, 2009] Schultz-Jones, B. (2009). Examining information behavior through social networks: An interdisciplinary review. *Journal of Documentation*.

[Soumya George and Joseph, 2014] Soumya George, K. and Joseph, S. (2014). Text classification by augmenting bag of words (bow) representation with co-occurrence feature. *IOSR Journal of Computer Engineering*, 16(1):34–38.

[Victorino et al., 2020] Victorino, J., Rudas, J., Reyes, A. M., Pulido, C., Chaparro, L. F., Narváez, L. A., Martinez, D., and Gómez, F. (2020). Spatial temporal patterns of aggressive behaviors. a case study bogotá, colombia. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 685–691. IEEE.

[Villegas et al., 2022] Villegas, Á. J. R., Pabón, J. S. M., Rubio, M. D., Quintero, S., Vargas, J. G., and García, H. (2022). Spatio temporal sparsity in homicide prediction models. *IEEE Access*, 10:14359–14367.

[Wang et al., 2012] Wang, X., Gerber, M. S., and Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer.

[Wang et al., 2021] Wang, Z. J., Kale, A., Nori, H., Stella, P., Nunnally, M., Chau, D. H., Vorvoreanu, M., Vaughan, J. W., and Caruana, R. (2021). Gam changer: Editing generalized additive models with interactive visualization. *arXiv preprint arXiv:2112.03245*.

[Yadav and Sheoran, 2018] Yadav, R. and Sheoran, S. K. (2018). Crime prediction using auto regression techniques for time series data. In *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–5. IEEE.

[Yadav and Sarkar, 2018] Yadav, S. and Sarkar, M. (2018). Enhancing sentiment analysis

using domain-specific lexicon: A case study on gst. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1109–1114. IEEE.

[Zhang et al., 2015] Zhang, Y., Chen, M., and Liu, L. (2015). A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 681–685. IEEE.

[Zhu, 2020] Zhu, Q. (2020). On the performance of matthews correlation coefficient (mcc) for imbalanced dataset. *Pattern Recognition Letters*, 136:71–80.

[Zschech et al., 2022] Zschech, P., Weinzierl, S., Hambauer, N., Zilker, S., and Kraus, M. (2022). Gam (e) changer or not? an evaluation of interpretable machine learning models based on additive model constraints. *arXiv preprint arXiv:2204.09123*.