ANÁLISIS DE LA CONTRIBUCIÓN DE LOS FONEMAS A LA PREDICCIÓN DE LA VALENCIA EMOCIONAL EN TWEETS EN ESPAÑOL E INGLÉS

TRABAJO DE GRADO PARA OPTAR POR EL TÍTULO DE
MAGISTER EN ANALÍTICA PARA LA INTELIGENCIA DE NEGOCIOS

AUTOR
GABRIEL ALEJANDRO BERNAL ROJAS

DIRECTOR
ING. JORGE ANDRÉS ALVARADO VALENCIA
PhD EN INGENIERÍA, MS EN ANALYTICS, MS EN EDUCACIÓN

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN ANALÍTICA PARA LA INTELIGENCIA DE NEGOCIOS
BOGOTÁ, D.C.

2020

TABLA DE CONTENIDO

Índice de Tablas

Índice de Figuras

Resumen

Aunque tradicionalmente se ha asumido que el sonido de las palabras y su significado se relacionan de forma arbitraria, distintos hallazgos empíricos respaldan la hipótesis de que las unidades fonológicas básicas del lenguaje guardan una relación sistemática con aspectos semánticos, incluyendo la connotación afectiva y actitudinal de las palabras (Adelman, Estes, & Cossu, 2018; Aryani, Conrad, Schmidtke, & Jacobs, 2018; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Monaghan, Shillcock, Christiansen, & Kirby, 2014; Schmidtke, Conrad, & Jacobs, 2014). A partir de estas premisas, se buscó identificar si las unidades fonológicas del español y el inglés contribuyen a la predicción de la valencia emocional en un corpus de tweets. Para esto, se entrenó un conjunto de modelos de regresión lineal múltiple, cuyo desempeño fue evaluado a partir de la correlación y los indicadores de error calculados partir de las valencias predichas y las observadas en los *datasets* de prueba proporcionados por el concurso SemEval-2018 (Mohammad, Bravo-Márquez, Salameh, & Kiritchenko, 2018). Se encontró que la adición de los recursos fonológicos a un conjunto de predictores léxicos (*Bag of Words* de los Tweets, normalizada con el método TF-IDF) tiene un efecto reducido pero consistente sobre las métricas globales de ajuste, y en ambos idiomas permite discriminar con mayor precisión las valencias observadas cercanas a los valores medios, así como las valencias inferiores asociadas a contenidos afectivos negativos.

Este proyecto representa un aporte para la discusión teórica que sustenta las técnicas de análisis de sentimiento, poniendo a prueba hipótesis derivadas de planteamientos psicolingüísticos, en relación con el concepto de simbolismo de los sonidos (*sound symbolism*) o iconicidad fonológica. Los hallazgos de este proyecto tienen relevancia aplicada, dado el creciente uso de estas técnicas en distintos ámbitos, como el estudio de las actitudes y las emociones en escenarios electorales, de servicio al cliente o de diagnóstico de la salud mental, entre otros.

Palabras clave: Arbitrariedad del signo lingüístico, fonema, modelos de regresión lineal, procesamiento del lenguaje natural, análisis de sentimiento, valencia emocional.

Introducción

Planteamiento del problema

En esencia, las palabras son combinaciones de letras (vocales y consonantes), las cuales, al representar ciertos sonidos, constituyen las unidades sub léxicas de un idioma. Así, las palabras remiten a un signo lingüístico conformado por un significado (su representación mental) y un significante (su representación acústica). Desde una perspectiva tradicional se ha considerado que la relación entre significante y significado es arbitraría, como lo planteó de Saussure hace un siglo (1916/2005), es decir "sonidos sin sentido se combinan entre sí para crear palabras con sentido".

No obstante, se ha discutido la posibilidad de que los sonidos asociados a las palabras estén vinculados con los aspectos semánticos del lenguaje, pues ha venido acumulándose un cuerpo creciente de evidencia empírica que pone de manifiesto diferentes interacciones entre sonido y significado. De especial interés son las relaciones entre los contenidos afectivos de las palabras y su representación acústica, un caso especial de modalidad cruzada que implica interacciones entre las emociones y el lenguaje, asociaciones enmarcadas dentro del fenómeno conocido en la literatura como simbolismo de los sonidos (*sound symbolism*) o iconicidad fonológica (Majid, 2012; Monaghan, Shillcock, Christiansen, & Kirby, 2014; Rummer, Schweppe, Schlegelmilch, & Grice, 2014; Schmidtke, Conrad, & Jacobs, 2014).

Por otro lado, gracias a la expansión de la web 2.0, hoy en día se cuenta con una enorme cantidad de datos que además se producen en tiempo real, y que incluyen información no estructurada con contenidos pictóricos, audiovisuales y de texto (Paredes, Colomo, Salas, & Valencia, 2017; Song, Kim, Lee, Kim, & Youn, 2017; Yadollahi, Shahraki, & Zaiane, 2017). Esta situación ha promovido el desarrollo de técnicas analíticas y computacionales que buscan procesar de forma automatizada grandes volúmenes de información desde los paradigmas del *Big Data* y el *Machine Learning*, dentro de las cuales se destaca el llamado análisis de sentimientos u opiniones, que busca identificar y clasificar contenidos afectivos a partir del análisis de textos.

Específicamente, la minería de emociones a partir del análisis de información textual tiene un gran potencial por sus posibles aplicaciones en campos como la medición de la satisfacción de los consumidores, la selección de materiales para la enseñanza electrónica, el diseño de sistemas de recomendación en las interacciones hombre-computador, la predicción de ventas de productos específicos o incluso el diagnóstico de trastornos mentales (Yadollahi, Shahraki, & Zaiane, 2017). Estas técnicas se basan en el análisis de diferentes componentes del lenguaje oral o escrito,

tomando como unidades básicas palabras o combinaciones de estas en frases, oraciones y párrafos, así como elementos gramaticales y sintácticos de la lengua. En todo caso, se ha explorado poco la posibilidad de incluir en los algoritmos de Procesamiento del Lenguaje Natural (PNL) unidades sub léxicas fonológicas, en coherencia con los hallazgos descritos al comienzo de este documento.

Desde la perspectiva analítica, Soleymani *et al.* (2017) mencionan la importancia de un enfoque multimodal para las tareas de análisis de sentimiento, es decir, un enfoque que considere información de distintos dominios (por ejemplo, lenguaje hablado y escrito, además de imágenes y videos provenientes de interacciones humano-máquina o humano-humano), teniendo en cuenta que las emociones humanas se expresan a través de una variedad de canales (ver Tabla 1). No obstante, a pesar de las aplicaciones potenciales de un enfoque multimodal, esta aproximación apenas está comenzando a desarrollarse. Algunos ejemplos de aplicaciones de análisis de sentimiento multimodal son el análisis de contenidos de *bloggeros* y *youtubers*, a partir de los cuales se puede analizar el lenguaje hablado y la expresión facial.

Tabla 1. *Métodos típicos en análisis de sentimiento.*

Modalidad	Métodos y características típicos
Texto	Diccionarios basados en lexicón, bolsa de palabras (*bag-of-words*), encajes de palabras (*Word embeddings*), en combinación con clasificadores como las Máquinas de Soporte Vectorial (SVM, por su sigla en inglés) o las redes neuronales profundas.
Lenguaje hablado	Atributos paralingüísticos, por ejemplo, tono de voz, en combinación con clasificadores como SVM o las redes neuronales profundas recurrentes.
Visual	Conceptos visuales de nivel medio, correspondientes a parejas adjetivo-pronombre, que presentan fuerte carga de sentimientos, a través de redes neurales convolucionales. También se analizan expresiones faciales, unidades de acción facial o estética visual.
Multimodal	Fusión multimodal de texto, expresión facial y características paralingüísticas.

Adaptada de Soleymani et al., (2017).

En el escenario anterior, la minería de texto orientada al análisis de sentimiento en información textual es un campo que se empezó a desarrollar durante los años 90 del siglo pasado (Soleymani, et al., 2017), cuando surgió el interés por la detección de tópicos en datos textuales, lo que dio lugar a la búsqueda de métodos para la detección de estados subjetivos. De esta forma, a partir de la década del 2000 se empezaron a desarrollar métodos analíticos supervisados y no supervisados. En cuanto a los métodos *supervisados*, están orientados a la "construcción de modelos predictivos de sentimiento a partir de bases de datos anotadas con las cuales se automatiza el aprendizaje" (p.

6), con base en la cuantificación de atributos o palabras del texto, que se entrenan a través de algoritmos como las Máquinas de Soporte Vectorial (SVM, por su sigla en inglés) o los clasificadores Naïve Bayes; los algoritmos de este tipo de métodos generalmente se entrenan en dominios específicos –por ejemplo, Twitter–, por lo cual resulta pertinente investigar la posibilidad de desarrollar modelos que funcionen en distintos contextos. Por otro lado, los métodos *no supervisados* son útiles cuando no existen datos anotados o estos son escasos; la estrategia principal consiste en acudir al conocimiento experto para codificar/calificar 'manualmente' en lexicones frases o palabras con su significado emocional, que después alimentan los recursos computacionales generados para las tareas de minería. Como es natural, estos dos enfoques metodológicos pueden combinarse para producir resultados de mayor precisión. A pesar de estos avances, los autores citados nos recuerdan que

> "El problema del análisis de sentimiento en texto está lejos de estar resuelto, aunque los desarrollos de los últimos años hacen que la aplicación del análisis de sentimientos sea una realidad. No obstante, debemos reconocer que aún no existe una solución universal y que el desempeño de las herramientas varía ampliamente según el contexto, la formalidad y el tipo de texto que se analice. El conocimiento experto y la validación siguen siendo ingredientes necesarios en la aplicación del análisis de sentimiento y una cuidadosa selección de herramientas puede asegurar una aproximación correcta y válida para la cuantificación de sentimientos en textos" (p. 7).

De acuerdo con lo anterior, en este proyecto se buscó establecer la contribución de los *features* fonéticos a la predicción de la valencia emocional, en un corpus de tweets en español e inglés; adicionalmente, se analizó si los fonemas identificados como predictores en esta tarea de regresión lineal múltiple son consistentes con los hallazgos reportados por otros investigadores. En resumen, el problema de investigación está enfocado en poner a prueba una hipótesis acerca de la posible asociación entre aspectos sub léxicos del lenguaje (específicamente, los aspectos fonológicos) y su connotación afectiva, que permitan desarrollar insumos para la generación de recursos que sustenten modelos de *machine learning* para el análisis de sentimientos. Con este fin, se comparó el desempeño de un modelo de regresión lineal múltiple que incluía como predictores los atributos léxicos (*Bag of Words* normalizado mediante TF-IDF) de los datasets de tweets en español e inglés provistos por el concurso SemEval-2018, con el desempeño del mismo modelo de regresión, pero agregando los *features* fonéticos extraídos de estos corpus.

En síntesis, el objetivo general del proyecto fue establecer el grado de contribución de los *features* fonéticos –en un corpus de Tweets en español e inglés– a la predicción de la valencia emocional. Como objetivos específicos se plantearon los siguientes:

- Identificar la capacidad específica de cada fonema en español e inglés, sin combinarlos con otros atributos, para predecir la valencia emocional de los tweets de SemEval-2018.
- Establecer si adicionar los fonemas como *features* adicionales al léxico extraído de los corpus de tweets mejora la capacidad predictiva de un modelo de regresión lineal múltiple.
- Contribuir al campo de conocimientos de la lingüística computacional y del PLN, analizando la interacción entre los aspectos fonológicos y semánticos del lenguaje, específicamente en lo relativo al afecto y las emociones.
- Brindar recursos de información actualizados, que sean útiles en las tareas de minería de texto y emociones en español.

Trabajos relacionados

En principio, resulta relevante ilustrar algunas aplicaciones de las técnicas de análisis de sentimientos en diferentes escenarios. En el ámbito hispanohablante, se puede mencionar el trabajo de Molina, Martínez, & Martín (2015), en el cual se desarrolló un recurso léxico para el análisis de sentimientos llamado CRiSOL a partir del cual se generó un clasificador de polaridad aplicable a comentarios de hoteles. Otro ejemplo de aplicación de las técnicas de análisis de sentimientos lo presentan Purao, Desouza, & Becker (2012), quienes utilizaron técnicas de minería de texto con el objetivo de analizar las posibles fallas en proyectos del sector público, a partir de los reportes de los líderes y las partes interesadas (*stakeholders*); este estudio mostró que este enfoque de análisis representa una forma plausible de obtener indicadores acerca del avance de los proyectos.

Acerca de la interacción entre el lenguaje y las emociones en diferentes niveles de la estructura lingüística, es pertinente citar a Nygaard & Queen (2008) o Roche, Peters, & Dale (2015), quienes han mostrado experimentalmente que ciertos aspectos prosódicos como la entonación y el ritmo pueden incidir sobre el procesamiento de información lingüística de contenido emocional. Igualmente, en términos de la interacción entre emociones y otros procesos psicológicos, Gil, Hattouti, & Laval (2016) encontraron que en niños de 5 a 9 años se evidencia un efecto de modalidad cruzada entre la información facial (de tipo visual) y la información prosódica proveniente del lenguaje (de tipo auditivo), que afecta la comprensión de la comunicación

emocional. Además, mediante la utilización de técnicas de neuro-imagen Zheng, Huang, & Zhang (2013) llegaron a conclusiones análogas. Estos estudios evidencian que las personas empleamos información proveniente de distintas fuentes (visuales y auditivas) para decodificar la información emocional a nivel psicológico y cerebral.

Desde la perspectiva de la psicología experimental, en el trabajo de Rummer, Schweppe, Schlegelmilch, & Grice (2014) mediante la presentación de material audiovisual se indujo un estado de ánimo positivo o negativo en una muestra de participantes, para evaluar la producción posterior de ciertos fonemas en la escritura de palabras inventadas. Los resultados indicaron que cuando las personas se encontraban en un estado de ánimo negativo escribieron con mayor frecuencia palabras que contenían el fonema /o:/, mientras que los participantes que se encontraban en un estado de ánimo positivo utilizaron más veces el fonema /i:/, demostrando que el estado de ánimo influye sobre la elección de ciertas vocales. Estos hallazgos se explican desde una perspectiva psicofisiológica, argumentando que los músculos faciales involucrados en la expresión de determinadas emociones son los mismos que se emplean al articular las vocales objeto de estudio (hipótesis articulatoria).

En todo caso, la idea de contar palabras como un indicador de emocionalidad ha sido cuestionada, pues el supuesto de que el uso de determinadas palabras refleja ciertos estados emocionales es debatible. Por ejemplo, Kross, et al. (2019) reportan evidencia obtenida con una muestra de 185 estudiantes universitarios, mediante técnicas de simulación (*bootstrapping*) y análisis de correlación de conteos de palabras en redes sociales con medidas de afecto autorreportado (antes o después del uso de redes sociales), encontrando resultados poco alentadores que controvierten la validez de las técnicas de minería de datos aplicadas en el análisis de sentimiento. Los investigadores concluyeron que contar el uso de palabras emocionales en Facebook no es un índice preciso de cómo se sienten los usuarios. Además de los problemas en el manejo del ruido (por ejemplo, ambigüedades, ironías o negaciones) o del análisis contextual de la información (una misma palabra puede tener sentidos opuestos si están en distintas frases) (Soleymani, et al., 2017), estos indicios señalan la importancia de identificar otros aspectos del lenguaje escrito que puedan brindar información precisa y válida sobre la experiencia emocional de las personas, por lo que el análisis de los atributos fonológicos resulta una veta prometedora de exploración sustentada en el concepto de iconicidad fonológica que se discute más adelante.

De forma específica, en relación con el objetivo del presente proyecto resulta relevante exponer resultados obtenidos por otros investigadores al abordar la misma tarea (regresión de la valencia emocional en tweets en el concurso SemEval-2018). De acuerdo con los promotores del concurso (Mohammad, Bravo-Márquez, Salameh, & Kiritchenko, 2018), esta tarea de predicción fue abordada principalmente a través de modelos de regresión lineal basados en una variedad de *features* que incluyeron *embeddings* de palabras y de oraciones, lexicones, o n-gramas de caracteres y palabras, entre otros, logrando correlaciones máximas entre las valencias predichas y el *gold standard* del concurso de ,873 en inglés y ,795 en español. Por ejemplo, Duppada, Jain, & Hiray (2018) lograron el mejor desempeño frente a la tarea con los datasets en inglés, para lo cual utilizaron *features* de emojis y contenido semántico (*word embeddings* y lexicones), a partir de los cuales entrenaron modelos de regresión con métodos de *Random Forest* y ensambles. Con una aproximación similar (combinación de *features* y ensambles de modelos), pero a partir de otros datasets de Twitter y Facebook en inglés, Akhtar, Ghosal, Ekbal, Bhattacharyya, & Kurohashi (2018) alcanzaron correlaciones de hasta ,635 y ,727 para la variable de valencia. Por otra parte, al abordar la tarea en español, Kuijper, van Lenthe, & van Noord (2018) utilizaron lexicones y *word embeddings* para entrenar modelos basados en ensambles de redes neuronales y máquinas de soporte vectorial, alcanzando desempeños máximos de ,766 en la correlación de Pearson con los *gold standard* del concurso.

Uno de los referentes más destacados en relación con este estudio es el trabajo de Adelman, Estes, & Cossu (2018), quienes partiendo del concepto de simbolismo de los sonidos (*sound symbolism*) demostraron que en un modelo de regresión con un diccionario de casi 37.000 unidades en diferentes idiomas (inglés, español, holandés, alemán y polaco) el primer fonema de una palabra era el mejor predictor de su valencia emocional; igualmente, bajo un paradigma experimental, encontraron que, a diferencia de los fonemas positivos, los negativos eran emitidos con mayor velocidad. Sus hallazgos son analizados teóricamente asumiendo que el simbolismo de los sonidos es un fenómeno desarrollado a lo largo del proceso de evolución de la especie, por lo que estas asociaciones podrían tener un carácter adaptativo por asociarse a señales de alerta y peligro. En todo caso, como se comentará en la sección de discusión de este documento, estos investigadores se basaron en las palabras como unidades de análisis (mientras que nosotros lo hicimos con tweets)

De otra parte, aunque el presente proyecto se basa en el procesamiento de información textual, resulta pertinente presentar los avances que se han dado en el análisis de sentimientos a partir del

lenguaje hablado, puesto que la hipótesis que se pondrá a prueba tiene relación con los aspectos fonológicos del lenguaje, que obviamente se relacionan con sus atributos sónicos. Entre las características del habla que han sido estudiadas se encuentran los parámetros físicos como el tono (*pitch*) o la velocidad y, más específicamente, las variaciones intra individuales en dichos parámetros, teniendo en cuenta que dado el carácter idiosincrásico de la voz humana los cambios de entonación y resultan ser indicadores importantes de emocionalidad, lo que se evidenció en el estudio de Tyagi & Chandra (2015). En el mismo sentido, Mairesse, Polifroni, & Di Fabbrizio (2012) utilizando revisiones habladas de 84 hablantes, encontraron que la información prosódica mejora el desempeño de modelos predictivos basados únicamente en información textual.

Finalmente, se presentan algunas aplicaciones de las técnicas de análisis de sentimiento en el contexto colombiano. En el ámbito de la política, Alvarado, Caicedo, Carrillo, Forero, & Urueña (2016) realizaron un trabajo para explorar los sentimientos frente a las elecciones de la alcaldía de Bogotá, cuyos resultados fueron comparados con algunos sondeos de intención de voto, logrando un nivel medio de precisión a través de algoritmos eficientes. De forma más genérica, Henríquez, Pla, Hurtado, & Guzmán (2017) exponen un sistema de análisis basado en la aplicación de Máquinas de Soporte Vectorial (SVM) que en pruebas experimentales arrojaron muy buenos resultados.

Marco conceptual

> "La estructura y el uso del lenguaje natural se basan en el supuesto de que los participantes de la conversación comparten experiencias y conocimientos similares, así como una forma de razonar, sentir y actuar. El gran desafío del problema del procesamiento automático de textos es usar un lenguaje natural sin restricciones para intercambiar información con una criatura de una naturaleza totalmente distinta: el computador" (Bolshakov & Gelbukh, 2004, Computational Lingüistics: Models, Resources, Applications, página 15).

Para contextualizar el proyecto, a continuación, se despliegan algunos elementos teórico conceptuales de referencia. Para esto, inicialmente se expondrán algunos elementos fundamentales de la lingüística, señalando su propósito como disciplina científica y definiendo algunos conceptos fundamentales; posteriormente, se delimitará el campo de la psicolingüística, enfatizando en el fenómeno de la modalidad cruzada (*cross modality*), específicamente en lo relativo a la interacción entre los procesos del lenguaje y la emoción; y, para terminar, se revisará el ámbito del

Procesamiento del Lenguaje Natural (PNL) y de la lingüística computacional, como marco metodológico que sustenta el ejercicio investigativo que se realizará.

Lingüística.

Se puede definir la lingüística como la ciencia del lenguaje natural, esto es, del lenguaje que se da en la interacción de los hablantes 'reales', y constituye, así mismo, un marco general que integra y articula distintos campos de conocimiento, como se expresa en la Figura 1 (Bolshakov & Gelbukh, 2004). Adicionalmente, de acuerdo con Gil (1999), el objeto de estudio de la lingüística está delimitado por los fenómenos relativos al sistema de la lengua, como fenómeno **social**, en tanto remite a la interacción entre los individuos, y como fenómeno **individual**, en tanto se ubica en la "mente" de las personas. En este sentido, la lingüística es una rama de la semiología, ocupada del estudio de los signos en general, que a su vez es una rama de estudio de la psicología social (y, por lo tanto, de la psicología en general).

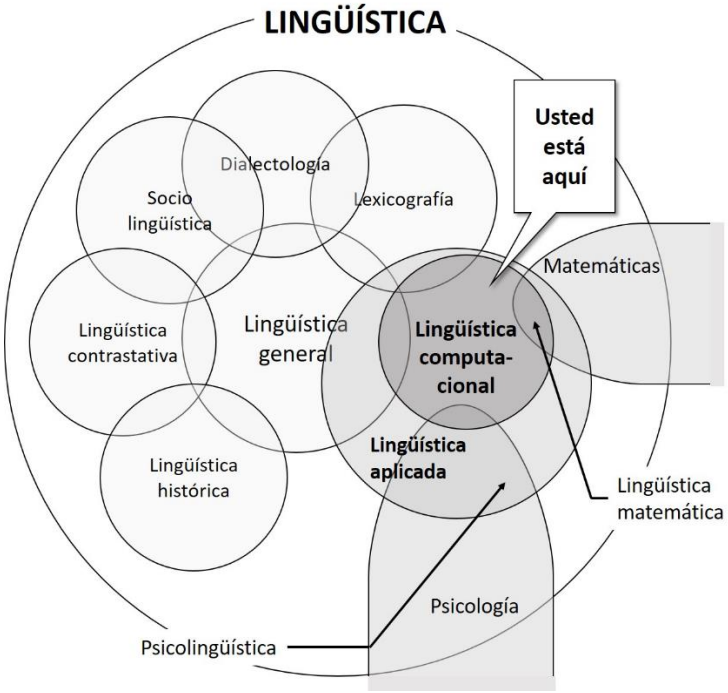


Figura 1. *Estructura de las ciencias lingüísticas (traducida de Bolshakov & Gelbukh, 2004).*

Según Bolshakov & Gelbukh (2004), en este marco la lingüística general es el núcleo alrededor del cual se agrupan las demás disciplinas y cuyo propósito es descubrir leyes universales, en

relación con aspectos fonológicos, morfológicos, semánticos y pragmáticos del lenguaje hablado y escrito. De las ciencias complementarias en el diagrama, vale la pena destacar en este proyecto la *lexicografía*, ocupada de estudiar el lexicón (conjunto de palabras) y desarrollar métodos para la compilación de diccionarios, y la *psicolingüística*, interesada en el estudio del comportamiento lingüístico desde una perspectiva psicológica, incluyendo la interacción de diferentes procesos (como la emoción) con el lenguaje natural. También es importante destacar la *lingüística matemática* (entendida como teoría de las gramáticas formales o como la intersección entre lingüística y matemáticas), una de cuyas ramas es la lingüística cuantitativa o estadística, así como la *lingüística aplicada*, orientada al uso de las ideas de la lingüística general en la práctica humana.

Ya específicamente, la *lingüística computacional* se define como el procesamiento automatizado del lenguaje natural con el objetivo de desarrollar programas computarizados que procesen palabras y textos del lenguaje natural. La definición anterior se puede complementar con la propuesta por la ACL (Asociación para la Lingüística Computacional, por su sigla en inglés) (2019), que la define como el "estudio científico del lenguaje desde una perspectiva computacional… [para] proveer modelos computacionales de varios tipos de fenómenos lingüísticos. Estos modelos pueden estar 'basados en el conocimiento' ('hechos a mano') o 'basados en los datos' ('estadísticos' o 'empíricos')". La aplicación concreta de la lingüística computacional se denomina Procesamiento del Lenguaje Natural (PLN), campo que será abordado más adelante en el presente documento.

El estructuralismo y la arbitrariedad del signo según Saussure.

Una de las teorías más lingüísticas más destacadas del siglo XX se enmarca dentro de las llamadas corrientes estructuralistas, cuyo principal exponente es Ferdinand de Saussure, quien define la lengua como "un sistema de signos que expresan ideas" (1916/2005). Para este importante teórico del lenguaje, la unidad básica de análisis es el signo lingüístico, una combinación mental de un concepto (significado o representación mental de un significado) y una imagen acústica (significante o representación mental de un sonido que expresa un significado). De acuerdo con lo anterior, las palabras, o sus raíces etimológicas constituyen signos lingüísticos; así, por ejemplo, al concepto 'felino doméstico' le corresponde la imagen acústica 'gato'.

Para mayor precisión, Majid (2012) se refiere a los fonemas como las vocales y consonantes de un idioma, que remiten a un sonido específico que las distingue como letras (por ejemplo, el sonido

de la ‘C’ es diferente al sonido de la ‘K’). Dada la variedad lingüística del mundo, el número de fonemas varía según el idioma; por ejemplo, el español cuenta con cinco sonidos vocálicos y entre 17 y 19 sonidos consonánticos (Franch y Blecua, 1980, citados por González-Díaz y cols, 2014), pero claro, esto puede variar según la región o la cultura de los hablantes.

Ahora bien, para de Saussure (1916/2005), uno de los principios de rigen el funcionamiento de la lengua es la arbitrariedad del signo lingüístico, según el cual el vínculo entre significado y significante no es motivado, es decir, no existe ninguna relación *esencial* o ‘lazo natural’ entre significado y significante. Un argumento a favor de esta postura es que en diferentes lenguas el mismo concepto puede ser (y de hecho es) expresado con diferentes significantes (*gato* en castellano, *cat* en inglés, *chat* en francés o *katze* en alemán). De aquí se deriva un segundo principio, el de la inmutabilidad del signo, según el cual, dado que la lengua es heredada de generación en generación y teniendo en cuenta que hay cambios histórico-sociales, no tiene sentido que el signo cambie, por lo cual no hay razón para que el signo no sea arbitrario (“decimos *hombre* y *perro* porque antes de nosotros se ha dicho *hombre* y *perro* [de Saussure, 1916])”.

Estas ideas marcaron de manera definitiva la teorización lingüística a lo largo del siglo XX, implicando una visión jerárquica del lenguaje, como sistema simbólico organizado en función de una serie de componentes discretos que se articulan en diferentes niveles de análisis (oraciones, palabras, fonemas). En este escenario, es de especial interés el campo específico de la fonología, que se interesa por la manera como se articulan los sonidos para dar significación a los enunciados y que se diferencia de la fonética, ocupada del estudio de los mecanismos de producción y percepción de los sonidos del lenguaje (González Díaz, y otros, 2014). En conclusión, es bastante notable el hecho de que el nivel de representación fonológica de una lengua particular esté formado por un conjunto finito (y relativamente pequeño) de unidades mínimas sin significado propio (fonemas), las cuales se articulan en unidades significativas (monemas o morfemas).

Lingüística computacional y Procesamiento del Lenguaje Natural (PLN).

La lingüística computacional tiene como objetivo “desarrollar sistemas computacionales que simulen, total o parcialmente, las destrezas y habilidades de un hablante real (v.gr.: decodificar un texto escrito, identificar los sonidos del habla, dotar de sentido a un texto…)” (Tordera Yllescas, 2011, pág. 341). Entre tanto, de acuerdo con Fisher, Garnsey, & Hughes (2016), el Procesamiento del Lenguaje Natural –PNL– es una parte del dominio de la inteligencia artificial que se constituye

como un “campo de estudio inherentemente interdisciplinario basado en dominios tan diversos como la psicología, la lingüística y la neurociencia… [que] abarca un rango de técnicas computacionales para analizar y representar textos en uno o más niveles de análisis lingüístico con el fin de permitir el procesamiento del lenguaje humano para un rango de tareas o aplicaciones particulares” (p. 157).

La relevancia actual de este campo de estudio tiene que ver con el hecho de que desde mediados del siglo pasado ha habido un interés creciente en el procesamiento automatizado del lenguaje natural. Dos razones, por lo menos, justifican este interés: la estandarización de procesos de producción textual puede asociarse a la automatización de procesos productivos en los cuales es necesaria la transmisión de información a través del lenguaje escrito (es más barato ‘entrenar’ una máquina que un humano) y la toma de decisiones informadas requiere del análisis de grandes cantidades de información que sería humanamente imposible de procesos (Bolshakov & Gelbukh, 2004). Desde esta perspectiva, mientras que la psicolingüística pretende describir y explicar las habilidades de un hablante-oyente real, la lingüística computacional se interesa en formalizar este conocimiento para un “hablante-oyente digital” (Tordera Yllescas, 2011).

De lo anterior se deriva que el objetivo de esta disciplina es desarrollar productos informáticos, con base en elementos teóricos de la lingüística y otras ciencias que han aportado a la comprensión del fenómeno lingüístico. En otras palabras, el desarrollo de estas tecnologías informáticas se basa en el conocimiento existente sobre lexicología, morfología y sintaxis. Por citar algunos ejemplos, el PLN ha permitido desarrollar sistemas de reconocimiento de voz, de lectura de texto o de traducción (Tordera Yllescas, 2011). Halvorsen (1988, citado por Tordera Yllescas, 2011) identifica tres campos de aplicación de la lingüística computacional teórica: 1) el tratamiento del habla, 2) el análisis, la generación e interpretación del lenguaje natural, y 3) la traducción automática. El presente proyecto se inscribe dentro de la segunda línea.

Vale la pena mencionar que las relaciones interdisciplinares de la lingüística computacional se inscriben en la metáfora computacional formulada desde los vínculos entre la psicología cognitiva y la inteligencia artificial, que busca trasladar el conocimiento sobre el funcionamiento de la mente a los computadores. Sin embargo, se debe tener en cuenta que esta relación entre el lenguaje natural y el lenguaje digital es solo de semejanza y no de analogía completa, pues las inteligencias artificiales carecen de la intencionalidad y subjetividad (los aspectos idiosincrásicos) del lenguaje humano (Tordera Yllescas, 2011). Justamente, uno de los principales desafíos relacionados con la

producción de lenguaje artificial tiene que ver con la simulación de los aspectos prosódicos, que pueden llegar a determinar la inteligibilidad de los mensajes que intercambian entre sí los hablantes-oyentes.

En relación con este último punto, un aspecto que se afecta directamente al presente proyecto por constituir un desafío a resolver tiene que ver con el hecho de que las unidades acústicas del lenguaje (fonemas o sílabas, por ejemplo), no presentan patrones estándar, es decir, el habla y sus características acústicas resultan sumamente complejas que pueden variar entre hablantes o incluso para un mismo individuo. Puede ocurrir, por ejemplo, que la acústica de determinados segmentes fonéticos puede cambiar según el contexto del contenido comunicativo.

Por otro lado, Henríquez, Pla, Hurtado, & Guzmán (2017) señalan que el análisis de sentimientos tiene como objetivo identificar las propiedades o características de una entidad y determinar la polaridad afectiva expresada de cada aspecto de esa entidad. El análisis se puede realizar a partir de aspectos específicos del lenguaje como su sintaxis (*aspect-based sentiment analysis*) o con base en características o variables de la información (*feature-based sentiment analysis*).

Con respecto al análisis de sentimientos, para Yadollahi, Shahraki, & Zaiane (2017), es un campo de la "computación afectiva" que tiene como fin "detectar, analizar y evaluar los estados mentales de los humanos hacia diferentes eventos, temas, servicios o cualquier otro interés,… se busca hacer minería de opiniones, sentimientos y emociones con base en observaciones de las acciones de las personas que pueden ser capturadas usando su escritura, expresión facial, lenguaje oral, música, movimientos, etc." (p. 25). Los mismos autores diferencian entre la minería de opiniones y la minería de emociones, como tareas específicas del análisis de sentimientos, pues estos conceptos tienen una connotación diferente.

La Figura 2 esquematiza las tareas de análisis de sentimientos, de acuerdo con los autores arriba citados. Según esta clasificación de tareas, la minería de opiniones se enfoca en **detectar subjetividad** (identificar si un texto es objetivo –basado en hechos–, o subjetivo –basado en opiniones), **clasificar la polaridad de opiniones** (determinar si un texto expresa opiniones positivas, negativas o neutrales), **detectar opiniones *spam*** (reconocer opiniones falsas con intenciones maliciosas), **resumir opiniones** (sintetizar un grupo amplio de opiniones desde diferentes perspectivas) y **detectar expresión de argumentos** (identificar estructuras argumentativas y relaciones entre ellas). Por otro lado, la minería de emociones busca **detectar**

emociones (identificar si un texto transmite determinadas emociones), **clasificar polaridad emocional** (determinar la valencia de las emociones expresadas en un texto), **clasificar emociones** (clasificación más fina de las emociones contenidas en un texto) y **detectar causas de las emociones** (hacer minería de factores que puedan provocar determinadas emociones).

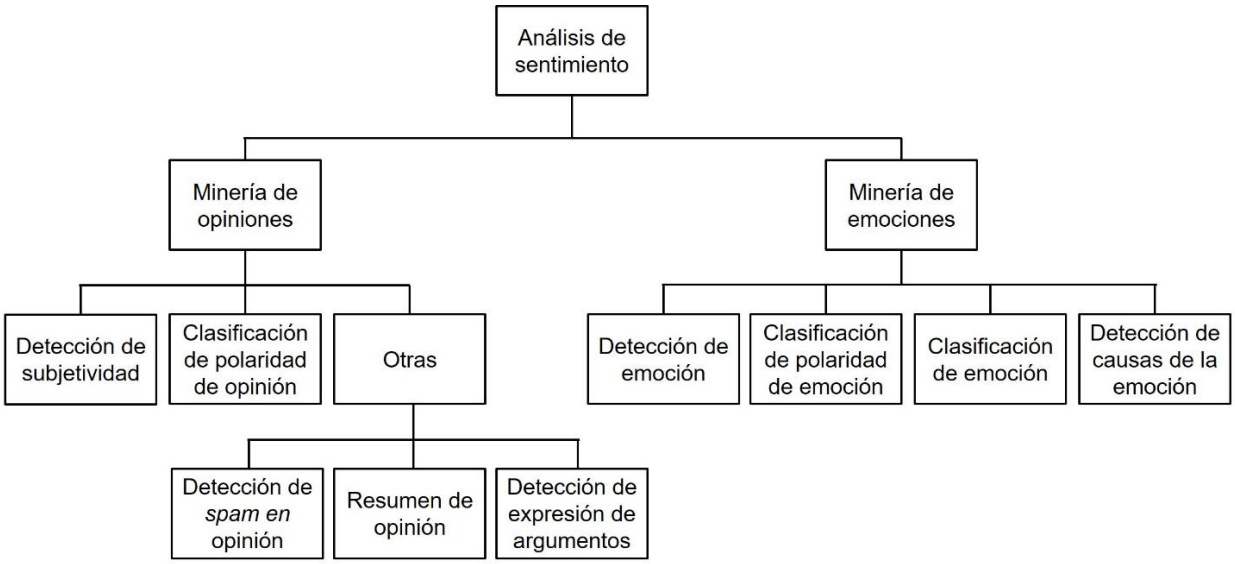


Figura 2. Clasificación de tareas de análisis de sentimiento (Traducida de Yadollahi, Sharahki & Zaiane, 2017).

Soleymani et al. (2017) aportan a la delimitación de este campo con las siguientes precisiones. La minería de opiniones consistiría en la extracción automática de juicios subjetivos con respecto a cualquier aspecto de un objeto particular (incluyendo la polaridad del sentimiento implicado), mientras que el análisis de sentimientos busca identificar el sentimiento asociado a la opinión. En tanto, las tareas de reconocimiento de emociones pretenden identificar de forma automatizada reacciones emocionales episódicas, normalmente en una persona específica, y pueden apoyar la identificación de la polaridad de un sentimiento frente a determinado atributo de una entidad de interés. Estos autores también indican que una de las tareas que se realizan en el análisis de sentimiento es la representación de la intensidad del sentimiento, aunque discuten si los sentimientos como disposiciones afectivas pueden ser representados como emociones discretas.

Liu y Zhang (2012, citados por Soleymani et al., 2017) indican que el análisis de sentimientos implica la identificación de cuatro componentes de un sentimiento: entidad (objeto, situación o individuo hacia el cual se dirige el sentimiento), aspecto (atributo frente al cual se presenta el

sentimiento), opinante (individuo que expresa el sentimiento) y sentimiento (experiencia frente al atributo/aspecto con determinada valencia afectiva). Un sistema de análisis de sentimiento debe permitir extraer dichos componentes de forma correcta.

Emoción, cognición y lenguaje.

En la literatura científica no existe un consenso acerca de qué son las emociones. En todo caso, se pueden agrupar las teorías de la emoción como cognitivas o somáticas. Las primeras consideran la cognición como parte esencial de las emociones y específicamente de la experiencia emocional subjetiva; así, por ejemplo, la emoción puede definirse como una reacción afectivamente importante, que involucra afecto, conciencia y valoración de un objeto emocional, preparación para la acción y activación automática (Frijda, 1994, citado por Lopatovska y Arapakis, 2011), o bien como una sincronización de diferentes procesos corporales, perceptuales y cognitivos (Scherer, 2005, citado por Lopatovska y Arapakis, 2011). Por otro lado, las teorías somáticas afirman que las emociones son producidas por procesos corporales, más que por juicios cognitivos. Estas perspectivas consideran las emociones como sistemas motivacionales primarios y asumen una postura evolucionaria, en la que se destaca el carácter adaptativo e instintivo de las emociones (por ejemplo, las posturas de Tomkins, 1984, Plutchik, 1980 o Ekman, 1984, citados por Lopatovska y Arapakis, 2011).

Otra forma de clasificar las teorías sobre las emociones (relacionadas con la anterior agrupación) es identificar enfoques de origen evolucionario y enfoques que conciben las emociones como fenómenos totalmente determinados por la sociedad y la cultura. En todo caso, la tendencia científica se orienta hacia una visión ecléctica e interaccionista que comprende condiciones ambientales, reacciones fisiológicas y experiencias subjetivas, enfatizando la función adaptativa de las emociones tanto para la supervivencia de la especie en el largo plazo, como para las relaciones sociales inmediatas (Stryker, 2004). Una consecuencia de los debates arriba mencionados es la conocida controversia acerca del carácter universal o culturalmente específico de las emociones. Esta tensión teórica, a su vez, implica la pregunta acerca de si existe un grupo de emociones que pueda identificarse independientemente del contexto sociológico y antropológico (Thamm, 2004).

Adicionalmente, Soleymani et al. (2017) indican que los términos afecto, sentimiento (en inglés *feeling* o *sentiment*), emoción y opinión han sido usados de forma intercambiable, y que existen

muchas acepciones de emoción y sus conceptos relacionados, dependiendo de la teoría específica desde la cual se aborden. A continuación, se presentan algunas definiciones, con base en la revisión de Yadollahi, Shahraki, & Zaiane (2017) y Soleymani et al. (2017).

- **Afecto:** Término abarcador para referirse de forma general a los sentimientos, emociones o estados de ánimo.
- **Sentimiento:** Actitud, pensamiento o juicio impulsado por una emoción; representación o experiencia subjetiva de una emoción. Los sentimientos implican a una persona que los experimenta, una polaridad (positiva o negativa) y un objeto.
- **Emoción:** Respuesta discreta a un evento particular interno o externo, significativo para el organismo. Se considera un evento de corta duración y alta intensidad. Como fenómenos de corto plazo, las emociones requieren detonantes, involucran valoraciones cognitivas, reacciones corporales, tendencias de acción, expresiones y sentimientos subjetivos. Para efectos prácticos en tareas de análisis de sentimiento, Munezero et al. (2014, citados por Soleymani et al., 2017) diferencian emociones y sentimientos a partir de su duración, siendo las primeras de corto plazo y los segundos de largo plazo.
- **Actitud:** Se refiere a un juicio o pensamiento valorativo con respecto a algo (por ejemplo, un individuo o un producto), que tiene una carga afectiva negativa, positiva o neutral.
- **Estado de ánimo:** Estado afectivo difuso menos intenso que una emoción, pero más duradero y prolongado en el tiempo (pueden durar horas o días). Para los estados de ánimo no existen detonantes aparentes.
- **Opinión:** Juicio incierto y abierto al debate que no necesariamente tiene una carga emocional. Refiere a un tópico, un portador, un argumento (*claim*) y un sentimiento; implica una interpretación personal de la información y, a diferencia de las emociones, no necesariamente se manifiesta en el comportamiento o la expresión.

Otra discusión relevante acerca de la naturaleza de las emociones tiene que ver con su naturaleza discreta o continua (Lopatovska & Arapakis, 2011; Yadollahi, Shahraki, & Zaiane, 2017). Los teóricos discretos conciben la existencia de un conjunto de emociones básicas (seis o más) universalmente expresadas y reconocidas, que pueden ser identificadas incluso en otros primates y que dependen de sistemas neuronales diferenciados. Entre tanto, los teóricos continuos

plantean que las emociones básicas se definen a partir de la existencia de dos o más dimensiones (por ejemplo, valencia positiva-negativa, placer-displacer, nivel de *arousal*, dominancia-sumisión, etc.), que dependen de un sistema neurofisiológico común e interconectado. En términos de las implicaciones para el presente proyecto, se puede anticipar que las tareas de minería de datos podrían orientarse a la predicción (tratando las dimensiones emocionales como variables escalares) o a la clasificación (concibiendo las emociones como variables categóricas). Una precisión conceptual importante para este estudio tiene que ver con la diferencia entre los atributos de *arousal* y valencia (en el cual se basa nuestro análisis); el primer concepto remite a una reacción emocional básica, esto es, al nivel de activación fisiológica que un estímulo elicita, mientras que el segundo concepto se refiere a una valoración cognitiva subjetiva de qué tan placentero o displacentero resulta un estímulo (Stadthagen-Gonzalez, Imbault, & Pérez Sánchez, 2017)

En todo caso, existe cierto acuerdo en torno a la idea de un grupo finito de emociones básicas, que no están compuestas por otras emociones, aunque Lopatovska & Arapakis (2011) señalan que en la ciencia computacional el modelo de Ekman de seis emociones es el más comúnmente empleado. La Tabla 2 presenta algunas de las clasificaciones más conocidas, desde las dos perspectivas presentadas.

Tabla 2. *Diferentes modelos de las emociones básicas propuestos por los teóricos.*

Teórico	Año	Emociones básicas	Tipo
Ekman	1972	Ira, disgusto (asco), miedo, alegría, tristeza, sorpresa	Discreto
Plutchik	1986	Ira, anticipación, disgusto, miedo, alegría, tristeza, sorpresa, confianza	Dimensional
Shaver	1987	Ira, miedo, alegría, amor, tristeza, sorpresa	Discreto
Lovheim	2011	Ira, disgusto, angustia (*distress*), miedo, alegría, interés, vergüenza, sorpresa	Dimensional

Traducida de Yadollahi, Shahraki & Zaiane (2017).

Con relación a la interface entre lenguaje y emoción, resulta pertinente remitirse al concepto de iconicidad fonológica (o simbolismo de los sonidos). Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan (2015) indican que, en contraste con la postura tradicional sobre la arbitrariedad del signo lingüístico, en el lenguaje juegan un rol relevante tanto la arbitrariedad como la no arbitrariedad. Específicamente, la no arbitrariedad del lenguaje puede darse en diferentes aspectos de la forma lingüística, como la estructura del discurso, la sintaxis y la morfología. Los autores

argumentan que, dada la variedad lingüística en los distintos idiomas humanos, la evidencia de apoyo para esta característica debe asumir una perspectiva transversal a través de distintas lenguas.

Igualmente, Dingemanse et al. (2015) diferencian dos fenómenos específicos de la no arbitrariedad y la arbitrariedad. En primer término, está la iconicidad (semejanza entre aspectos de forma y significado), la cual implica la existencia de analogías de los procesos perceptuomotores con el aprendizaje y la comprensión del lenguaje, que podrían tener un carácter universal y por lo tanto no arbitrario. Ejemplos clásicos de la iconicidad son las llamadas onomatopeyas –palabras que imitan sonidos–, las cuales pueden presentar similitudes en distintos idiomas (la onomatopeya para el ladrido de un perro en español es *guau*, en japonés *wan wan* y en inglés es *woof* o *bow wow*), y los ideófonos –palabras que permiten transmitir ideas o sensaciones (como los sonidos empleados para referirse a los aumentativos o los diminutivos). Por otra parte, se encuentra la sistematicidad (relación estadística entre los patrones de sonido de un grupo de palabras y el uso de las mismas), que puede ser arbitraria, en la medida en que determinadas regularidades fonológicas pueden ser específicas de determinados idiomas. El análisis de corpus o diccionarios ha revelado, por ejemplo, la consistencia de ciertas pistas fonológicas y prosódicas que permiten distinguir sustantivos de verbos o indicar aspectos semánticos como la concreción.

De forma más puntual, y con base en el concepto de iconicidad, se ha planteado que existe un código de frecuencia asociado a ciertas letras: la /i:/ es comúnmente empleada para expresar el concepto de pequeñez o una actitud de sumisión y produce un espectro sónico distinto al de la /o:/, que usualmente se emplea para expresar agresividad o una actitud amenazante, incluso en primates diferentes al *homo sapiens sapiens*, lo cual se sustenta en la hipótesis del *feedback* articulatorio, según el cual existe una relación bidireccional entre las emociones y la actividad de los músculos faciales que producen el lenguaje hablado, pues la activación de estos músculos, en particular el zigomático mayor y el orbicular labial (asociados a la producción del sonido de la /i:/ y de la /o:/, respectivamente), influye sobre el flujo sanguíneo afectando la temperatura cerebral y, por lo tanto, la liberación de neurotransmisores asociados a las emociones (Rummer, Schweppe, Schlegelmilch, & Grice, 2014).

En todo caso, es importante anotar que la hipótesis de la iconicidad fónico emocional también ha sido puesta a prueba con resultados negativos. Por ejemplo, en el campo de la literatura, en el trabajo de Kraxenberger & Menninghaus (2016) no se pudo confirmar la conexión entre la

frecuencia u ocurrencia de ciertas clases de vocales o consonantes y la percepción del tono afectivo de poemas escritos en alemán.

Esta discusión también se puede enmarcar en el concepto de modalidad cruzada (*cross modality*), que se refiere a los efectos del procesamiento psicológico de información con diferentes códigos. Es un fenómeno que se ha estudiado comúnmente en los procesos senso-perceptuales (por ejemplo, en estudios de modalidad cruzada se puede determinar si ciertas imágenes evocan sensaciones olfativas). En el contexto del presente proyecto, se refiere a posibles interacciones entre códigos lingüísticos y códigos emocionales o afectivos. Un concepto útil para abordar este proyecto es el que proponen Aryani, Jacobs y Conrad (2013), quienes se refieren a esta interacción fonológico-semántica como iconicidad fonológica.

Método

Datos

Los datos empleados en este estudio fueron tomados de la tarea 1 del concurso SemEval-2018: *Affect in Tweets* (Mohammad, Bravo-Márquez, Salameh, & Kiritchenko, 2018), orientada a la identificación del contenido emocional de un conjunto de tweets. Específicamente, se tomó la subtarea 3 (*V-reg*), la cual planteaba el problema de predecir la intensidad de la valencia que mejor representara el estado emocional del autor de un tweet. Para esta tarea de regresión, la variable objetivo fue una puntuación real continua, con valores comprendidos entre 0 (más negativo) y 1 (más positivo). Si bien los datasets proporcionados para las diferentes subtareas se encontraban en inglés, español y árabe, para este proyecto solamente fueron analizados los tweets en los primeros dos idiomas. Los conjuntos de entrenamiento y desarrollo (prueba), anotados con la valencia emocional, se distribuían como se presenta en la Tabla 3.

Tabla 3. *Distribución en entrenamiento y prueba (development).*

Dataset	Español	Inglés
Entrenamiento	1.566 (87%)	1.181 (72%)
Development	229 (13%)	449 (28%)

Elaboración propia.

Para la anotación de los datasets, los compiladores de SemEval emplearon el método de *Escalamiento Mejor menos Peor* (B-WS, por su sigla en inglés), un método comparativo que mostró confiabilidad y capacidad discriminativa. En este método, un grupo de anotadores convocados y seleccionados en una plataforma de *crowdsourcing*, evaluaron tuplas de 4 tweets, los cuales tenían que ordenar según la propiedad de interés (intensidad de la valencia emocional); el ranking obtenido para cada tweet fue transformado a un valor real, calculado como la proporción de veces que fue calificado con la mayor intensidad menos la proporción de veces que fue calificado con la menor intensidad. Finalmente, las puntuaciones resultantes se escalaron en un rango de 0 a 1. Específicamente, la confiabilidad de estas puntuaciones fue verificada mediante el método de división por mitades (Kuder y Richardson, 1937, citados por Mohammad, Bravo-Márquez, Salameh, & Kiritchenko, 2018), consistente en dividir aleatoriamente las puntuaciones de las tuplas en dos grupos y luego obtener la correlación media entre las dos series, con resultados

entre 0,82 y 0,92, valores cercanos a 1 que se pueden interpretar como un alto grado de consistencia o reproducibilidad de las puntuaciones de valencia obtenidas para los tweets.

Con respecto al tratamiento de los datos, se emplearon los lenguajes de programación Python y R, siguiendo el esquema general en un proceso de *machine learning*: pre procesamiento, entrenamiento y evaluación. A continuación, se describe cada una de estas etapas.

Pre procesamiento de los tweets

En términos de la primera fase, las tareas realizadas fueron 1) limpieza y tokenización por palabras de los Tweets, 2) transcripción fonética de los tokens obtenidos, y 3) vectorización TF-IDF de los tokens de palabras y los tokens de fonemas extraídos en los pasos anteriores. Para la limpieza y tokenización se empleó la biblioteca NLTK -Natural Language Toolkit (Bird, Loper, & Klein, 2009) de Python, con la cual se creó un algoritmo para eliminar de todos los tweets los siguientes elementos: menciones a url's y usuarios (@), números (en dígito y en letra), caracteres especiales, puntuaciones y paréntesis; adicionalmente, el contenido fue normalizado, dejando todo en minúsculas, eliminando la acentuación (tíldes y diéresis) y removiendo las *stop words* respectivas de cada idioma. Hay que aclarar que en la tokenización no se aplicaron técnicas de estematización o lematización, pues dado que el objetivo fue valorar la contribución de los fonemas en la predicción de la valencia emocional, se consideró relevante conservar posibles inflexiones en las palabras (por ejemplo, diminutivos o aumentativos) que potencialmente aportarían a la connotación afectiva de los tweets.

Para la fonetización de los tokens obtenidos en el paso anterior se empleó EPITRAN, un sistema multilingüe G2P (grapheme to phomene), open source y con una licencia MIT (Mortensen, Dalmia, & Littell, 2018) para Python. Entre las ventajas de este sistema se encuentran su amplía cobertura (61 idiomas), así como la posibilidad de transcripción a través de diferentes estándares fonéticos (IPA, X-SAMPA); para este proyecto se empleó el alfabeto fonético internacional –IPA.

Finalmente, en la primera etapa se realizó un proceso de vectorización de las palabras y los fonemas. Este proceso de *feature extraction* del contenido léxico del corpus de entreamiento y prueba se realizó con el algoritmo TF–IDF (*Term Frequency – Inverse Document Frequency*), disponible en la librería de quanteda (Benoit, et al., 2018) para R. Este índice se aplicó para establecer la relación entre cada token (palabra o fonema) y el conjunto completo de tweets empleados de entrenamiento y prueba, para establecer la especificidad o comunalidad de cada

token, lo que implica que los términos con valores altos de TF–IDF son *distintivamente* frecuentes en un tweet, en comparación con el resto de tweets de la colección completa (Lavin, 2019). En este algoritmo, la frecuencia de un término (TF) corresponde a la ocurrencia específica de un token dentro de un tweet, mientras que la frecuencia del documento (DF) implica la cantidad de ocurrencias de un token específico en la colección total de documentos, por lo que la frecuencia inversa de documento (IDF) es una medida de qué tan común es un token en el total de tweets. Es importante aclarar que el cálculo de la ponderación TF-IDF se realizó por separado para cada dataset (entrenamiento y prueba), con el fin de minimizar los riesgos de sobreajuste de los modelos entrenados posteriormente. Específicamente, la frecuencia de un término (token de palabras o fonemas) por documento (tweet), fue calculada como

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k n_{t,d}} \tag{1}$$

siendo t un token específico y d un tweet del dataset; por lo tanto, la frecuencia del término (TF) corresponde a la cantidad de veces que aparece el token en relación con el conjunto total de tokens del tweet.

Entre tanto, la frecuencia inversa del documento para un término dado ($IDF_{(t)}$) fue computada como

$$IDF_{(t)} = \log\left(\frac{N}{DF_t}\right) + 1 \tag{2}$$

donde N corresponde al número total de documentos (tweets) en el dataset y DF_t es la frecuencia del documento para el término t (el número de tweets del dataset que contienen el término t); en el algoritmo de quanteda se suma un valor de 1 a esta ecuación para no ignorar los términos con un IDF de cero, es decir, que ocurren en todos los tweets del dataset.

Finalmente, para cada término se obtiene el TF–IDF como el producto de los dos anteriores, mediante la expresión

$$TF - IDF_{(t,d)} = TF_{t,d} * IDF_t \tag{3}$$

La vectorización se realizó por unigramas, donde cada palabra o cada fonema constituyó una variable predictora, es decir, se crearon *features* de un solo token. Con respecto a la bolsa de palabras del corpus de tweets, el algoritmo de vectorización se parametrizó para crear un diccionario únicamente con las palabras que tuvieran una frecuencia mínima de dos (2) en el data set de entrenamiento; esto, con el propósito de reducir la dispersión (*sparsity*) del conjunto de datos resultante (en cada idioma se identificaron inicialmente más de 4.000 palabras), por tratarse un problema regresión del tipo $p >> n$ (mayor número de variables en comparación con el número de observaciones), en el cual el riesgo de sobreajuste es muy alto por el alto grado de dimensionalidad. Para terminar, se debe añadir que no se utilizaron recursos adicionales (por ejemplo, lexicones disponibles o *feature extraction* de aspectos sintácticos de los tweets), pues el propósito del estudio fue analizar la posible contribución de los *features* fonéticos a la capacidad predictiva de los modelos creados únicamente a partir del léxico propio del corpus analizado.

Entrenamiento de los modelos de predicción

En la segunda fase se procedió al entrenamiento de diferentes modelos de predicción, basados en el uso de técnicas de regresión lineal mútliple. El factor experimental que varió durante la fase de entrenamiento fue el grado de reducción en la cantidad de variables predictoras. De esta forma, se partió de 1) una línea de base en la cual se entrenaron modelos de regresión con la totalidad de las palabras y fonemas identificadas en la fase de vectorización; posteriormente, 2) se entrenaron modelos de regresión después de aplicar técnicas de *feature selection* mediante regularización por redes elásticas, sin establecer un número fijo de variables a seleccionar; finalmente, 3) se entrenaron modelos de regresión después de aplicar técnicas de *feature selection* mediante regularización por redes elásticas, pero esta vez limitando el número de variables predictoras a una cantidad determinada. El propósito de esta reducción gradual en el número de variables fue disminuir el sobreajuste de los modelos obtenidos, así como aumentar la interpretabilidad de los resultados obtenidos.

De esta forma, en cada nivel de reducción de variables se entrenaron modelos solamente con palabras, solamente con fonemas, o combinando palabras y fonemas. Como se explicó anteriormente, el propósito de esta modelación fue identificar la capacidad específica de cada fonemá en español o en inglés para predecir la valencia emocional de los tweets (a partir de sus coeficientes de regresión β), así como establecer el grado de contribución de la inclusión de los

fonemas en esta tarea de predicción, con respecto a un modelo simple de *Bag of Words* normalizado mediante TF-IDF.

Tabla 4. *Modelos entrenados como parte de las variaciones experimentales del proyecto.*

Grado de reducción de los predictores	**Modelos**
Modelos con todos los tokens extraídos en la vectorización TF-IDF	1. Unigramas de palabras 2. Unigramas de fonemas 3. Unigramas de palabras y fonemas
Modelos con *Feature Selección* sin definir un número fijo de variables	4. Unigramas de palabras 5. Unigramas de fonemas 6. Unigramas de palabras y fonemas
Modelos con *Feature Selección* definiendo un número fijo de variables	7. Unigramas de palabras 8. Unigramas de palabras y fonemas (partiendo de todas las palabras y fonemas) 9. Unigramas de palabras y fonemas (partiendo de las palabras seleccionadas y agregando los fonemas)

Elaboración propia.

En términos generales, los modelos lineales están orientados a la regresión de una variable objetivo, la cual se espera que sea una combinación lineal de los *features* de entrada (palabras o fonemas). De esta forma, si $\hat{y}$ es el valor predicho (la valencia afectiva de un tweet, en este caso), el modelo se puede expresar como

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \cdots + w_{\mathrm{p}} x_{\mathrm{p}} \tag{4}$$

siendo w_0 el intercepto del modelo y $w = (w_1, \ldots, w_{\mathrm{p}})$ el vector de coeficientes de regresión estimado para las variables de entrada (los TF-IDF de los tokens de palabras o fonemas). De esta forma, el algoritmo de scikit-learn ajusta un modelo lineal con estas constantes, que minimiza la suma de los residos cuadráticos entre los valores observados y los predichos en la variable objetivo. Formalmente

$$\min_{w} \|Xw - y\|_2^2 \tag{5}$$

Para esto, se emplea el método ordinario de mínimos cuadrados (OLS, por su sigla en inglés), el cual, además de asumir la linealidad de la relación entre los componentes de la matriz X y la

variable objetivo, parte del supuesto de independencia entre los *features* utilizados para la definición del modelo y, por tanto, para la predicción de la variable objetivo (ausencia de multicolinealidad).

Ahora bien, al abordar un problema de predicción mediante un modelo de regresión lineal múltiple es posible encontrar situaciones en las cuales se dispone de un gran número de variables (algo común en las tareas de procesamiento del lenguaje natural, dada la amplía variedad léxica de los distintos idiomas), por lo que resulta importante determinar cuáles variables explicativas deberían entrar al modelo de predicción (González Vidal, 2015). Esto es así, si se tiene en cuenta que al incluir un número muy grande variables se corre el riesgo de sobreajustar el modelo, lo que implica que su capacidad de predicción en datos nuevos va a disminuir (poco sesgo, mucha varianza), mientras que si se dejan muy pocas variables el nivel de ajuste del modelo disminuye, pero es probable que al predecir nuevos resultados con datos diferentes a los del entrenamiento los resultados sean más consistentes (mayor sesgo, menor varianza). Además de enfrentar el clásico dilema de sesgo-varianza en el contexto del *machine learning*, la reducción del número de variables independientes en un modelo de regresión lineal múltiple disminuye el grado de multicolinealidad, por lo que aumenta la confiabilidad del modelo.

Si bien existe una gran variedad de técnicas para realizar este proceso de reducción, en este proyecto se empleó un método de mínimos cuadrados penalizados (regresión por regularización), los cuales imponen una penalización para reducir el valor de algunos de los valores del vector de coeficientes w, con el fin de evitar el sobreajuste del modelo e identificar las variables de mayor capacidad explicativa. Una de las preguntas con respecto a este tipo de métodos tiene que ver con el valor del parámetro de penalización, lambda (λ), el cual depende del algoritmo de reducción que se aplique (González Vidal, 2015).

Por una parte, los modelos lineales tipo Lasso (*Least Absolute Shrinkage and Selection Operator*), introducidos por Tibshirani en 1996 (Guerra de la Corte, 2016), realizan el proceso de selección de variables al añadir un factor de penalización a un modelo de mínimos cuadrados que lleve a cero (0) los coficientes w pequeños (norma L1), lo que corresponde a un proceso de optimización que formalmente puede expresarse como

$$\min_{w} \frac{1}{2n\ casos} \|Xw - y\|_2^2 + \lambda \|w\|_1 \qquad (6)$$

Por su parte, la regularización tipo Ridge realiza el proceso de regularización al aplicar una penalidad basada en la norma L2, con lo cual el parámetro de penalización es un valor cuadrático que permite minimizar el valor de los coeficientes del modelo, sin que los lleve necesariamente a cero como en una regresión Lasso, por lo que la función de costo en la una regresión Ridge se experesa como

$$\min_{w} \frac{1}{2n\ casos} \|Xw - y\|_2^2 + \lambda \|w\|_2^2 \quad (7)$$

Ahora bien, una red elástica es un método de regularización que combina los dos términos de penalización, por lo cual incluye las ventajas de los métodos Ridge (aumentar la capacidad explicativa de ciertas variables y disminuir la de otras, especialmente cuando hay una situación de multicolinealidad) y Lasso (disminuir el número de variables, al reducir algunos coeficientes a cero) (Zou & Hastie, 2005). Así, una regularización por redes elásticas permite reducir la dimensionalidad en un problema de aprendizaje automático, con la siguiente función de costo

$$\min_{w} \frac{1}{2n\ casos} \|Xw - y\|_2^2 + \lambda \|w\|_1 + \lambda \|w\|_2^2 \quad (8)$$

Dado que la regresión por redes elásticas es un método de reducción de variables, además de ayudar a encontrar una solución equlibrada entre sesgo y varianza, presenta diferentes ventajas. Por ejemplo, al seleccionar solamente algunas variables, la interpretabilidad del modelo mejora y se disminuye el grado de multicolinealidad entre los *features* empleados como predictores (Melkumova & Shatskikh, 2017).

Estos procedimientos de reducción de la dimensionalidad se llevaron a cabo con la librería glmnet de R (Friedman, Hastie, & Tibshirani, 2010), la cual reduce el número de *features* a partir de un modelo matemático de gradiente descendiente, que busca minimizar el indicador de error definido por el usuario, que en el presente proyecto fue el Error Cuadrático Medio (MSE, por su sigla en inglés). Para la identificación del valor óptimo de penalización en la red elástica (hiperparámetro lambda λ), se parametrizó el algoritmo con un proceso de validación cruzada (con 10 *foldings*, los definidos por defecto en la librería, extraídos del dataset de entrenamiento) y una secuencia de 1.000 valores de λ con el fin de identificar el óptimo para la función objetivo descrita

anteriormente. Dado que el algoritmo tiene un componente aleatorio, el proceso de regularización se repitió con cinco valores semilla distintos para verificar la consistencia de los resultados; dado que los resultados fueron consistentes (por ejemplo, el orden relativo de los coeficientes de regresión), en la sección de resultados se reportan los últimos valores obtenidos.

Evaluación del desempeño de los modelos de regresión

La tercera fase de procesamiento consistió en la evaluación del desempeño de los modelos extraídos. Para esto, se calcularon diferentes métricas de ajuste, tanto en el conjunto de entrenamiento como en el conjunto de prueba. Para empezar, dado que, en el concurso SemEval-2018 la métrica de evaluación fue la correlación de Pearson entre las puntuaciones de valencia asignadas por el modelo y las puntuaciones de valencia del *gold standard* (contenidas en el dataset de evaluación que no fue abordado en este proyecto), para el presente proyecto esta fue la primera métrica calculada. También se tuvieron en cuenta otras métricas sensibles al nivel de precisión en la predicción de la variable objetivo, la cual, al ser un escalar continuo entre 0 y 1 se considera como de "grano fino" (*fine grained*); es decir, el objetivo de la tarea vas más allá de la clasificación en categorías discretas más "gruesas", pues pretende acercarse lo más posible al valor más cercano posible al valor del *gold standard*.

El coeficiente de correlación de Pearson es una medición de la relación lineal entre dos variables que supone la normalidad de la distribución de las variables de origen y fluctúa entre -1 y 1. Se puede decir que a medida que su valor se acerca a uno (1) absoluto, la relación entre las variables es más fuerte (positiva o negativamente). Su cálculo se realiza a través de la expresión

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \tag{9}$$

donde $\bar{x}$ y $\bar{y}$ corresponden a la media de los vectores x y y, que en este caso equivalen a los valores reales y a los predichos por el modelo.

Además, se calcularon las métricas correspondientes al Error Medio Absoluto (MAE, por su sigla en inglés), la Raíz del MSE (RMSE, por su sigla en inglés) y el coeficiente de determinación (R^2) o Varianza explicada. El MAE es un indicador del grado de diferencia entre la predicción y el valor observado, que asume que las dos variables son expresiones del mismo fenómeno; así,

para n observaciones, equivale al promedio de los errores absolutos (en la misma escala de medida original) y se obtiene como

$$MAE(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| \tag{10}$$

El MSE es un estimador del promedio de los residuos cuadráticos. Por ser el segundo momento de la distribución de los errores de predicción incorpora la varianza y el sesgo de los estimadores del modelo; al calcular su raíz cuadrada se obtiene el RMSE, el cual se puede interpretar como el error estándar de la predicción (la desviación estándar de la distribución de los errores de predicción). Con n observaciones, su cálculo se realiza con la expresión

$$MSE(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2 \tag{11}$$

En cuanto al coeficiente de determinación (R^2) corresponde a la varianza explicada, es decir, la proporción de la varianza de la variable objetivo que es predecible a partir de los predictores incluidos en el modelo de regresión; en otras palabras, es una medida de qué tan bien los valores observados de la variable objetivo son replicados por el modelo. Este indicador de bondad de ajuste de un modelo puede arrojar valores negativos, implicando que la media de los valores empleados en el entrenamiento corresponde a un modelo de mejor ajuste (el modelo estimado puede ser arbitrariamente peor). De esta forma, si $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ es la sumatoria de los residuos cuadráticos y $\bar{y}_i$ el promedio de los valores observados, el R^2 se calcula como

$$R^2(y_i, \hat{y}_i) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{12}$$

Finalmente, se calculó el Criterio de Información de Akaike (AIC, por su sigla en inglés), un indicador de la calidad relativa de un modelo, en comparación con otros modelos. Este cálculo se basó en la expresión

$$AIC = -2 * \log(\mathrm{L}) + 2\mathrm{k} \quad (13)$$

Donde L corresponde a la función de verosimilitud estimada a partir de la desviación entre el modelo obtenido y el modelo teórico esperado (o modelo nulo) y k al número de parámetros del modelo (incluyendo el intercepto). Dado que el AIC también se puede interpretar como un indicador del grado de parsimonia de un modelo, entre más bajo sea su valor, mejor será un modelo (como su valor puede oscilar entre –∞ e ∞, será más bajo entre más se acerque a –∞).

Resultados

Para empezar, se presentan algunas estadísticas descriptivas de la variable objetivo (Tabla 5). Como es de esperar, la media de las puntuaciones es cercana al valor medio de la escala (,5) en todos los datasets, con una desviación estándar equivalente al 20% de la escala y un error estándar que, por el tamaño de muestra, es menor en los conjuntos de entrenamiento. Con respecto a la distribución de esta variable, se encuentra un patrón cercano a la normalidad estadística, como se presenta en la Figura 3.

Tabla 5. *Estadísticas descriptivas de la valencia emocional en los diferentes datasets.*

Dataset	N	Media	DE	EE	IC (95%)	
Español – Entrenamiento	1.562	,500	,214	,005	,489	,510
Español – Prueba	229	,516	,218	,014	,487	,544
Inglés – Entrenamiento	1.181	,501	,209	,006	,489	,512
Inglés - Prueba	449	,485	,226	,011	,464	,506

Elaboración propia.

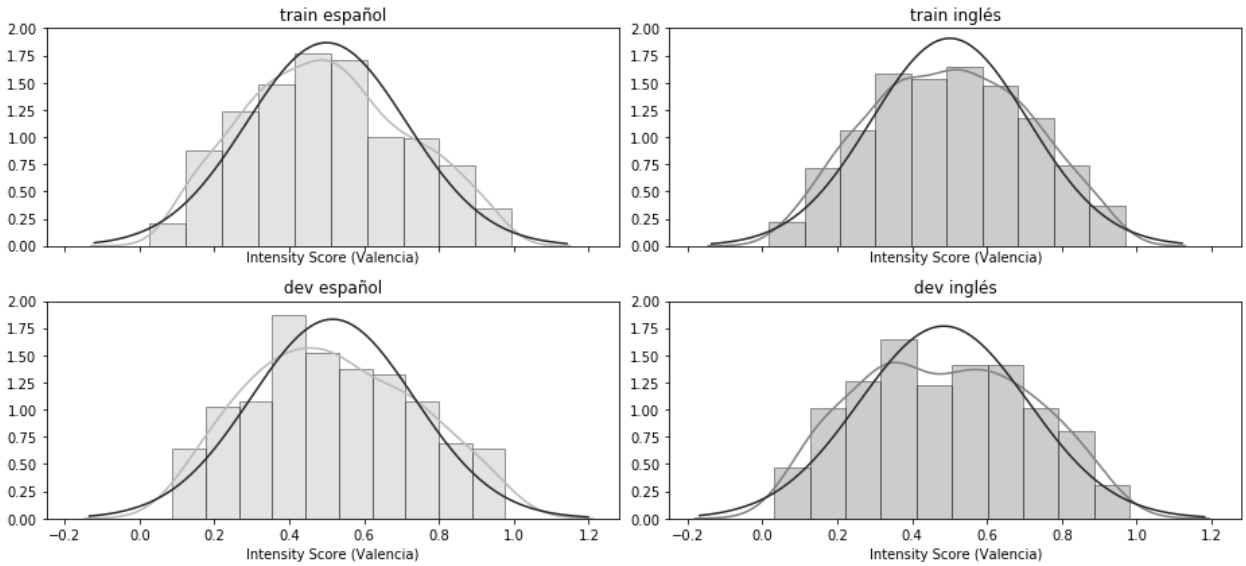


Figura 3. Distribución de las puntuaciones de valencia emocional en los diferentes datasets.

A continuación, se describen los resultados obtenidos para los diferentes modelos de regresión evaluados.

Primer conjunto de modelos: con todos los unigramas de palabras y fonemas extraídos en la vectorización

Al aplicar el algoritmo de vectorización con normalización TF-IDF en los datasets, se extrajeron 4.174 unigramas de palabras en español y 4.342 en inglés; no obstante, como se explicó anteriormente, el diccionario se redujo solamente a las palabras con una ocurrencia mínima de dos en el dataset, por lo cual quedaron 1.403 palabras en español y 1.437 en inglés. Igualmente, se identificaron 27 unigramas de fonemas en español y 38 en inglés.

A partir de estas variables predictoras, sin realizar ningún tipo de *feature selection* por regularización, se generaron los modelos de regresión que pretendían servir como línea de base. La Tabla 6 contiene las métricas empleadas para la evaluación del desempeño de los modelos en cada uno de los datasets.

Tabla 6. *Métricas de evaluación del desempeño del modelo base y los modelos experimentales, sin aplicar técnicas de regularización.*

Modelo	Dataset	Pearson	R^2	RMSE	MAE	AIC
1. Unigramas de palabras	Español – Entrenamiento	,97	,95	,05	,03	-2327,59
	Español – Prueba	,04	,00	1,04	,61	-
	Inglés – Entrenamiento	1,00	1,00	,01	,00	-4500,57
	Inglés – Prueba	,06	,00	63,21	32,15	-
2. Unigramas de fonemas	Español – Entrenamiento	,31	,10	,20	,17	-492,40
	Español – Prueba	,24	,06	,21	,17	-
	Inglés – Entrenamiento	,33	,11	,20	,16	-401,16
	Inglés – Prueba	,11	,01	,23	,19	-
3. Unigramas de palabras + unigramas de fonemas	Español – Entrenamiento	,98	,96	,04	,02	-2534,01
	Español – Prueba	,04	,00	1,01	,61	-
	Inglés – Entrenamiento	1,00	1,00	,01	,00	-5603,22
	Inglés – Prueba	,04	,00	1716,15	770,94	-

Elaboración propia.

Como era de esperar, las métricas muestran que hay un sobreajuste, consecuencia del alto número de características de entrada sin proceso de depuración. Por ejemplo, los valores del error medio en los modelos 1 y 3 generados con los tweets en inglés, exceden de lejos el valor máximo de la escala de valencia. No se profundiza el análisis de estos resultados, pues carecen de validez.

Segundo conjunto de modelos: con *feature selection* de unigramas de palabras y fonemas

A continuación, se expone el desempeño de los modelos después de aplicar la técnica de regularización por redes elásticas para eliminar las unidades lingüísticas que no contribuían a la predicción de la valencia emocional de los tweets, y de esta forma reducir la complejidad del modelo para lograr una mayor parsimonia e interpretabilidad, así como una menor probabilidad de sobreajuste. Como se explicó en la sección de método, dado que la elección del híper parámetro de penalización lambda (λ) puede ser arbitraria, el algoritmo de cvglmnet añadió un componente de validación cruzada para iterar el ajuste del modelo hasta encontrar el valor más cercano al óptimo posible, con el fin de reducir el valor de la función de costo (los residuos cuadráticos, en el caso de un modelo de regresión lineal).

En esta segunda etapa de análisis se evaluaron tres tipos de modelos: solamente con unigramas de palabras, solamente con unigramas de fonemas y con la combinación de unigramas de palabras y fonemas. Al aplicar los diferentes vectorizadores en los datasets, se extrajeron 319 unigramas de palabras en español y 382 en inglés (Modelo 4), 18 unigramas de fonemas en español y 20 en inglés (Modelo 5), y 338 unigramas de palabras + fonemas en español y 336 en inglés (Modelo 6). La Tabla 7 presenta las métricas obtenidas.

Tabla 7. *Métricas de evaluación del desempeño del modelo base y los modelos experimentales, después de aplicar técnicas de regularización.*

Modelo	Dataset	Pearson	R^2	RMSE	MAE	AIC
4. Unigramas de palabras	Español - Entrenamiento	,78	,62	,14	,12	-599,50
	Español – Prueba	,59	,35	,18	,15	-
	Inglés – Entrenamiento	,86	,75	,12	,10	-728,51
	Inglés – Prueba	,51	,26	,19	,16	-
5. Unigramas de fonemas	Español – Entrenamiento	,30	,09	,20	,17	-29,64
	Español – Prueba	,23	,05	,21	,18	-
	Inglés – Entrenamiento	,32	,10	,20	,16	-34,97
	Inglés – Prueba	,07	,01	,23	,19	-
6. Unigramas de palabras + unigramas de fonemas	Español – Entrenamiento	,80	,64	,14	,11	-634,94
	Español – Prueba	,59	,35	,18	,15	-
	Inglés – Entrenamiento	,85	,72	,12	,10	-638,24
	Inglés – Prueba	,48	,23	,20	,16	-

Elaboración propia.

Una primera observación con respecto a estos resultados es que en, en general, se logra reducir notablemente el nivel de sobreajuste, aunque el problema persiste en alguna medida. En relación con el objetivo de este proyecto (determinar el aporte de los fonemas a la predicción de la valencia), se encuentra un desempeño ligeramente superior en el modelo combinado al compararlo con el modelo de solo palabras, los que se ve reflejado específicamente en los valores del AIC (las demás métricas tienen un comportamiento muy similar). En el modelo combinado se conservaron 10 fonemas en español y 7 fonemas en inglés (ver Tablas 8 y 9), los cuales, al parecer, contribuyen a mejorar la predicción especialmente en los valores medios, como se puede ver en la Figura 4.

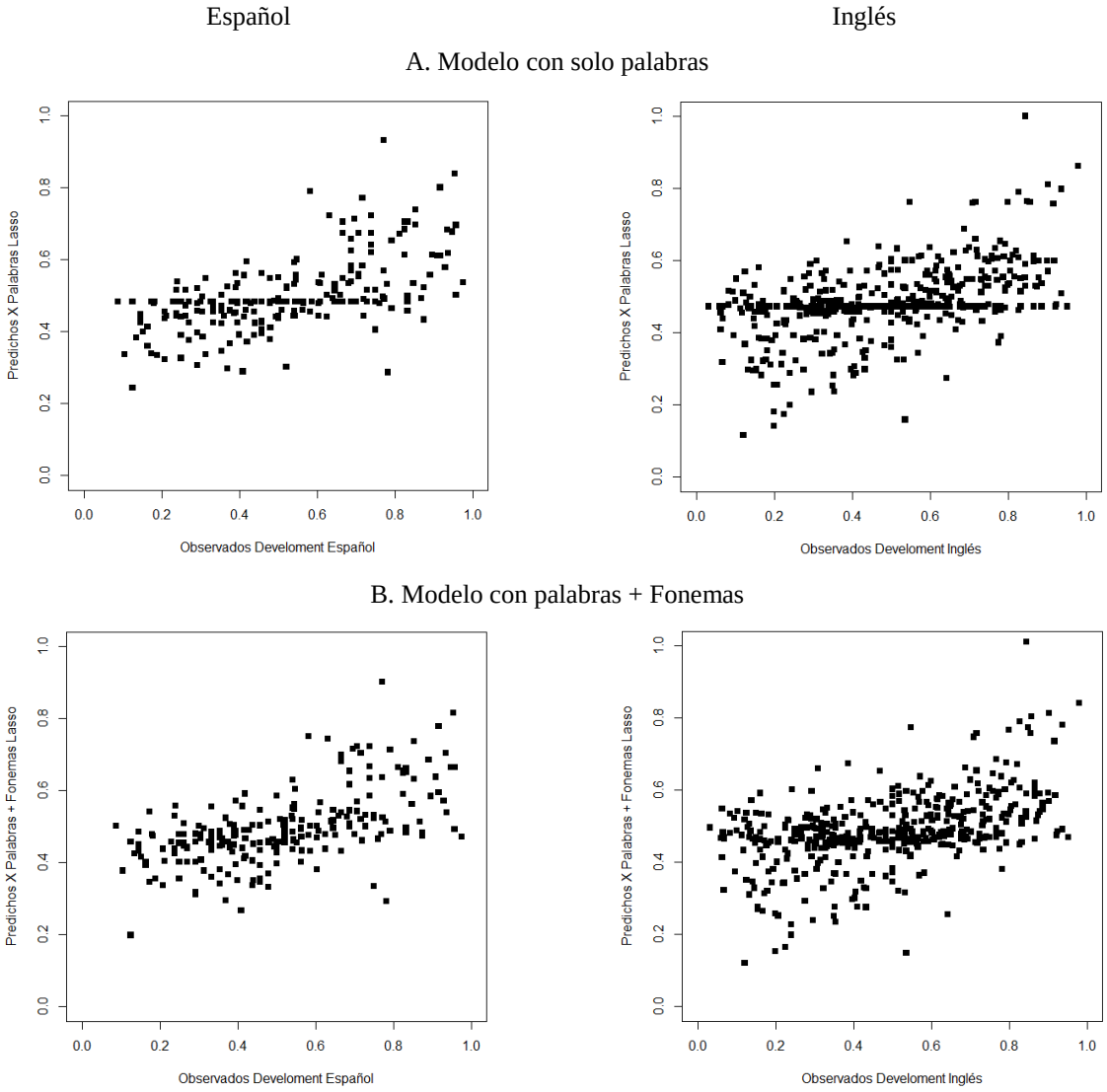


Figura 4. Valores observados vs predichos, en los datasets de prueba (development) en español y en inglés, a partir de los modelos de regresión después de la regularización.

Una posible interpretación de los resultados expuestos es que, aunque no se aumente el poder explicativo del modelo al agregar los fonemas, estos últimos podrían sustituir algunas palabras en los modelos base. Las tablas 8 y 9 presentan los fonemas seleccionados (y eliminados) mediante el mecanismo de *feature selection* para el modelo 5 (solo fonemas) y para el modelo 6 (palabras + fonemas), indicando los coeficientes de regresión en cada caso.

Tabla 8. *Fonemas seleccionados y eliminados en español, indicando sus coeficientes de regresión.*

Fonema	Sonido	Rasgos distintivos mínimos (https://es.wikipedia.org/wiki/Fonema)	Beta solo fonemas	Beta palabras + fonemas
t͡s	ts	Africada alveolar sorda	,072	Eliminado
x	j, g	Fricativo velar (grafías g y j)	,055	,047
ɡ	gu	Obstruyente velar sonoro (grafías g y gu)	,036	Eliminado
f	f	Labial, fricativo, sordo, oral	,031	Eliminado
b	b, v	Obstruyente bilabial sonoro (grafías: b, v y w)	,029	Eliminado
i	i, y	Vocálico palatal y apertura mínima	,025	Eliminado
a	a	Vocálico de apertura máxima	,014	Eliminado
t͡ʃ	ch	Africada pos alveolar sorda	,0003	Eliminado
ɾ	r	Vibrante simple (grafía -r-, -r)	-,004	-,015
j	i	Vocálico	-,011	Eliminado
t	t	Oclusivo (coronal-)alveolar sordo	-,014	Eliminado
u	u	Vocálico velar de apertura mínima	-,024	-,012
k	c, k, q	Oclusivo velar sordo (grafías c, qu y k)	-,024	-,001
t͡ɬ	ths	Africada lateral alveolar sorda	-,034	Eliminado
p	p	Oclusivo (bi)labial sordo	-,067	-,042
r	rr	Vibrante múltiple (grafía -rr-, r-)	-,068	-,030
n	n	Nasal (coronal-)alveolar	-,075	-,027
d	d	Obstruyente coronal-alveolar sonoro (alófonos: [d], [ð])	-,089	-,083
s	c, s, z	Fricativo (coronal-)alveolar (grafía s)	Eliminado	Eliminado
e	e	Vocálico palatal de apertura media	Eliminado	-,026
o	o	Vocálico velar de apertura media	Eliminado	Eliminado
l	l	Lateral (coronal-)alveolar	Eliminado	-,001
ʝ	ll, y	Sonorante palatal (grafía y, en zonas yeístas también ll)	Eliminado	Eliminado
m	m	Nasal labial (alófono usual: [m], ante a una f: [ɱ])	Eliminado	Eliminado
w	u	Vocálico	Eliminado	Eliminado
ɲ	ñ	Nasal palatal	Eliminado	Eliminado
ʃ	h	Fricativo, laringal	Eliminado	Eliminado

Elaboración propia.

Tabla 9. *Fonemas seleccionados y eliminados en inglés, indicando sus coeficientes de regresión.*

Fonema	Ejemplos Sonidos	Rasgos distintivos mínimos	Beta solo fonemas	Beta palabras + fonemas
l	leg, little	Alveolar lateral	,270	,117
j	yes, yellow	Palatal aproximante	,118	Eliminado
i	see, heat	Vocálico	,078	Eliminado
m	man, lemon	Nasal plosiva (nasal)	,071	Eliminado
t͡ʃ	check, church	Palato-alveolar africada	,064	,050
ʊ	put, could	Vocálico	,034	,010
θ	think, both	Inter-dental fricativa	,026	Eliminado
d͡ʒ	just, large	Palato-alveolar africada	,016	Eliminado
v	voice, five	Labio-dental fricativa	,011	Eliminado
ɡ	give, flag	Velar plosiva	,003	Eliminado
ɔ	call, four	Vocálico	-,004	-,0001
p	pet, map	Nasal plosiva	-,016	Eliminado
ɹ	real, grow	Alveolar aproximante	-,026	Eliminado
ð	the, mother	Inter-dental fricativa	-,026	Eliminado
s	sun, miss	Alveolar fricativa	-,030	Eliminado
n	no, ten	Alveolar plosiva (nasal).	-,040	Eliminado
æ	cat, black	Vocálico	-,043	-,004
t	tea, getting	Alveolar plosiva	-,077	Eliminado
k	cat, back	Velar plosiva	-,085	Eliminado
d	did, lady	Alveolar plosiva	-,093	-,015
e	met, bed	Vocálico	Eliminado	Eliminado
ɹ̩	bird, heard	Vocálico	Eliminado	Eliminado
ŋ	sing, finger	Velar plosiva (nasal)	Eliminado	Eliminado
z	zoo, lazy	Alveolar fricativa	Eliminado	Eliminado
ɛ	turn, learn	Vocálico	Eliminado	Eliminado
ə	away, cinema	Vocálico	Eliminado	Eliminado
a	law, caught	Vocálico	Eliminado	Eliminado
ɑ	arm, father	Vocálico	Eliminado	Eliminado
u	blue, food	Vocálico	Eliminado	Eliminado
w	wet, window	Velar aproximante	Eliminado	Eliminado
o	boat, low	Vocálico	Eliminado	Eliminado
ɪ	hit, sitting	Vocálico	Eliminado	Eliminado
ʌ	cup, luck	Vocálico	Eliminado	-,010
b	bad, lab	Nasal plosiva	Eliminado	Eliminado
f	find, if	Labio-dental fricativa	Eliminado	Eliminado
h	how, hello	Glotal fricativa	Eliminado	Eliminado
ʃ	she, crash	Palato-alveolar fricativa.	Eliminado	Eliminado
ʒ	pleasure, vision	Palato-alveolar fricativa.	Eliminado	Eliminado

Elaboración propia.

Aunque los fonemas por sí solos tienen una capacidad predictiva muy pequeña, es importante notar que, al combinarlos con las palabras, los fonemas que se conservan en los modelos de regresión mantienen el signo (y en inglés el orden relativo), aunque llama la atención que en los modelos combinados aparecen algunos fonemas que habían sido eliminados. También es interesante agregar que en español se conservan prácticamente solo fonemas negativos y que en los dos idiomas solamente hay coincidencia en un fonema (d), el cual, en ambos casos, tiene el valor negativo más alto.

Tercer conjunto de modelos: con *feature selection* de unigramas de palabras y fonemas, limitado a un número fijo de tokens

El último conjunto de modelos se basó nuevamente en una selección de variables con redes elásticas, pero limitando el número de variables predictoras a una cantidad determinada. Si bien existe controversia en relación con este punto, una de las reglas de oro sugiere un mínimo de 20 casos por variable (Austin & Steyerberg, 2015); así, puesto que el dataset de entrenamiento en español estaba conformado por 1.562 tweets y el dataset en inglés por 1.181 tweets, se parametrizó el algoritmo de regularización para elegir los 78 atributos más importantes en español y los 59 más importantes en inglés. Además de reducir el riesgo de sobreajuste y ganar en parsimonia, esto facilita la interpretación de los resultados, en términos de las unidades léxicas y fonológicas que aportan a la predicción de la valencia emocional. Para este tercer conjunto de modelos se hicieron las siguientes variaciones: un primer modelo base solamente con unigramas de palabras (modelo 7); un segundo modelo en el cual se iniciaba con todas las palabras y todos los fonemas, para seleccionar las 78 o 59 unidades más importantes en español y en inglés, respectivamente (modelo 8); y un tercer modelo, en el cual se tomaban las 78 o 59 palabras seleccionadas en el modelo 7 y se combinaron con los 18 fonemas seleccionados en español o con los 20 fonemas seleccionados en inglés (modelo 9). La siguiente tabla presenta las métricas de desempeño alcanzadas.

Tabla 10. *Métricas de evaluación del desempeño del modelo base y los modelos experimentales, después de aplicar técnicas de feature selection con un número restringido de atributos.*

Modelo	Dataset	Pearson	R^2	RMSE	MAE	AIC
7. Unigramas de palabras	Español - Entrenamiento	,60	,36	,18	,15	-139,84
	Español – Prueba	,53	,28	,19	,16	-
	Inglés – Entrenamiento	,63	,39	,18	,15	-105,24
	Inglés – Prueba	,41	,17	,21	,18	-
8. Unigramas de palabras + unigramas de fonemas	Español – Entrenamiento	,62	,39	,18	,15	-137,18
	Español – Prueba	,51	,26	,19	,16	-
	Inglés – Entrenamiento	,64	,41	,18	,15	-104,60
	Inglés – Prueba	,38	,14	,21	,18	-
9. Unigramas de palabras + unigramas de fonemas	Español – Entrenamiento	,70	,49	,15	,12	-159,04
	Español – Prueba	,54	,30	,18	,15	-
	Inglés – Entrenamiento	,74	,54	,14	,11	-132,02
	Inglés – Prueba	,44	,20	,21	,17	-

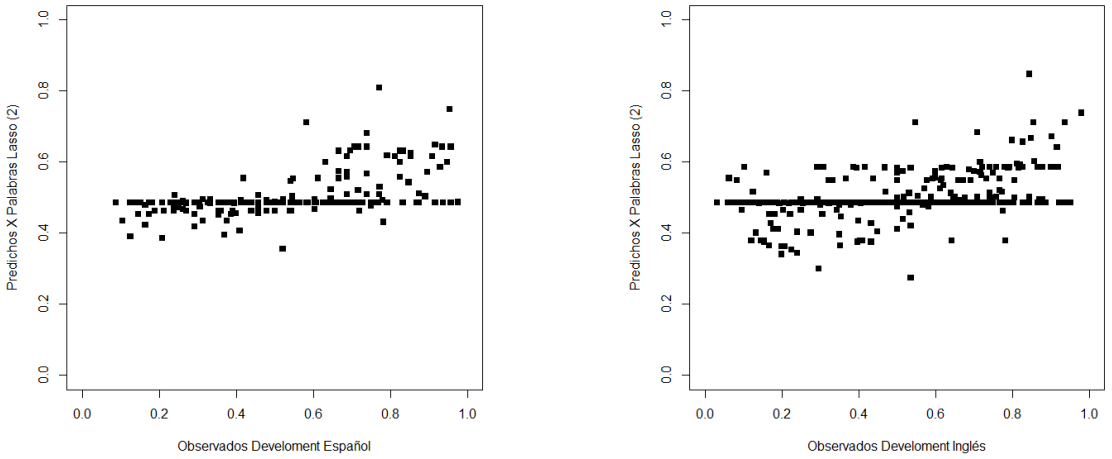
Elaboración propia

Con respecto al modelo de solo palabras (7) y al primer modelo combinado (8), se encuentran unas métricas inferiores a las obtenidas en la etapa anterior (disminuyen la correlación y el coeficiente de determinación, mientras que aumentan los indicadores de error y el AIC), aunque también hay un menor grado de sobreajuste. Ahora bien, al añadir los fonemas a los modelos de predicción, se encuentran mejoras en el desempeño; por ejemplo, además de las palabras en el modelo 8 fueron seleccionados cinco fonemas en español (uno positivo: x, y cuatro negativos: n, r, p, d) y tres en inglés (todos positivos: l, t͡ʃ y j). Sin embargo, en este modelo no se observa una mejora notable en los indicadores de ajuste. Finalmente, en el modelo 9, que combinó las palabras identificadas en el modelo 7 y los fonemas seleccionados en el modelo 5, sí se evidencia un mejor desempeño pues aumentan la correlación y el R^2, mientras que disminuyen los errores y el AIC.

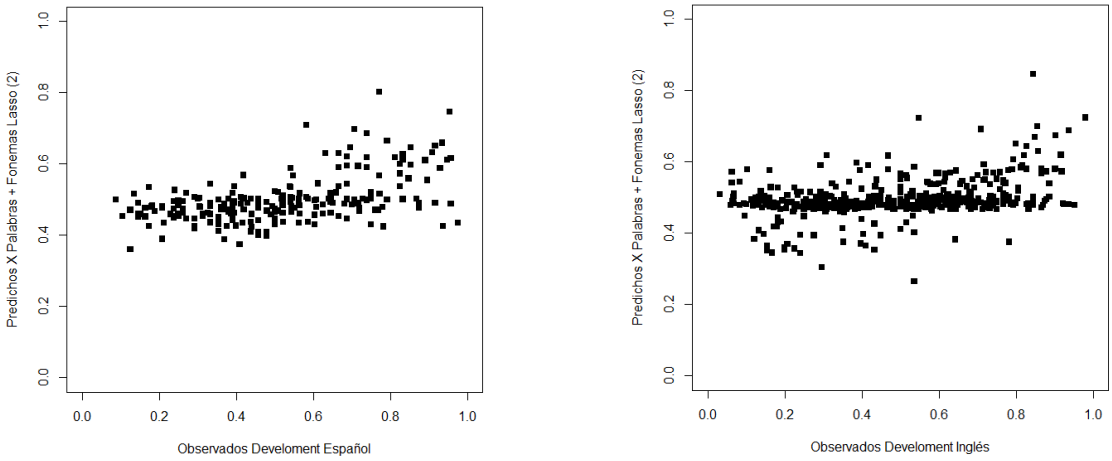
Con respecto a la distribución de las predicciones y su relación con los valores observados (Figura 5), se vuelve a observar el efecto de agregar los fonemas. Por una parte, se logra una mayor diferenciación de las valencias en los niveles intermedios de la escala, que tienden a concentrarse casi exclusivamente en ,5 cuando se usan solamente palabras (posible contribución a la desambiguación cuando solamente hay atributos léxicos). Igualmente, parece haber más valores predichos por debajo del punto medio, lo que sugiere que el aporte de los fonemas podría ser más importante en relación con los estados afectivos negativos.

Español Inglés

A. Modelo con solo palabras



B. Modelo con palabras + Fonemas (partiendo de todos los tokens de palabras y fonemas)



C. Modelo con palabras + Fonemas (partiendo de los tokens de palabras y añadiendo los fonemas)

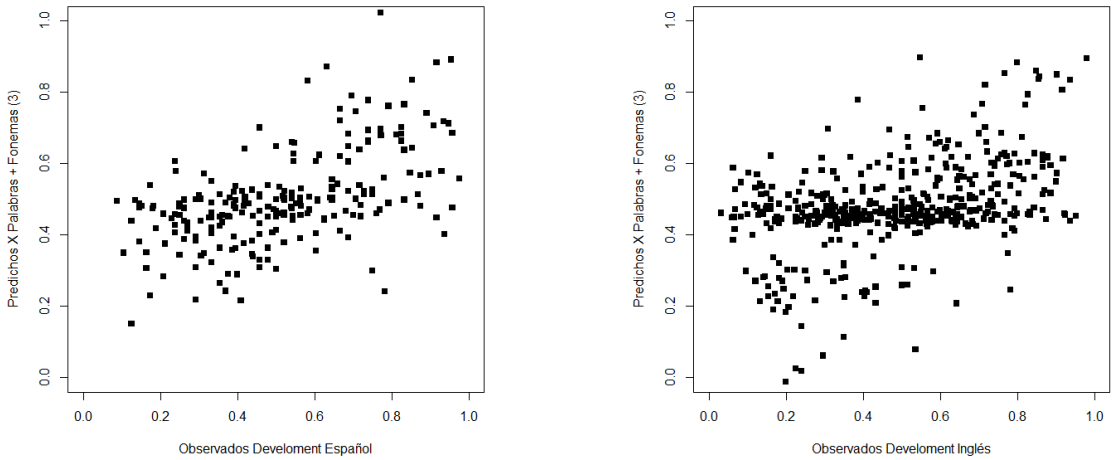


Figura 5. Valores observados vs predichos, en los datasets de prueba (development) en español y en inglés, a partir de los modelos de regresión después de la regularización (2).

Finalmente, al comparar el comportamiento de los fonemas por sí solos (modelo 5) o en combinación con las palabras (modelo 9), se observa que en español los siguientes fonemas cambian de signo: t (de ,031 a -,014), g (de -,003 a ,036), f (de -,008 a ,031), a (de -,054 a ,014), mientras que en inglés lo hacen los siguientes fonemas: s (de ,029 a -,03), n (de ,0001 a -,4), g (de -,005 a ,003) y θ (de -,007 a ,026). El resto de los coeficientes de los fonemas muestra coherencia en los signos, en incluso en las magnitudes.

Discusión

Una primera observación relevante es que se evidencia un nivel de sobreajuste en los modelos de regresión entrenados (más acentuado en inglés que en español), aunque este va disminuyendo a medida que se reduce el número de atributos de entrada. Una posible explicación para este comportamiento tiene que ver con el grado de normalización de los tokens de palabras extraídos a partir del dataset de entrenamiento, pues, como se indicó anteriormente, para conservar los detalles fonológicos que pudieran estar asociados a ciertas inflexiones del lenguaje (como aumentativos o diminutivos) no se realizó un proceso de lematización y estemizado; por esta razón, es posible que al vectorizar el diccionario del conjunto de prueba (*development*) a partir del contenido del dataset de entrenamiento el grado de coincidencia de las palabras disminuyera, por ejemplo por palabras combinadas que no aparecen en el dataset de prueba o mal escritas en alguno de los conjuntos, afectando el poder de generalización de los modelos de regresión entrenados. Igualmente, un modelo de Bag of Words (aunque vectorizado mediante el TF-IDF) tiene un alcance limitado, especialmente cuando los textos expresan contenidos emocionales de forma indirecta o implícita (Alhamdan, 2019). En todo caso, a favor de los resultados obtenidos se debe indicar que las palabras identificadas como predictores en los dos idiomas, así como sus coeficientes de regresión (ver Apéndice 1), son coherentes con la valencia correspondiente (palabras positivas con signo positivo y palabras negativas con signo negativo), lo que asegura un nivel de validez de los resultados obtenidos.

Dicho lo anterior, es posible afirmar que los resultados de la presente investigación resultan consistentes con la hipótesis del simbolismo de los sonidos, en tanto, por sí solos, los fonemas mostraron cierta capacidad predictiva de la valencia emocional de dos corpus de tweets en español e inglés. Además, al utilizar los fonemas como variables adicionales en un modelo basado en el contenido léxico de los datasets, se encuentra una contribución pequeña pero consistente en los dos idiomas; específicamente, esta contribución se evidencia en dos aspectos: la inclusión de los fonemas permite 1) diferenciar mejor las valencias medias de los tweets y 2) pronosticar mejor las valencias negativas de los tweets. Esta pequeña contribución resulta de valor, pues anteriormente se ha fallado al intentar identificar pistas vocales que permitan diferenciar confiablemente los niveles de valencia (a diferencia del efecto de los fonemas sobre el grado de *arousal*), lo que se explicaría por la diferencia entre estos dos atributos: el *arousal* remite al estado fisiológico que acompaña la reacción a un estímulo, por lo que podría asociarse al estado emocional del emisor

del lenguaje, mientras que la valencia remite al proceso evaluativo y, por lo tanto, cognitivo y de orden superior, lo que haría más difícil su detección a un nivel sub léxico mediante un mapeo acústico consistente (Aryani, Conrad, Schmidtke, & Jacobs, 2018).

Desde otra perspectiva, los resultados obtenidos en el presente estudio tienden a ser coherentes con los reportados por Adelman, Estes, & Cossu (2018), quienes trabajaron con una muestra de palabras en cinco idiomas (12.847 en inglés, 13.935 en español, 4.270 en holandés, 2.900 en alemán y 2.902 en polaco). Estos investigadores emplearon modelos de regresión lineal jerárquica, verificando la capacidad predictiva de la valencia emocional de las palabras, al controlar la importancia afectiva (*arousal*) y los aspectos léxicos (longitud, frecuencia y diversidad contextual de las palabras); igualmente, se demostró que esta capacidad predictiva se asocia a *features* específicos de los fonemas, como su posición o la forma de articulación de las consonantes. En todo caso, es importante aclarar que estos hallazgos fueron obtenidos tomando como unidad de análisis los componentes léxicos (se predijo la valencia de las palabras), más que estructuras de mayor complejidad como frases, oraciones o –como en el presente proyecto– tweets. Además, esta concordancia en los resultados es relevante, si se tiene en cuenta que la muestra de palabras identificadas en el nuestro estudio fue mucho menor a la de los autores que se están referenciando.

En el estudio de Adelman, Estes, & Cossu, los valores del coeficiente de determinación múltiple (R^2) para los modelos de regresión lineal con los fonemas en inglés y en español están alrededor de ,014 (1,4% de la varianza explicada, solo por los fonemas, para las valencias de las palabras). Si se observan los resultados obtenidos en la presente investigación a través de los modelos con *feature selection* se encuentra que, empleando únicamente los fonemas para predecir las valencias de los tweets, los coeficientes de determinación son legeramente mejores en español (R^2 = 5% en el conjunto de prueba), y similares en inglés (R^2 = 1%). Además, en nuestro estudio resulta importante anotar que el valor del coeficiente de determinación aumentó cuando al modelo de solo palabras se le adicionaron los fonemas como variables predictoras (en español pasa del 28% al 30% en el modelo con el segundo filtro a través de la regularización y en inglés del 17% al 20%). Otro contraste con este estudio es que Adelman y su equipo identificaron un 36% de fonemas en español y un 45% de fonemas en inglés, los cuales tuvieron coeficientes de regresión para la valencia de las palabras estadísticamente significativos, contra un 67% de fonemas en español y un 53% en inglés, retenidos como predictores de la valencia de los tweets en la presente investigación. Sin embargo, hay que tener en cuenta que ellos emplearon una metodología con un

enfoque estadístico “tradicional”, mientras que nosotros nos basamos en una técnica más característica del *machine learning* (la regresión por redes elásticas, que no selecciona las variables por un criterio de significancia estadística, sino buscando reducir los coeficientes de regresión de las variables con menor capacidad explicativa a partir del parámetro de penalización).

Por otro lado, es posible que al asignar una valencia afectiva global a un tweet se esté soslayando la contribución de los fonemas. Esta línea argumentativa se puede apoyar en el hecho de que en un mismo tweet pueden existir múltiples contenidos emocionales, incluso contradictorios entre sí, como de hecho se plantea en otra de las tareas del concurso SemEval-2018. Por ejemplo, Jabreel & Moreno (2019), al abordar esta tarea de clasificación multi etiqueta lograron un desempeño superior al identificado en el estado del arte (*accuracy* = ,59), mediante técnicas de *deep learning* basadas en la asociación específica de cada elemento del tweet (incluyendo la identificación del usuario, así como palabras, emoticones y signos de puntuación) con 11 emociones diferentes. En este sentido, se hace necesario verificar si la contribución de los fonemas es más específica en relación con emociones concretas (bien sea en términos categóricos o dimensionales), más que con la dimensión general de valencia emocional. En síntesis, puede que la dimensión de valencia emocional de un tweet tenga poca sensibilidad al efecto fonológico global, pues este podría “diluirse” al combinar palabras que tienen una orientación emocional diferente.

Llevando este argumento un paso adelante, si un tweet puede contener múltiples etiquetas, también sería un espacio multidimensional. Además, cada fonema podría representar por sí mismo “un vector de atributos fonéticos” (Aryani, Conrad, Schmidtke, & Jacobs, 2018), asociados a sus propiedades afectivas. Lo que implicaría que la valencia anotada en los *datasets* de entrenamiento y prueba podría tener problemas de validez. Se debe precisar que, si bien la calificación asignada por los anotadores de SemEval-2018 mostró un grado de confiabilidad aceptable (medido a partir de la técnica de Kuder-Richardson, como se explicó en la sección de método), esto no implica que tenga validez, pues desde un punto de vista conceptual, la confiabilidad se refiere a la precisión y consistencia de la medida, mientras que la validez remite a la capacidad de la medida de reflejar realmente el atributo de interés (AERA, APA, NCME, 2014); es decir, es posible que la valencia asignada por diferentes calificadores resulte consistente y estable, pero eso no significa que realmente represente la valencia emocional del contenido del tweet, teniendo en cuenta que finalmente los anotadores fueron personas legas sin conocimientos expertos en psicología de las

emociones, un campo que, como otros campos de la psicología, presenta grandes controversias teóricas. En otras palabras, la valencia emocional asignada podría representar la valencia emocional "percibida" por los anotadores, más que la valencia emocional "real" asociada a los autores de los tweets.

En este sentido, los siguientes ejemplos ilustran cómo en un mismo tweet, a pesar de tener una valencia global más positiva o más negativa (recuérdese que valores menores a ,5 indican valencia negativa), pueden coincidir palabras de orientación positiva y negativa; es interesante notar, además, que en inglés la orientación de la valencia predicha cambia con respecto a la original observada.

Tabla 11. *Ejemplos de Tweets, que incluyen palabras con diferente valencia emocional (las palabras de coeficiente positivo aparecen subrayadas y las de coeficiente negativo en negrita).*

Tweet	**Dataset de prueba (development)**	
	Valencia observada	**Última valencia predicha**
*"este Tito me hace **llorar** de la <u>risa</u>"*	,848	,574
*"Que **triste** lo mio, Ni <u>amigas</u> tengo"*	,104	,350
*"boys are so **awful** but also so <u>amazing</u>.... right?"*	,536	,399
*"And I'm really pissed the fuck off because I do a <u>good</u> job of keeping my kid well because I don't like to see her sick and **sad**"*	,121	,531

Elaboración propia.

A propósito de las diferencias encontradas entre resultados obtenidos en español y en inglés, estas se pueden analizar en dos sentidos. Por una parte, el desempeño global de los modelos resultó mejor en español, y, por otro lado, son diferentes los fonemas identificados y sus coeficientes de regresión específicos (en signo y en magnitud), hallazgos que sugieren que el efecto de la fonética sobre la iconicidad fonológica, a pesar de ser problablemente un mecanismo desarrollado a lo largo de la evolución filogenética del *homo sapiens sapiens*, es un fenómeno que no tiene una única expresión universal, sino que puede manifestarse a través de patrones de simbolismo de los sonidos específicos para distintos idiomas (sin hablar de la variedad dialectal que puede haber en un mismo idioma). Estas premisas se ven sustentadas por nuestro estudio, como se observa en la Tabla 12,

que presenta los fonemas con coeficientes de regresión estadísticamente significativos para preedecir la valencia de las palabras en el estudio de Adelman, Estes, & Cossu (2018), en comparación con los coeficientes de regresión obtenidos para los fonemas que coinciden en el presente estudio, 8 en español y 7 en inglés, aunque llama la atención que en español hubo tres cambios de signo en el coeficiente (señalados en negrita dentro de la tabla).

Tabla 12. *Comparación entre los coeficientes de regresión estadísticamente significativos para los fonemas del estudio de Adelman, Estes, & Cossu (2018) y los coeficientes de regresión para el presente estudio (modelo de solo fonemas).*

Idioma	Fonema	Beta presente estudio	Beta del estudio de Adelman et al.	Sonido	Rasgos distintivos mínimos
Español	x	,055	,094*	j, g	Fricativo velar
	ɡ	,036	,177*	gu	Obstruyente velar sonoro
	b	,029	,135**	b, v	Obstruyente bilabial sonoro
	t͡ʃ	**-,0003**	**,155***	**ch**	**Africada pos alveolar sorda**
	ɾ	**-,004**	**,078****	**r**	**Vibrante simple**
	j	-,011	-,158**	i	Vocálica semiconsonante (alófono)
	n	**-,075**	**,058***	**n**	**Nasal (coronal-) alveolar**
	d	-,089	-,279**	d	Obstruyente coronal-alveolar sonoro
Inglés	t͡ʃ	,064	,122*	check, church	Palato-alveolar africada
	i	,078	,099**	see, heat	Vocálica
	ʊ	,034	,250*	put, could	Vocálica
	v	,011	,144**	voice, five	Labio-dental fricativa
	s	-,030	-,087**	sun, miss	Alveolar fricativa
	ɹ	-,026	-,061*	real, grow	Alveolar aproximante
	d	-,093	-,188**	did, lady	Alveolar plosiva

* Estadísticamente significativos
** Estadísticamente significativos, aún después de una correción Bonferroni
Elaboración propia.

De la anterior tabla se destacan varios aspectos. Para empezar, los resultados de nuestro estudio tienen un alto grado de coincidencia con los de Adelman et al., en cuanto al signo de los coeficientes asociados a los fonemas (excepto por tres fonemas en españól: / t͡ʃ /, / ɾ / y /n/, aunque los coeficientes de los dos primeros son muy cercanos a cero en nuestro estudio). Puede que el orden relativo de estos coeficientes sea distinto en los dos estudios, pero resulta muy interesante el hecho de que el aporte a la predicción esté orientado mayormente en la misma dirección. En

todo caso, en los dos idiomas predominan los sonidos consonánticos y en ambas investigaciones los únicos fonemas comunes para los dos idiomas fueron el sonido de la /d/ y de la / t͡ʃ /, aunque los resultados para el segundo fonema son contradictorios, pero es llamativo que en los dos estudios es fonema más negativo en los dos estudios corresponda al de la /d/.

A este respecto, existen hallazgos opuestos. Por ejemplo, Aryani, Conrad, Schmidtke, & Jacobs (2018) encontraron que poemas escritos en alemán los sonidos sibilantes (como el que produce la s) y las consonantes plosivas (como la d) generan un *arousal* negativo, lo que coincide con los resultados en el dataset en inglés de SemEval-2018, aunque en español otra plosiva (la b) mostró un coeficiente positivo. De forma opuesta, Auracher, Albers, Zhai, Gareeva, & Stavniychuk (2011), a partir de poemas y hablantes nativos de diferentes idiomas (alemán, ruso, ucraniano, chino), encontraron que, independientemente del idioma, cuando hay una frecuencia relativamente alta de sonidos plosivos es más probable que un poema sea valorado como una expresión de sentimientos positivos, mientras que la mayor frecuencia relativa de sonidos nasales (n y m) se asoció a sentimientos negativos (como ocurrió en el dataset en inglés de SemEval-2018 en español, donde la n fue negativa). De otra parte, Aryani et al. reportan que las palabras con vocales cortas, como la asociada al fonema /ɪ/ generan un *arousal* y una evaluación de valencia más negativa, mientras que Rummer, Schweppe, Schlegelmilch, & Grice (2014), encontraron que bajo estados de ánimo positivos las personas tienden a emitir con mayor probabilidad el sonido de la /i:/, que sería un vocal larga (los dos resultados coinciden con lo observado en el dataset en inglés). En suma, parace que, además de las diferencias interidiomáticas, hay aspectos específicos de los fonemas (como su forma de articulación o su posición dentro de una palabra, por ejemplo) que pueden producir el efecto asociado al simbolismo de los sonidos o la iconicidad fonológica.

Para terminar esta discusión en torno a la especificidad idiomática, algunos autores controvierten la perspectiva de sistemas fonológicos universales, como la planteada en la Teoría de las Características Distintivas propuesta por Chomsky & Halle (1968), quienes señalan la supuesta existencia de unas manifestaciones innatas de estas características (por ejemplo, ciertos inventarios de sonidos). En contraste, Blevins (2017) sostiene que “un amplío estudio de los patrones de sonido de los idiomas del mundo permite ver que esos supuestos universales no son más que tendencias universales…, son propiedades emergentes de las gramáticas fonológicas; son aprendidas y específicas de cada idioma” (p. 55). Este razonamiento resulta coherente con lo expuesto por Aryani, Conrad, Schmidtke, & Jacobs (2018), quienes afirman que las pistas

fonéticas podrían no ser universales, teniendo en cuenta que cada idioma despliega distintas variaciones de las características fonéticas, dependiendo de su inventario fonético y de determinadas reglas fono tácticas. Esta diversidad fonológica se ve reflejada, además, en la cantidad de fonemas que se identificaron en la actual investigación: 27 para el español (de los cuales se seleccionaron 18) y 37 para el inglés (de los cuales se seleccionaron 20). De acuerdo con este razonamiento, el simbolismo de los sonidos y la iconicidad fonológica no necesariamente seguirían un patrón totalmente universal, como se deduciría de los resultados obtenidos con los corpus de tweets de SemEval-2018.

En relación con las implicaciones prácticas, nuestros resultados constituyen un aporte importante para tareas que implique la regresión de una variable continua (tarea de grano fino), en la medida en que los fonemas parecen contribuir a desambiguar los valores medios de valencia y, especialmente, a realizar predicciones más precisas en los niveles inferiores de la escala (es decir, las valencias asociadas a contenidos afectivos negativos). Además, si el efecto de los fonemas resulta redundante en relación con el contenido léxico del corpus, se podría ganar en tiempo computacional si se releva este contenido léxico por ciertas correspondencias fonológicas.

Para terminar, es importante señalar líneas futuras de investigación. Una sugerencia consiste en determinar si es posible potenciar la capacidad predictiva de los fonemas al entrenar modelos de mayor complejidad, incluyendo otros *features* que interactúen con el contenido acústico de los tweets (por ejemplo, *embeddings* de palabras) y utilizando otros algoritmos (como técnicas de *random forest* o ensambles de modelos) (Alhamdan, 2019). Igualmente, para que la contribución marginal identificada sea más alta resulta pertinente controlar ciertos atributos de los fonemas como su posición o su forma de articulación –como en el estudio de Adelman et al. (2018)– y ciertos atributos de las palabras como su categoría semántica o sus formas conjugadas –como en el estudio que hicieron Stadthagen-Gonzalez, Imbault, & Pérez Sánchez (2017) para establecer normas de valencia y *arousal* de un corpus de palabras en español. Por otra parte, es posible que el efecto de los fonemas esté asociado también con aspectos paralingüísticos como la prosodia o la entonación, los cuales han mostrado capacidad predictiva en tareas de análisis de sentimiento (Mairesse, Polifroni, & Di Fabbrizio, 2012), teniendo en cuenta que este efecto solo sería detectable a partir del lenguaje hablado, por lo que una forma de validar los presentes hallazgos consistiría en estudiar la capacidad predictiva de los fonemas a partir de muestras sonoras de lenguaje.

Referencias

Adelman, J., Estes, Z., & Cossu, M. (2018). Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*(175), 122-130. doi:10.1016/j.cognition.2018.02.007

AERA, APA, NCME. (2014). *Standards for educational and psychological testing.* Washington DC: AERA.

Akhtar, M., Ghosal, D., Ekbal, A., Bhattacharyya, P., & Kurohashi, S. (2018). A Multi-task Ensemble Framework for Emotion, Sentiment and Intensity Prediction. *arXiv:1808.01216v2 [cs.CL]* .

Alhamdan, I. (2019). Predicting the Emotional Intensity of Tweets. Tesis. Rochester Institute of Technology.

Alvarado, J. A., Caicedo, L. E., Carrillo, A. C., Forero, J. D., & Urueña, J. C. (2016). Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales (Trabajo de grado para optar por el título de Ingeniría Industrial). Bogotá: Pontificia Universidad Javeriana.

Aryani, A., Conrad, M., Schmidtke, D., & Jacobs, A. (2018). Why 'piss' is ruder than 'pee'? The role of sound in affective meaning making. *PLoS ONE, 13*(6), e0198430. doi:10.1371/journal.

Aryani, Arash; Jacobs, Arthur M.; Conrad, Markus. (2013). Extracting salient sublexical units from written texts: “Emophon,” a corpus-based approach to phonological iconicity. *Frontiers in Psychology, 4*(654), 1-15. doi:10.3389/fpsyg.2013.00654

Association for Computational Linguistics. (2019). *What is the ACL and what is Computational Linguistics?* Retrieved 12 01, 2019, from https://www.aclweb.org/portal/what-is-cl

Auracher, J., Albers, S., Zhai, Y., Gareeva, G., & Stavniychuk, T. (2011). P Is for Happiness, N Is for Sadness: Universals in Sound Iconicity to Detect Emotions in Poetry. *Discourse Processes, 48*, 1-25. doi:10.1080/01638531003674894

Austin, P., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*(68), 627-636. doi:10.1016/j.jclinepi.2014.12.014

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 774. doi:10.21105/joss.00774

Bird, S., Loper, E., & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

Blevins, J. (2017). What are grammars made of? En B. Samuels, *Beyond Markedness in Formal Phonology* (págs. 48-68). Amsterdam: John Benjamins Publishing Company. doi:10.1075/la.241.03ble

Bolshakov, I. A., & Gelbukh, A. (2004). *Computational linguistics. Models, Resources, Applications.* México: IPN – UNAM – Fondo de Cultura Económica.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English.* New York NY: Harper & R ow.

de Saussure, F. (1916/2005). *Curso de lingüística general.* (A. Alonso, Trad.) Buenos Aires: Losada.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences, 19*(10), 603-614. doi:10.1016/j.tics.2015.07.013

Duppada, V., Jain, R., & Hiray, S. (2018). SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets. *Proceedings of The 12th International Workshop on Semantic Evauation*, (pp. 18-23).

Fisher, I., Garnsey, M., & Hughes, M. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Int. Syst. in Accounting, Finance and Management*(23), 157-214. doi:10.1002/isaf.1386

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33*(1), 1-22. doi:10.18637/jss.v033.i01

Gil, J. M. (1999). *Introducción a las teorías lingüísticas del siglo XX*. Santiago de Chile: Red Internacional del Libro (RIL Editores).

Gil, S., Hattouti, J., & Laval, V. (2016). How children use emotional prosody: Crossmodal emotional integration? *Developmental Psychology, 52*(7), 1064-1072. doi:https://doi.org/10.1037/dev0000121

González Díaz, M. J., Koza, W. A., Méndez, B., Píppolo, C., Rivero, S. A., Rodrigo, A., . . . Tramallino, C. P. (2014). *Estudios de Lenguaje: Niveles de Representación Lingüística.* Iniciativa Latinoamericana de Libros Abiertos (LATin).

González Vidal, A. (2015). *Selección de variables: Una revisión de métodos existentes (Trabajo de fin de máster - Máster en Técnicas Estadísticas).* Universidade da Coruña, La Coruña, España.

Guerra de la Corte, A. (2016). *Técnicas de selección de variables en minería estadística de fatos (Trabajo fin de grado, Estadística e Investigación Operativa).* Universidad de Sevilla, Sevilla, España.

Henríquez, C., Pla, F., Hurtado, L.-F., & Guzmán, J. (2017). Análisis de sentimientos a nivel de aspecto usando ontologías y aprendizaje automático. *Procesamiento del Lenguaje Natural*(59), 49-56.

Jabreel, M., & Moreno, A. (2019). A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets. *Applied Science, 9*(1123). doi:10.3390/app9061123

Kraxenberger, M., & Menninghaus, W. (2016). Mimological Reveries? Disconfirming the Hypothesis of Phono-Emotional Iconicity in Poetry. *Frontiers in Psychology, 7*(1779), 1664-1078. doi:10.3389/fpsyg.2016.01779

Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., Vickers, B., . . . Jonides, J. (2019). Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook. *Emotion, 19*(1), 97-107. doi:10.1037/emo0000416

Kuijper, M., van Lenthe, M., & van Noord, R. (2018). UG18 at SemEval-2018 Task 1: Generating Additional Training Data for redicting Emotion Intensity in Spanish. *Proceedings of The 12th International Workshop on Semantic Evaluation.*, (pp. 279-285).

Lavin, M. (2019). Analyzing Documents with TF-IDF. *The Programming Historian, 8*. Obtenido de https://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf

Lopatovska, I., & Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Information Processing & Management, 47*(4), 575-592. doi:10.1016/j.ipm.2010.09.001

Mairesse, F., Polifroni, J., & Di Fabbrizio, G. (2012). Can Prosody Inform Sentiment Analysis? Experiments on Short Spoken Reviews. *The 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012).* doi:10.1109/ICASSP.2012.6289066

Majid, A. (2012). Current Emotion Research in the Language Sciences. *Emotion Review, 4*(4), 432-443. doi:10.1177/1754073912445827

Melkumova, L., & Shatskikh, S. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*(201), 746-755.

Mohammad, S., Bravo-Márquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval 2018 Task 1: Affect in Tweets. *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018)* (pp. 1-17). New Orleans, LA, USA: Association for Computational Linguistics.

Molina, M., Martínez, E., & Martín, M. T. (2015). CRiSOL: Base de Conocimiento de Opiniones para el Español. *Procesamiento de Lenguaje Natural*(55), 143-150.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of The Royal Society*(369), 20130299-20130299. doi:10.1098/rstb.2013.0299

Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for many languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (págs. 2710-2714). París, Francia: European Language Resources Association (ELRA).

Nygaard, L., & Queen, J. (2008). Communicating emotion: Linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance, 34*(4), 1017-1030. doi:https://doi.org/10.1037/0096-1523.34.4.1017

Paredes, M., Colomo, R., Salas, M., & Valencia, R. (2017). Sentiment analysis in spanish for improvement of products and services: A deep learning approach. *Sci. Program.*(1), 1-6. doi:10.1155/2017/1329281

Purao, S., Desouza, K., & Becker, J. (2012). Investigating Failures in Large-Scale Public Sector Projects with Sentiment Analysis. *e-Service Journal, 8*(2), 84-105. doi:10.2979/eservicej.8.2.84

Roche, J., Peters, B., & Dale, R. (2015). "Your tone says it all": The processing and interpretation of affective language. *Speech Communication*(66), 47-64. doi:https://doi.org/10.1016/j.specom.2014.07.004

Rummer, R., Schweppe, J., Schlegelmilch, R., & Grice, M. (2014). Mood Is Linked to Vowel Type: The Role of Articulatory Movements. *14*(2), 246-250. doi:10.1037/a0035752

Schmidtke, D. S., Conrad, M., & Jacobs, A. M. (2014). Phonological iconicity. *Frontiers in Psychology, 5*(80), 1-6. doi:10.3389/fpsyg.2014.00080

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing, 65*, 3-14. doi:10.1016/j.imavis.2017.08.003

Song, J., Kim, K., Lee, B., Kim, S., & Youn, H. (2017). A novel classification approach based on Naïve Bayes for Twitter sentiment analysis. *KSII Transactions on Internet and Information Systems, 11*(6). doi:10.3837/tiis.2017.06.011

Stadthagen-Gonzalez, H., Imbault, C., & Pérez Sánchez, M. (2017). Norms of valence and arousal for 14,031 spanish words. *Behavioral Research*(49), 111-123. doi:10.3758/s13428-015-0700-2

Stryker, S. (2004). Integrating emotion into identity theory. In J. Turner (Ed.), *Advances in group processes, Vol. 21. Theory and research on human emotions* (pp. 1-23). Elsevier Science/JAI Press. doi:10.1016/S0882-6145(04)21001-3

Thamm, R. (2004). Towards a universal power and status theory of emotion. In J. Turner (Ed.). Elsevier Science/JAI Press. doi:10.1016/S0882-6145(04)21008-6

Tordera Yllescas, J. C. (2011). *Lingüística computacional: Tecnologías del habla.* Valencia: Publicacions de la Universitat de València.

Tyagi, A. K., & Chandra, N. J. (2015). A Proposed Approach with Analysis of Speech Signals for Sentiment Detection. *2015 Fifth International Conference on Communication Systems and Network Technologies*, (págs. 339-344). doi:10.1109/CSNT.2015.97

Yadollahi, A., Shahraki, G., & Zaiane, O. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*(50), 1-33. doi:10.1145/3057270

Zheng, Z., Huang, X., & Zhang, Q. (2013). Emotional prosody modulates the recognition of affective word: An ERP study. *Acta Psychologica Sinica, 45*(4), 427-437. doi:10.3724/SP.J.1041.2013.00427

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society, Series B, 67*(Part 2), 301-320. doi:10.1111/j.1467-9868.2005.00503.x

Apéndices

Apéndice 1. Palabras identificadas mediante el algoritmo de regularización con un número de palabras determinadas.

Español		Inglés	
Palabra	**Beta**	**Palabra**	**Beta**
feliz	,1157	hilarious	,0958
felicidad	,0775	happy	,0851
genial	,0743	love	,0646
orgullosa	,0702	smile	,0424
sonrisa	,0623	rejoice	,0422
emocionada	,0608	thank	,0404
lindo	,0562	wonderful	,0384
afortunada	,0465	laughter	,0379
amo	,0417	morning	,0346
emocionado	,0399	amazing	,0282
jajajaja	,0389	exhilarating	,0210
risa	,0381	optimism	,0185
alegria	,0368	cheerful	,0174
jajajajajaja	,0368	joyous	,0164
gracias	,0366	big	,0151
jajaja	,0345	smiling	,0130
orgulloso	,0238	delight	,0129
jaja	,0215	joyful	,0123
amigos	,0201	glad	,0096
sonreir	,0193	beautiful	,0090
encantada	,0190	thanks	,0076
domingo	,0173	excited	,0075
radiante	,0146	fun	,0066
agradable	,0142	happiness	,0044
chistoso	,0118	funny	,0044
jajajajajajaja	,0116	lively	,0040
exito	,0114	coys	,0040
fortuna	,0111	lol	,0038
jugando	,0111	breezy	,0033
queda	,0107	hearty	,0026
sorpresa	,0102	anybody	,0022
afortunado	,0097	playful	,0016
viene	,0096	comedy	,0003
mega	,0088	fond	,0002
graciosa	,0087	maybeoneday	,00002
ahi	,0051	good	,00002
suerte	,0044	hopefullyfunny	,000003
siempre	,0038	irritation	-,0006
casita	,0034	supposed	-,0009

Español		Inglés	
Palabra	**Beta**	**Palabra**	**Beta**
contento	,0032	annoyed	-,0010
cumpleaños	,0028	furious	-,0015
jajajaj	,0009	nightmare	-,0036
semana	,0008	point	-,0051
videos	,0004	offended	-,0077
enojo	-,0005	swear	-,0097
pesadilla	-,0006	worst	-,0103
cansado	-,0020	tired	-,0150
detesto	-,0023	irritate	-,0150
mal	-,0031	fuming	-,0161
depresion	-,0035	terrible	-,0195
impotencia	-,0041	sadly	-,0248
asco	-,0049	outrage	-,0312
etc	-,0051	unhappy	-,0361
cansada	-,0056	horrible	-,0363
llorar	-,0069	anger	-,0432
ofendido	-,0077	sadness	-,05078
nadie	-,0095	awful	-,0590
zona	-,0111	sad	-,0647
decepcion	-,0125	depressing	-,0653
lamentable	-,0136	depression	-,07034
terrorismo	-,0139		
ninguna	-,0144		
mierda	-,0167		
romper	-,0189		
ironia	-,0221		
hirviendo	-,0230		
triste	-,0247		
resentida	-,0249		
lamentablemente	-,0261		
chester	-,0284		
odio	-,0286		
deprimida	-,0327		
rabia	-,0329		
insultar	-,0383		
horror	-,0392		
tristeza	-,0400		
puta	-,0423		
indignada	-,0441		
horrible	-,0553		

ANEXO 2

CARTA DE AUTORIZACIÓN DE LOS AUTORES
(Licencia de uso)

Bogotá, D.C., junio de 2020 _____

Señores
Biblioteca Alfonso Borrero Cabal S.J.
Pontificia Universidad Javeriana
Cuidad

Los suscritos:

Gabriel Alejandro Bernal Rojas	, con C.C. No	**79'802.825**
	, con C.C. No	
	, con C.C. No	

En mi (nuestra) calidad de autor (es) exclusivo (s) de la obra titulada:
Análisis de la contribución de los fonemas a la predicción de la valencia emocional en tweets en español e inglés
(por favor señale con una "x" las opciones que apliquen)
Tesis doctoral [ ] Trabajo de grado [X] Premio o distinción: Si [ ] No [X]
cual:
presentado y aprobado en el año 2020 , por medio del presente escrito autorizo (autorizamos) a la Pontificia Universidad Javeriana para que, en desarrollo de la presente licencia de uso parcial, pueda ejercer sobre mi (nuestra) obra las atribuciones que se indican a continuación, teniendo en cuenta que en cualquier caso, la finalidad perseguida será facilitar, difundir y promover el aprendizaje, la enseñanza y la investigación.

En consecuencia, las atribuciones de usos temporales y parciales que por virtud de la presente licencia se autorizan a la Pontificia Universidad Javeriana, a los usuarios de la Biblioteca Alfonso Borrero Cabal S.J., así como a los usuarios de las redes, bases de datos y demás sitios web con los que la Universidad tenga perfeccionado un convenio, son:

AUTORIZO (AUTORIZAMOS)	SI	NO
1. La conservación de los ejemplares necesarios en la sala de tesis y trabajos de grado de la Biblioteca.	X	
2. La consulta física (sólo en las instalaciones de la Biblioteca)	X	
3. La consulta electrónica – on line (a través del catálogo Biblos y el Repositorio Institucional)	X	
4. La reproducción por cualquier formato conocido o por conocer	X	
5. La comunicación pública por cualquier procedimiento o medio físico o electrónico, así como su puesta a disposición en Internet	X	
6. La inclusión en bases de datos y en sitios web sean éstos onerosos o gratuitos, existiendo con ellos previo convenio perfeccionado con la Pontificia Universidad Javeriana para efectos de satisfacer los fines previstos. En este evento, tales sitios y sus usuarios tendrán las mismas facultades que las aquí concedidas con las mismas limitaciones y condiciones	X	

De acuerdo con la naturaleza del uso concedido, la presente licencia parcial se otorga a título gratuito por el máximo tiempo legal colombiano, con el propósito de que en dicho lapso mi (nuestra) obra sea explotada en las condiciones aquí estipuladas y para los fines indicados, respetando siempre la titularidad de los derechos patrimoniales y morales correspondientes, de

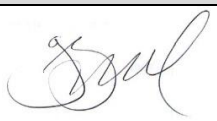
acuerdo con los usos honrados, de manera proporcional y justificada a la finalidad perseguida, sin ánimo de lucro ni de comercialización.

De manera complementaria, garantizo (garantizamos) en mi (nuestra) calidad de estudiante (s) y por ende autor (es) exclusivo (s), que la Tesis o Trabajo de Grado en cuestión, es producto de mi (nuestra) plena autoría, de mi (nuestro) esfuerzo personal intelectual, como consecuencia de mi (nuestra) creación original particular y, por tanto, soy (somos) el (los) único (s) titular (es) de la misma. Además, aseguro (aseguramos) que no contiene citas, ni transcripciones de otras obras protegidas, por fuera de los límites autorizados por la ley, según los usos honrados, y en proporción a los fines previstos; ni tampoco contempla declaraciones difamatorias contra terceros; respetando el derecho a la imagen, intimidad, buen nombre y demás derechos constitucionales. Adicionalmente, manifiesto (manifestamos) que no se incluyeron expresiones contrarias al orden público ni a las buenas costumbres. En consecuencia, la responsabilidad directa en la elaboración, presentación, investigación y, en general, contenidos de la Tesis o Trabajo de Grado es de mí (nuestro) competencia exclusiva, eximiendo de toda responsabilidad a la Pontifica Universidad Javeriana por tales aspectos.

Sin perjuicio de los usos y atribuciones otorgadas en virtud de este documento, continuaré (continuaremos) conservando los correspondientes derechos patrimoniales sin modificación o restricción alguna, puesto que de acuerdo con la legislación colombiana aplicable, el presente es un acuerdo jurídico que en ningún caso conlleva la enajenación de los derechos patrimoniales derivados del régimen del Derecho de Autor.

De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, "Los derechos morales sobre el trabajo son propiedad de los autores", los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. En consecuencia, la Pontificia Universidad Javeriana está en la obligación de RESPETARLOS Y HACERLOS RESPETAR, para lo cual tomará las medidas correspondientes para garantizar su observancia.

NOTA: <u>Información Confidencial:</u>
Esta Tesis o Trabajo de Grado contiene información privilegiada, estratégica, secreta, confidencial y demás similar, o hace parte de una investigación que se adelanta y cuyos resultados finales no se han publicado. Si [ ] No [X]
En caso afirmativo expresamente indicaré (indicaremos), en carta adjunta, tal situación con el fin de que se mantenga la restricción de acceso.

NOMBRE COMPLETO	No. del documento de identidad	FIRMA
Gabriel Alejandro Bernal Rojas	79'802.825	

FACULTAD: Ingeniería
PROGRAMA ACADÉMICO: Maestría en Analítica para la Inteligencia de Negocios

ANEXO 3
BIBLIOTECA ALFONSO BORRERO CABAL, S.J.
DESCRIPCIÓN DE LA TESIS O DEL TRABAJO DE GRADO
FORMULARIO

TÍTULO COMPLETO DE LA TESIS DOCTORAL O TRABAJO DE GRADO	
Análisis de la contribución de los fonemas a la predicción de la valencia emocional en tweets en español e inglés	
SUBTÍTULO, SI LO TIENE	
AUTOR O AUTORES	
Apellidos Completos	**Nombres Completos**
Bernal Rojas	Gabriel Alejandro
DIRECTOR (ES) TESIS O DEL TRABAJO DE GRADO	
Apellidos Completos	**Nombres Completos**
Alvarado Valencia	Jorge Andrés

FACULTAD			
Ingeniería			
PROGRAMA ACADÉMICO			
Tipo de programa (seleccione con "x")			
Pregrado	Especialización	Maestría	Doctorado
		X	
Nombre del programa académico			
Analítica para la Inteligencia de Negocios			
Nombres y apellidos del director del programa académico			
Jorge Andrés Alvarado Valencia			
TRABAJO PARA OPTAR AL TÍTULO DE:			
Magister			
PREMIO O DISTINCIÓN *(En caso de ser LAUREADAS o tener una mención especial)*:			

CIUDAD	**AÑO DE PRESENTACIÓN DE LA TESIS O DEL TRABAJO DE GRADO**	**NÚMERO DE PÁGINAS**
Bogotá	**2020**	**58**

TIPO DE ILUSTRACIONES (seleccione con "x")						
Dibujos	Pinturas	Tablas, gráficos y diagramas	Planos	Mapas	Fotografías	Partituras
		X				

SOFTWARE REQUERIDO O ESPECIALIZADO PARA LA LECTURA DEL DOCUMENTO
Nota: En caso de que el software (programa especializado requerido) no se encuentre licenciado por la Universidad a través de la Biblioteca (previa consulta al estudiante), el texto de la Tesis o Trabajo de Grado quedará solamente en formato PDF.

MATERIAL ACOMPAÑANTE					
TIPO	**DURACIÓN (minutos)**	**CANTIDAD**	**FORMATO**		
			CD	DVD	Otro ¿Cuál?
Vídeo					
Audio					
Multimedia					
Producción electrónica					
Otro Cuál?					

DESCRIPTORES O PALABRAS CLAVE EN ESPAÑOL E INGLÉS

Son los términos que definen los temas que identifican el contenido. (*En caso de duda para designar estos descriptores, se recomienda consultar con la Sección de Desarrollo de Colecciones de la Biblioteca* Alfonso Borrero Cabal S.J *en el correo biblioteca@javeriana.edu.co, donde se les orientará).*

ESPAÑOL	INGLÉS
Arbitrariedad del signo lingüístico	Arbitrariness of linguistic sign
Simbolismo de los sonidos	Sound symbolism
Fonemas	Phonemes
Procesamiento del lenguaje natural	Natural Language Processing
Análisis de sentimiento	Sentiment Analysis

RESUMEN DEL CONTENIDO EN ESPAÑOL E INGLÉS

(Máximo 250 palabras - 1530 caracteres)

Aunque tradicionalmente se ha asumido que el sonido de las palabras y su significado se relacionan de forma arbitraria, distintos hallazgos empíricos respaldan la hipótesis de que las unidades fonológicas básicas del lenguaje guardan una relación sistemática con aspectos semánticos, incluyendo la connotación afectiva y actitudinal de las palabras (Adelman, Estes, & Cossu, 2018; Aryani, Conrad, Schmidtke, & Jacobs, 2018; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Monaghan, Shillcock, Christiansen, & Kirby, 2014; Schmidtke, Conrad, & Jacobs, 2014). A partir de estas premisas, se buscó identificar si las unidades fonológicas del español y el inglés contribuyen a la predicción de la valencia emocional en un corpus de tweets. Para esto, se entrenó un conjunto de modelos de regresión lineal múltiple, cuyo desempeño fue evaluado a partir de la correlación y los indicadores de error calculados partir de las valencias predichas y las observadas en los *datasets* de prueba proporcionados por el concurso SemEval-2018 (Mohammad, Bravo-Márquez, Salameh, & Kiritchenko, 2018). Se encontró que la adición de los recursos fonológicos a un conjunto de predictores léxicos (*Bag of Words* de los Tweets, normalizada con el método TF-IDF) tiene un efecto reducido pero consistente sobre las métricas globales de ajuste, y en ambos idiomas permite discriminar con mayor precisión las valencias observadas cercanas a los valores medios, así como las valencias inferiores asociadas a contenidos afectivos negativos.