Propuesta recibida en noviembre 2020 y aceptada para publicación en abril 2022.

La #felicidad en Twitter: ¿qué representa realmente?

#happiness in Twitter: What does it really represent?

Gemma Bel-Enguix
Instituto de Ingeniería
Universidad Nacional Autónoma de México

Helena Gómez-Adorno
Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas
Universidad Nacional Autónoma de México

Karla Mendoza Grageda
Facultad de Ciencias
Universidad Nacional Autónoma de México

Grigori Sidorov
Centro de Investigación en Computación
Instituto Politécnico Nacional

Juan Vásquez
Posgrado en Ciencia e Ingeniería de la Computación
Universidad Nacional Autónoma de México

Resumen

Existe un gran número de trabajos que tienen por objeto la clasificación de diversos tipos de documentos, desde textos literarios hasta interacciones informales en redes sociales como Twitter, de acuerdo a los sentimientos que pretenden evocar. Se pueden realizar clasificaciones muy variadas con base en los sentimientos que el autor considere. El objetivo de este artículo es clasificar una recopilación de tuits en diferentes contextos en los que la palabra 'feliz' o 'felicidad' se pueden emplear; por ejemplo publicidad, felicitaciones o como un simple sarcasmo. Para esto se hará uso de sistemas de aprendizaje supervisado y se emplearán varios métodos de procesamiento de lenguaje natural como tokenización, identificación de palabras funcionales y n-gramas.

Palabras clave

Feliz, Felicidad, Tuit, Aprendizaje Automático

Abstract

There are a number of works that deal with the classification of feelings evoked by various types of documents, from literaty texts to social networks like Twitter. Various classifications can be performed based on the sentiments considered by the author. The goal of this article is to classify a collection of tweets in different contexts in which the word 'happy' or 'happiness' can be used, for example advertising, congratulations or as a simple sarcasm. This will be done using supervised learning systems. Several natural language processing methods such as tokenization, functional word identification, and n-grams will be employed.

Keywords

Happy, Happiness, Tweet, Machine Learning

1. Introducción

Las redes sociales se han convertido en el medio de comunicación por excelencia en el siglo XXI. Al contrario de los denominados "mass media", que proliferaron y consiguieron un gran poder durante el siglo XX, las redes sociales tienen un carácter bidireccional, o incluso multidireccional. Esto significa que cada usuario es a la vez hablante y oyente, generador y receptor del contenido comunicativo. El hecho de que todos sean comunicadores contrasta con una de las principales características de los medios tradicionales: la exclusividad en la información y la difusión de la noticia. En este nuevo escenario, las redes sociales están más orientadas a la comunicación de actividades personales, sentimientos y opiniones, eso sí, en un canal que los difunde como si fueran noticias. En este contexto, la expresión de la emoción ha tomado un lugar preponderante en las plataformas de comunicación social.

Por otra parte, la generación de grandes cantidades de datos textuales y la aparición de técnicas de análisis computarizadas ha desembocado en un gran desarrollo de la minería de opinión y análisis de sentimientos (Turney, 2002). Los pri-

DOI: 10.21814/lm.14.1.342

meros trabajos en el área tenían como objetivo la clasificación de textos bajo las categorías de polaridad positivo/negativo. En diferentes contextos, esta polaridad se puede interpretar como buenas noticias vs. malas noticias (Koppel & Shtrimberg, 2004), me gusta vs. no me gusta (Kim & Hovy, 2006), estoy de acuerdo vs. no estoy de acuerdo (Bansal et al., 2008; Wojatzki et al., 2018), apoyo político vs. no apoyo (Thomas et al., 2006), o probablemente gane vs. probablemente no gane (Kim & Hovy, 2007) en el ámbito electoral.

Al principio, los trabajos se centraban en el análisis de reseñas y recomendaciones sobre diversos productos, desde películas a hoteles. Pero la gran penetración de las redes sociales en la vida cotidiana, ha acabado poniendo estas formas de comunicación en el centro de interés del *big data* por lo que respecta al lenguaje natural. Uno de los primeros trabajos que traslada el tema de clasificación de polaridad a análisis de grandes cantidades de datos extraídos de redes sociales es el de Go et al. (2009).

A veces, los escritos, sobre todo los originados en internet, no están orientados exclusivamente a dar la opinión sobre un tema en concreto, o a expresarla de una forma binaria. Por esto muchos autores optaron por identificar elementos de subjetividad y emociones en los textos (Banea et al., 2011; Morency et al., 2011; Mohammad & Kiritchenko, 2018). Durante la última década, muchos trabajos han seguido esta línea centrándose en el estudio de las redes sociales (Mohammad, 2012). El interés suscitado ha llevado a organizar competiciones en congresos internacionales (Mohammad et al., 2018; Naderi et al., 2018).

Aunque se han propuesto diversas formas de categorizar las emociones (Plutchik, 1980), existe un consenso general en considerar seis emociones básicas, siguiendo la clasificación de Ekman (1992; 1994). Estas son tristeza, miedo, enfado, disgusto, sorpresa y felicidad.

En general, dentro del marco descrito por Ekman, las emociones negativas han recibido más atención por parte de la comunidad científica. En cambio, este artículo va enfocado al estudio de la felicidad en Twitter. Las bases para esta investigación se encuentran en (Sidorov et al., 2016), donde se explica la metodología de la recopilación de tuits, basada en hashtags. Los mismos autores aplican estos criterios al filtrado y análisis de micro-mensajes con emociones negativas (Camacho Vázquez et al., 2018). En este trabajo se parte de un corpus que selecciona tuits con hashtags relacionados con la felicidad, cuyo léxico se incluye también dentro de este campo semántico. Sin embargo, nos preguntamos si el uso de una terminología relacionada con la felicidad implica que los tuits analizados transmiten efectivamente esta emoción o bien tienen otras connotaciones diferentes. Para verificar si efectivamente tienen otros significados, se propone realizar una detección manual y posteriormente implementar un sistema de aprendizaje que sea capaz de distinguir la verdadera emoción más allá del léxico y hashtags utilizados.

Para llevar a cabo esta investigación, el artículo está organizado de la siguiente forma. La sección 2 revisa el trabajo que se ha hecho hasta la fecha en esta misma línea de estudio. En la sección 3 se explica el proceso de compilación y etiquetado del corpus. Además, se detalla la metodología para la clasificación de tuits y se da cuenta de un primer acercamiento al léxico. La sección 4 reseña los principales de pasos de preprocesamiento textual que se han llevado a cabo. La sección 5 muestra los resultados obtenidos para cada uno de los experimentos. El artículo se cierra con las conclusiones (sección 6), donde se anotan también algunas líneas de trabajo futuro.

2. Trabajo Relacionado

El presente artículo aborda el estudio de la 'felicidad' en un corpus de tuits en español, proponiendo un sistema de aprendizaje automático que ayude a distinguir el verdadero significado de los micro-textos. Los artículos que sirven como referencia, o bien trabajan con la detección y análisis de las emociones, principalmente positivas, encontradas en diferentes corpus, o bien aplican técnicas de aprendizaje automático sobre colecciones textuales en español.

Aunque la mayor parte de la investigación en el área se realiza en inglés, existen interesantes contribuciones en español. Algunos artículos proponen algoritmos para la clasificación de polaridad (positivo-negativo-neutro) en diversos documentos. Por ejemplo, Vilares et al. (2013) clasifican textos subjetivos como positivos o negativos, basándose en diccionarios semánticos y en la estructura semántica de las oraciones.

Por otro lado, Gruzd et al. (2011) realizan un análisis sobre un conjunto de tuits recopilados desde el primer día de los juegos olímpicos hasta unos días antes del final donde, después de clasificarlos en positivo, negativo, neutro y ambos, llegan a la conclusión de que aquellos señalados como positivos son, en cantidad, tres veces más que aquellos señalados como negativos.

El trabajo de Mogilner et al. (2011), aunque en inglés, sigue una línea de interés muy cercana a la nuestra. Los autores recopilan blogs de cuyos

autores conocen la edad, derivados de las búsquedas de las frases 'yo siento' o 'me siento'. De esta recopilación filtran los que completan la frase de búsqueda con la palabra 'felicidad'. Después realizan un análisis sobre las palabras coocurrentes de esta frase distinguiendo dos categorías distintas *'excited happiness' y 'peaceful happiness'* de las cuales concluyen que los autores más jóvenes expresan una *excited happiness* mientras que los adultos están más cercanos a una *peaceful happiness*.

Kumar et al. (2015) trabajan con una recopilación de tuits y consideran cinco emociones derivadas de la clasificación de Ekman, de donde eliminan la sorpresa: felicidad, tristeza, disgusto, miedo e ira. Los autores evalúan adjetivos asignándoles un valor en cada uno de los sentimientos. Además consideran los adverbios y algunos verbos que modifican el adjetivo. De esta forma se calcula el valor del sentimiento para cada uno de los tuits recopilados.

Los ejemplos anteriores utilizan métodos usuales de sistemas de aprendizaje. Recientemente se ha procurado resolver problemas de clasificación con métodos basados en redes neuronales. Bakhtiyar et al. (2019) enfrentan el problema de distinguir entre 'felicidad del autor' y 'felicidad social'. Para ello proponen una transferencia inductiva semi-supervisada de aprendizaje, en la cual los resultados obtenidos en la tarea actual dependen de la transferencia de conocimientos obtenidos en tareas anteriores. Su propuesta consta de tres pasos. El primero tiene como propósito pre-entrenar el modelo del lenguaje base del modelo AWD-LSTM (*Average-SGD (Stochastic Gradient Descent) Weight-Dropped Long Short Term Memory*). En el segundo paso se ajusta el lenguaje utilizando datos específicos de la tarea; además, en lugar de mantener la misma tasa de aprendizaje para las capas del modelo AWD-LSTM, esta se va modificando en cada ajuste de capa. En el último paso se adapta un *gradual unfreezing heuristic* que se encarga de que no todas las capas sean ajustadas al mismo tiempo y, en lugar de eso, empieza 'descongelando' la última de ellas para ajustarla y seguir con las que la preceden. Al final se obtiene una exactitud de 93 % para la detección de 'felicidad social' y una exactitud de 87 % para la detección de 'felicidad del autor'.

Con todo este marco de referencia, el presente trabajo está estrechamente relacionado con un primer análisis realizado por Camacho Vázquez et al. (2018) sobre tuits con carga negativa. Dicho artículo explica cómo se implementan pruebas sobre dos categorías, tuits detectados como negativos y otros catalogados como neutrales. Los resultados de estos experimentos han demostrado que el uso de del sistema de aprendizaje de distribución multinomial (MNB) con frecuecia de término — frecuencia inversa de documento (TF-IDF) obtiene los mejores resultados con una puntuación $F_1 = 0{,}962$ tomando en cuenta unibigramas de palabras y $F_1 = 0{,}960$ tomando unigramas de palabras. Tras llevar a cabo diversas pruebas con cuatro categorías diferentes de tuits negativos y una categoría de tuits neutrales se muestra cómo las mayores puntuaciones se obtuvieron al usar el sistema de aprendizaje MNB con valores de frecuencias de términos dando una puntuación de $F_1 = 0{,}664$ usando unibigramas y una puntuación de $F_1 = 0{,}663$ usando unibitrigramas. Estos resultados son mejores comparados con las mismas pruebas tomando las palabras lematizadas.

De dichos experimentos se concluye que la combinación de diversas características, como se puede ver con los unibigramas y los unibitrigramas con frecuencias de términos, mejora los resultados obtenidos en las pruebas con sistemas de aprendizaje. Además se prueba que si se limitan las características más frecuentes combinadas por categorías a 1000 elementos también se obtiene una mejora en los resultados.

Por otra parte modelos basados en BERT también se han implementado para la tarea de clasificación de tuits en español. Por ejemplo, González et al. (2021) han propuesto un modelo basado en BERT y pre-entrenado con tuits en español. El objetivo de su modelo es mejorar los resultados del estado del arte en cuanto a tareas de clasificación de tuits en español. Asimismo Zeng et al. (2021) han propuesto un sistema de aprendizaje profundo llamado Senti-BSAS, el cual, por medio de un mecanismo de atención junto con un cálculo de sentimientos basado en un análisis léxico, clasifica la felicidad de una oración en dos clases: una clase en la cual el autor es el motivo de dicha felicidad, y otra en la cual la felicidad es generada por agentes externos al mismo.

Chiorrini et al. (2021) aplicaron los modelos *Uncased BERT* y *Cased BERT* para el análisis de emociones de tuits. Este trabajo utiliza al *Tweet Emotion Intensity Dataset*, un corpus de tuits en inglés creado por Mohammad & Bravo-Marquez (2017). La *accuracy* de *BERT uncased* resultó en 0.89, mientras que la F_1 fue de 0.89. Respecto al modelo *BERT cased*, obtuvieron una *accuracy* de 0.90 y una F_1 de 0.91.

Respecto al tratamiento de tuits en español, Rosá & Chiruzzo (2021) clasificaron tuits en ocho y seis clases. Este trabajo utiliza una red LSTM

alimentada con características generadas por BETO; específicamente, la red toma el token CLF y el centroide de la representación de cada token generado por BETO. Después de obtener las características, clasificaron los tuits respecto a los sentimientos *anger, disgust, fear, joy, sadness, surprise, others*. Después de obtener las características, realizaron dos distintas clasificaciones de sus tuits. La primera clasificación fue respecto a las clases *anger, disgust, fear, joy, sadness, surprise, others*, mientras que la segunda fue considerando *anger, disgust, fear, joy, sadness, surprise*. La clase *others* fue considerada para aquellos tuits que recibieron diferentes emociones de acuerdo a diferentes anotadores, o a tuits neutrales. Los resultados del modelo de clasificación entrenado con el corpus que contiene la clase *others* fueron: precisión de 0.6860 y $F_1 = 0{,}6620$, mientras que para el modelo entrenado con el corpus sin la clase *others* se obtuvo una precisión de 0.7447 y una $F_1 = 0{,}7170$.

Estos recientes trabajos de clasificación de tuits utilizando arquitecturas basadas en BERT demuestran el potencial de dichas arquitecturas para obtener mejores resultados que las técnicas clásicas para clasificación.

Aunque muy cercanos conceptualmente a nuestra investigación, estos trabajos no pueden compararse directamente con el nuestro. Por una parte, la mayor parte de trabajos tratan de la subcategorización de la felicidad, pero lo hacen desde presupuestos binarios. Otros, como Rosá & Chiruzzo (2021) proponen una clasificación multiclase, pero consideran únicamente las emociones primarias, no las distinciones dentro de cada una de ellas, lo que dificulta una comparación directa.

3. Compilación y etiquetado del corpus

La técnica de recopilación del corpus usada en este trabajo se basa en (Sidorov et al., 2016), donde se analizan formas de resolver algunos de los problemas que se presentan dentro de la detección de sentimientos en microblogs.

En el referido trabajo se recopilaron tuits que contienen un hashtag relacionado con una de las emociones: alegría, ira, tristeza, asco, miedo o sorpresa. Es importante mencionar que se descartaron todos aquellos tuits que contuvieran hashtags de más de una emoción, como el siguiente:

```
Última noche en casa de mis padres! #feliz
pero también última noche que duermo
con mis niños #triste.
```

En el trabajo mencionado se hace un análisis de los hashtags que se usan en los tuits relacionados con la emoción 'alegría'. Son los siguientes: #felicidad, #feliz, #alegria, #felicidades #alegre y #contento(a), de los cuales #feliz y #felicidad resultaron ser los más comunes.

Para nuestro trabajo, se han recopilado 10048 tuits del 26 de agosto al 2 de diciembre de 2016. Todos tienen el hashtag #felicidad o #feliz, además de cualquier otro de los que han sido identificados dentro de la categoría que define esta emoción. Este ha sido el único criterio de recopilación de corpus. Es decir, todos los tuits emitidos durante este período con este hashtag han sido recogidos. No existe ningún otro filtro. El corpus recopilado está disponible[1].

Por otra parte, Sidorov et al. (2016) realizaron un estudio del uso de los hashtags relacionados con la felicidad. Se llegó a la conclusión de que la alegría es la emoción que se expresa más frecuentemente dentro de Twitter. Sin embargo, se encuentra en múltiples contextos que, curiosamente, no siempre denotan alegría o felicidad. Mediante un análisis semántico hemos clasificado los tuits con los hashtags #felicidad, #feliz y #alegría en cinco categorías:

1. Alegría (A). Tuits que reflejan realmente alegría o felicidad: `que #feliz me haces, me sacas una #sonrisa sin hacer nada.... #esperanza!`
2. Publicidad (P). Hace referencia a comerciales que ofrecen felicidad si compras un producto determinado: `para ser #feliz primero se piensa en #seguridad y al pensar en seguridad se piensa en ocq security group`
3. Felicitación (F). Felicitaciones de cumpleaños, otros eventos personales o fechas señaladas: `eres ejemplo vivo de gente luchadora que ama a su país y #feliz cumpleaños te mando un beso gigante desde méxi https://t.co/o1rlvxygud`
4. Consejo (C). Se trata, en general, de mensajes de autoayuda o reflexiones (pseudo-)filosóficas: `descubre como vivir en el presente para poder ser #feliz y tener # https://t.co/uzkuigvtht salud sin depender de na https://t.co/tuguwxfomm`
5. Sarcasmo o no_alegría (N). Tuits con doble sentido que no denotan alegría en realidad: `literalmente... estás viendo la #felicidad "lo que ves es una proteína de miosina arrastrando una endorfina a lo... https://t.co/kjtcjq6vzn`

[1] https://github.com/GIL-UNAM/TwitterHappiness

3.1. Etiquetado

Los 10048 tuits que conforman el corpus final fueron proporcionados a tres personas para que etiquetaran, según su criterio, cada uno de los tuits en las categorías de 'alegría' (A), 'publicidad' (P), 'felicitación' (F), 'consejo' (C) y 'sarcasmo o no_alegría' (N).

En un 86.2 % de los casos, al menos dos etiquetadores estuvieron de acuerdo en considerar el tuit dentro del mismo grupo. Aquellos mensajes en los que ninguno de los etiquetadores coincidieron se clasificaron en 'No agreement' (NA). Representaron un 13.81 % del conjunto del corpus.

En la Figura 1 se observa el porcentaje de tuits que fueron englobados en cada categoría. Las clases más numerosas son las de 'alegría', 'consejos' y 'no agreement'. El alto porcentaje de NA indica hasta qué punto este tipo de clasificación está influenciada por parámetros subjetivos.

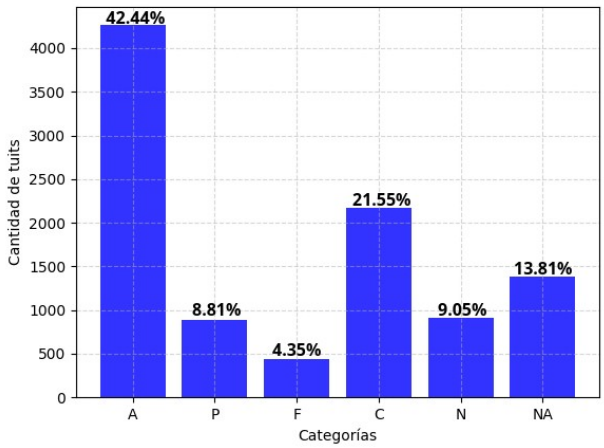

Figura 1: Gráfica de barras de la cantidad de tuits por categoría.

3.2. Análisis léxico

Una vez recopilado y etiquetado el corpus se ha realizado un análisis léxico comparando los resultados en cada una de las categorías. En primer lugar, después de extraer las palabras funcionales, se han seleccionado los 15 términos de contenido más frecuentes. Este número se ha establecido con el propósito de tener una muestra amplia de palabras pero suficientemente reducida para que pueda ser ilustrativa a la hora de su análisis. A continuación, se ha estudiado la frecuencia relativa de cada una de estas palabras en las diferentes categorías.

La Tabla 1 muestra los resultados obtenidos. A pesar de que *si* y *mas* se podrían considerar dentro de las palabras funcionales, en este caso se decidió conservarlas, pues podrían tener una estrecha relación con los sentimientos positivos. Hay que decir que la especial ortografía de la lengua tecleada no permite distinguir entre *si* y *sí*. Se puede observar que las palabras *feliz* y *felicidad* son las más frecuentes en todas las categorías y además están distribuidas equitativamente, con excepción de la palabra *felicidad* en la categoría F.

Palabra	**Categoría**					
	A	P	F	C	N	Total
feliz	7.71	6.42	10.71	6.51	6.99	7.33
felicidad	3.68	5.04	0.95	5.73	2.86	4.12
vida	0.80	0.58	0.76	2.00	0.36	1.01
hoy	1.01	0.75	1.43	0.43	0.61	0.82
día	0.72	0.62	1.80	0.66	0.60	0.77
si	0.40	0.54	0.32	1.26	0.69	0.66
mejor	0.59	0.49	0.55	0.62	0.31	0.55
gracias	0.88	0.25	0.76	0.13	0.14	0.53
amor	0.49	0.20	0.69	0.68	0.19	0.48
semana	0.42	0.36	0.21	0.27	0.34	0.45
siempre	0.39	0.12	0.53	0.72	0.19	0.44
mas	0.64	0.15	0.58	0.27	0.34	0.42
dios	0.40	0.00	1.02	0.43	0.12	0.40
cada	0.32	0.23	0.25	0.64	0.23	0.39
hace	0.38	0.15	0.21	0.35	0.28	0.32

Tabla 1: Frecuencia relativa de las primeras 15 palabras de mayor a menor aparición dentro de todo el corpus.

4. Procesamiento del corpus

Con el fin de poder trabajar con los textos de los tuits, estos se han sometido a un preprocesamiento que ha incluido: a) eliminación de los hiperenlaces, b) tokenizando, c) eliminación de los signos de puntuación y las palabras funcionales, d) extracción de raíces. Se han llevado a cabo experimentos combinando la existencia o no de los dos últimos procesos.

En este sentido, hay que mencionar que se intentaron otros métodos de preprocesamiento. Por ejemplo, se usó el lematizado en lugar de la extracción de raíces. Pero los resultados obtenidos fueron peores. Por ello, se optó por trabajar con esta última técnica.

En la Tabla 2 podemos ver una comparación entre un tuit antes y después de aplicar los pasos a, b y c del pre-procesamiento. Se observa que los hashtags se conservan, pues se espera que mejoren la posterior categorización, y se retira solamente el signo # de la palabra para estandarizar el vocabulario.

Tuit original	`#feliz #cumple a esa amiga hermosa que tengo la suerte de tener hace ya unos años. feliz de`
Tuit pre-procesado	`feliz cumple amiga hermosa suerte tener hace años feliz`

Tabla 2: Ejemplo de un tuit antes y después de pre-procesarse (pasos a, b, c).

Después del pre-procesamiento de los tuits podemos recabar la cantidad de palabras diferentes detectadas dentro de todo el corpus y la cantidad de palabras diferentes que resultan después de descartar las palabras funcionales, en la Tabla 3 se muestran dichas cantidades.

Cantidad de palabras distintas	19,357
Cantidad de palabras distintas descartando palabras funcionales	19,128

Tabla 3: Cantidad de palabras distintas dentro del corpus.

5. Clasificación automática de tipos de felicidad

Una vez compilado, etiquetado y pre-procesado el corpus, entrenamos algoritmos de aprendizaje para obtener modelos de clasificación de tipos de tuits con etiqueta de #felicidad. Los algoritmos utilizados son *Naive Bayes* (NB), *Logistic Regression* (LR), *Random Forest* (RF) y *Support Vector Machines* (SVM).

Para la representación vectorial de los tuits utilizamos n-gramas de palabras (con n de 1 a 6). Los n-gramas se forman después de realizar las operaciones de pre-procesamiento, es decir, cuando en el pre-procesamiento se eliminan las palabras funcionales éstas ya no forman parte de los n-gramas. Como podemos observar en las tablas 4 y 5, los n-gramas con $n = 1$ no parecen ser muy representativos ya que obtienen los resultados mas bajos cuando se realiza la evaluación del modelo de clasificación. Seguramente esto es a causa de que los vocabularios de cada categoría son bastante similares. Por ello se optó por incluir también secuencias más largas.

Evaluamos el impacto de las técnicas de pre-procesamiento descritas anteriormente en el rendimiento de los modelos de clasificación. Según lo que se aprecia en la Tabla 4, los mejores resultados se obtienen conservando las palabras funcionales. Esto resulta contraintuitivo, ya que estas palabras no aportan información semántica y, en principio, su eliminación debería aumentar el rendimiento de los modelos. Por el contrario, se espera que la extracción de raíces de las palabras disminuya el rendimiento de los modelos ya que, por ejemplo, 'felicidad' y 'felicidades' tienen la misma raíz. Si suponemos que 'felicidades' puede ser una palabra única dentro de la categoría 'F', al extraer la raíz dejará de serlo. En cambio, se obtienen mejores resultados cuando se realiza la extracción de raíces.

Por último, realizamos experimentos con dos esquemas de pesado frecuentemente utilizados en la literatura existente en el área: la frecuencia del término (TF) y la frecuencia del término por la frecuencia inversa de documento (TF-IDF). En este experimento se espera que, como la medida TF-IDF considera la especificidad de los términos dentro del corpus, mejore los resultados de la clasificación de los tuits.

Para evaluar la generalización de los modelos de clasificación utilizamos un método de validación cruzada con el fin de asegurar que los resultados fueran independientes de la partición de corpus en los conjuntos de prueba y entrenamiento. En particular, utilizamos un método de validación cruzada estratificada por capas que permite preservar la proporción de muestras para cada clase pues, como se puede observar en las cantidades de tuits por categoría y en la gráfica de barras, no se tiene instancias balanceados por clase.

Se realizaron experimentos con validación cruzada para 3, 5 y 10 capas y se obtuvo un promedio de exactitud para cada uno de los algoritmos de aprendizaje implementados. Después se calculó un promedio de rendimiento de todos los clasificadores por tipo de pre-procesamiento y esquemas de pesado. Una vez teniendo estos promedios se procedió a identificar el tipo de pre-procesamiento y esquema de pesado más adecuado para cada conjunto de características. Los resultados del proceso descrito se muestran en la Tabla 4 donde se reporta el promedio del rendimiento de los 4 clasificadores por conjunto de características. Como resultado podemos observar que para cada uno de los conjuntos de características, los mejores porcentajes se obtienen cuando se consideran las palabras funcionales, se aplica el proceso de extracción de raíces y se utiliza el esquema de pesado TF-IDF. Nótese que este resultado es contrario a nuestra hipótesis de trabajo, que señalaba que el proceso de extracción de raíces tendría previsiblemente un impacto negativo en los resultados.

Configuraciones			Conjunto de características					
Con Palabras Funcionales	Raíz	TF-IDF	1-grama	1-2-grama	1-3-grama	1-4-grama	1-5-grama	1-6-grama
✓	✓	✓	**70.44 %**	**71.18 %**	**71.09 %**	**71.15 %**	**71.08 %**	**71.12 %**
✓	✓	✗	70.39 %	70.36 %	70.38 %	70.53 %	70.43 %	70.47 %
✓	✗	✓	70.17 %	70.61 %	70.59 %	70.56 %	70.66 %	70.57 %
✗	✓	✓	68.91 %	69.81 %	69.77 %	69.79 %	69.92 %	69.84 %
✗	✓	✗	68.47 %	69.57 %	69.44 %	69.59 %	69.63 %	69.61 %
✗	✗	✓	68.48 %	69.28 %	69.19 %	69.11 %	69.14 %	69.11 %
✓	✗	✗	70.17 %	70.08 %	70.15 %	70.10 %	70.10 %	70.10 %
✗	✗	✗	67.64 %	68.18 %	68.01 %	68.07 %	68.07 %	68.00 %

Tabla 4: Promedios de exactitud de los clasificadores por conjuntos de características (n-gramas de palabras) y diferentes configuraciones.

Características	NB	LR	RF	SVM	Promedio
1-grama	69.91 %	69.54 %	70.12 %	72.18 %	70.44 %
1-2-grama	**70.91 %**	70.20 %	70.91 %	72.69 %	**71.18 %**
1-3-grama	70.69 %	70.15 %	70.96 %	72.57 %	71.09 %
1-4-grama	70.63 %	**70.39 %**	70.85 %	**72.74 %**	**71.15 %**
1-5-grama	70.70 %	70.11 %	**70.97 %**	72.53 %	71.08 %
1-6-grama	70.64 %	70.28 %	70.95 %	72.59 %	71.12 %
	70.58 %	70.11 %	70.79 %	**72.55 %**	

Tabla 5: Promedios de exactitud de clasificadores por conjuntos de características, utilizando las siguientes configuraciones: con palabras funcionales, con extracción de raíces y TF-IDF.

Para obtener el mejor modelo de clasificación, buscamos el algoritmo de clasificación que funcione mejor con características específicas. Se calculó un promedio de los porcentajes de validación cruzada de cada algoritmo de aprendizaje con los pre-procesamientos. Los resultados se pueden observar en la Tabla 5, donde señalamos los mejores porcentajes de exactitud para cada uno de los sistemas de aprendizaje los cuales son 70.91 % para NB, 70.39 % para LR, 70.97 % para RF y 72.74 % para SVM. Aquí se puede observar que diferentes conjuntos de características funcionan mejor con diferentes clasificadores, sin embargo en promedio, los 2-gramas y los 4-gramas tienen un mejor rendimiento. En cuanto al desempeño promedio de los clasificadores, se puede observar que las SVM obtienen mejores predicciones que el resto en términos de exactitud.

Como hipótesis adicional se consideraron n-gramas de caracteres y se realizaron los mismos experimentos obteniendo la Tabla 6 donde se reporta el promedio del rendimiento de los 4 clasificadores por conjuntos de características. Podemos observar que se obtienen promedios bajos en comparación con la Tabla 4. Por lo anterior, concluimos que para este problema las características extraídas a nivel de palabra funcionan mejor.

Considerando que el modelo SVM fue el algoritmo con mejores promedios de exactitud, en la Tabla 7 mostramos su reporte de clasificación donde podemos observar que la clase en la que se tuvo una mayor precisión y exhaustividad fue la clase de 'alegría' obteniendo un porcentaje del 81 % en predicciones correctas, mientras que la clase 'no_alegría' obtuvo el promedio más bajo en predicciones correctas lo que podríamos asociar con una deficiencia en la detección de sarcasmo dentro de los tuits. Finalmente, el promedio de precisión del algoritmo entre todas las clases fue de un 75 % mientras que su promedio de exhaustividad fue de un 59 %, esto nos da un porcentaje de predicciones correctas del 63 % hechas por el modelo SVM de acuerdo a la medida F_1 y una exactitud del 74 % en sus predicciones.

Adicionalmente, se muestran las características que han resultado más relevantes para la clasificación en las clases A (Figura 2), C (Figura 3), F (Figura 4), N (Figura 5) y P (Figura 6). La revisión de estas figuras ayuda a comprender cómo funciona el proceso de distinguir entre las diferentes clases. Por ejemplo, el clasificador no toma muy en cuenta bigramas como 'feliz cumpleaños' en la categoría A pero les confiere la mayor importancia en la categoría de F, donde se englo-

Configuraciones			Conjunto de características					
Con Palabras Funcionales	Raíz	TF-IDF	1-grama	1-2-grama	1-3-grama	1-4-grama	1-5-grama	1-6-grama
✓	✓	✓	53.25 %	**64.18 %**	**65.50 %**	66.53 %	67.05 %	67.37 %
✓	✓	✗	53.73 %	61.20 %	64.76 %	66.05 %	66.73 %	67.00 %
✓	✗	✓	**53.81 %**	61.77 %	65.46 %	**66.64 %**	**67.28 %**	**67.52 %**
✗	✓	✓	52.32 %	59.86 %	63.79 %	64.89 %	65.45 %	65.73 %
✗	✓	✗	52.31 %	59.19 %	62.63 %	64.09 %	64.53 %	64.87 %
✗	✗	✓	53.29 %	61.00 %	64.33 %	65.34 %	65.96 %	66.20 %
✓	✗	✗	53.73 %	61.20 %	64.76 %	66.05 %	66.73 %	67.00 %
✗	✗	✗	53.22 %	60.49 %	62.86 %	64.45 %	65.15 %	65.40 %

Tabla 6: Promedios de exactitud de los clasificadores por conjuntos de características (n-gramas de caracteres) y diferentes configuraciones

Classe	Precisión	Exhaustividad	F_1
A	0.75	0.89	0.81
C	0.69	0.80	0.74
F	0.85	0.58	0.69
N	0.66	0.22	0.33
P	0.82	0.46	0.59
Promedio	0.75	0.59	0.63
Exactitud			0.74

Tabla 7: Reporte de clasificación para el modelo Support Vector Machine.

ban los tuits de felicitación. En la categoría C, en cambio, palabras como 'aprender', 'consejo' o 'éxito' se encuentran entre las más relevantes, mientras que los n-gramas preferidos para N no tienen en realidad relación con lo que normalmente entendemos por felicidad. Por último, en la categoría de la publicidad (Figura 6), son importantes hashtags como 'siguemeytesigo' o términos como 'coaching'.

Junto con eso, es fácil observar en cada figura cómo algunas de las características consideradas parecen no tener mucho sentido. Además, se puede inferir que un pre-procesamiento más preciso podría tener efectos positivos en el resultado final de clasificación. Finalmente, todos los experimentos y resultados están disponibles en el repositorio de GitHub[2].

Para finalizar esta sección haremos una comparación entre el modelo que obtuvo mejores porcentajes en el rendimiento promedio y un modelo pre-entrenado. Para ello se ha considerado el modelo *BERT multilingual uncased* (Devlin et al., 2018) entrenado por cuatro épocas, y se ha aplicado al corpus con el siguiente preprocesamiento: eliminación de los hiperenlaces y eliminación de los signos de puntuación. Esto debido a que los modelos del lenguaje no necesitan extracción de raíces ni TF-IDF.

El modelo BERT obtuvo una exactitud final de 77.33 %. La exactitud por cada época de entrenamiento en el BERT se puede observar en la Tabla 8.

Época	Exactitud
1	0.75
2	0.79
3	0.77
4	0.77

Tabla 8: Reporte de exactitud obtenida por BERT para cada época de entrenamiento.

La Tabla 9 muestra el reporte de clasificación obtenido para el modelo BERT, mientras que la Tabla 7 muestra el reporte de clasificación para el modelo Support Vector Machine. La exactitud obtenida por BERT supera en 0.03 a la exactitud obtenida por el modelo SVM. Esto quiere decir que el modelo BERT tiene un mayor porcentaje de predicciones correctas. Sin embargo, si evaluamos la métrica de precisión podemos observar que el modelo SVM es 0.03 superior a BERT. La precisión por clase indica la proporción de instancias correctamente clasificadas en cada clase. Aquí podemos ver que aunque el modelo SVM no logra capturar muchas instancias de la clase **N** (solo el 22 % según la métrica de exhaustividad), las instancias clasificadas en esta clase son correctas en un 66 %, en comparación del 53 % del modelo de BERT. En contraparte, la métrica de exhaustividad nos dice que el modelo BERT logra identificar más instancias de cada clase en un promedio de 70 % respecto al modelo SVM que solo identifica un 59 % de instancias de cada clase, en promedio.

[2] https://github.com/GIL-UNAM/TwitterHappiness

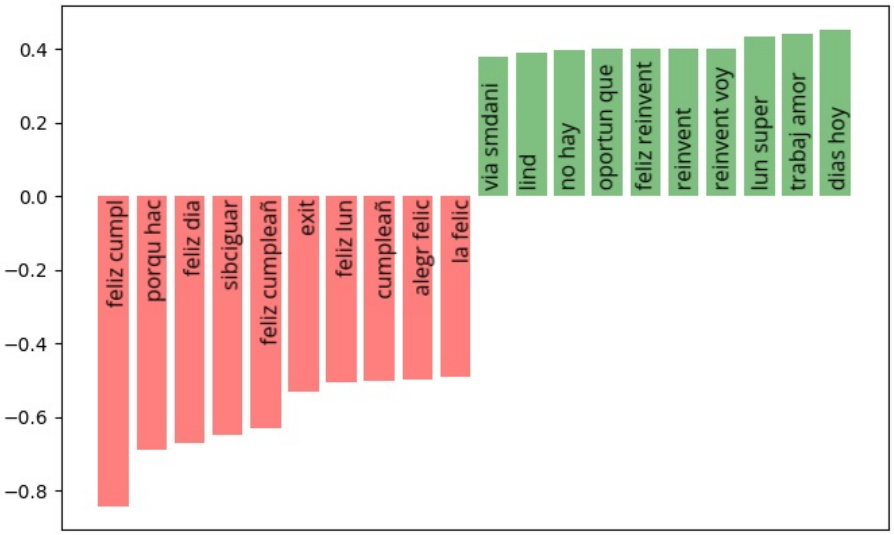

Figura 2: Top características de menor y mayor relevancia para la categoría A con SVM.

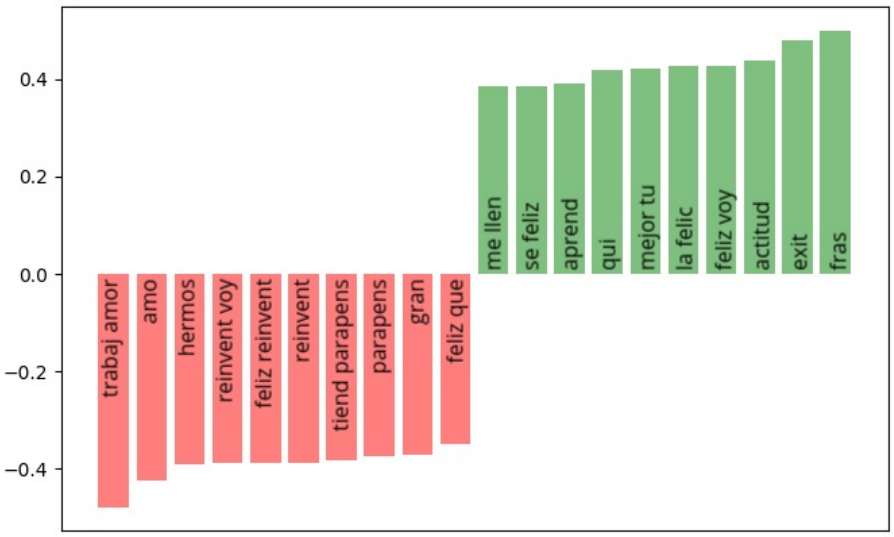

Figura 3: Top características de menor y mayor relevancia para la categoría C con SVM.

Classe	Precisión	Exhaustividad	F_1
A	0.83	0.84	0.83
C	0.77	0.85	0.81
F	0.73	0.74	0.74
N	0.53	0.43	0.48
P	0.74	0.65	0.69
Promedio	0.72	0.70	0.71
Exactitud			0.77

Tabla 9: Reporte de clasificación para el modelo BERT.

A pesar de los buenos resultados obtenidos en la evaluación del clasificador basado en el modelo *BERT multilingual uncased*, no es posible analizar directamente las características utilizadas por este modelo para realizar la clasificación (Yeh et al., 2020). Si bien existen técnicas para explicar cómo un clasificador realiza sus predicciones, basadas en el cálculo de un modelo local alrededor de la predicción a explicar (Rogers et al., 2021), estas no son necesariamente precisas (Pruthi et al., 2019). Serrano & Smith (2019) demuestran que los pesos de las atenciones no necesariamente corresponden con la importancia que tienen éstas dentro de los modelos.

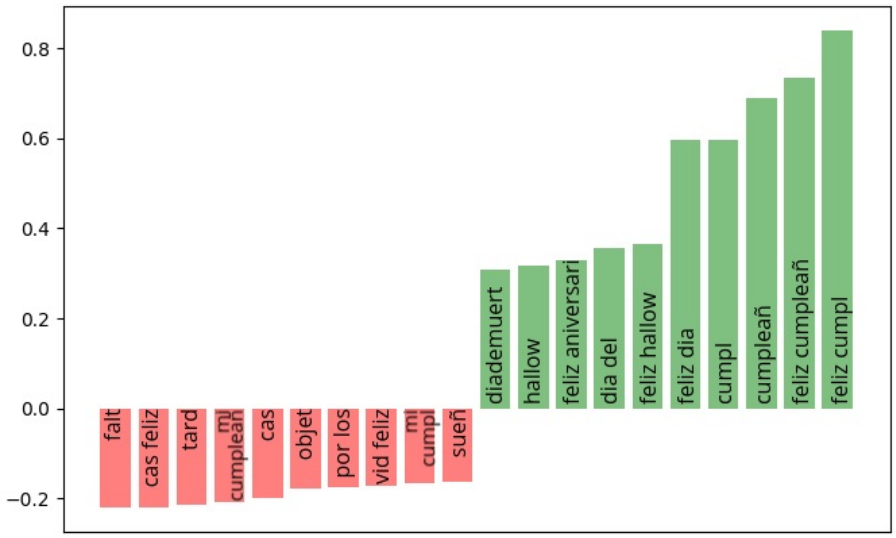

Figura 4: Top características de menor y mayor relevancia para la categoría F con SVM.

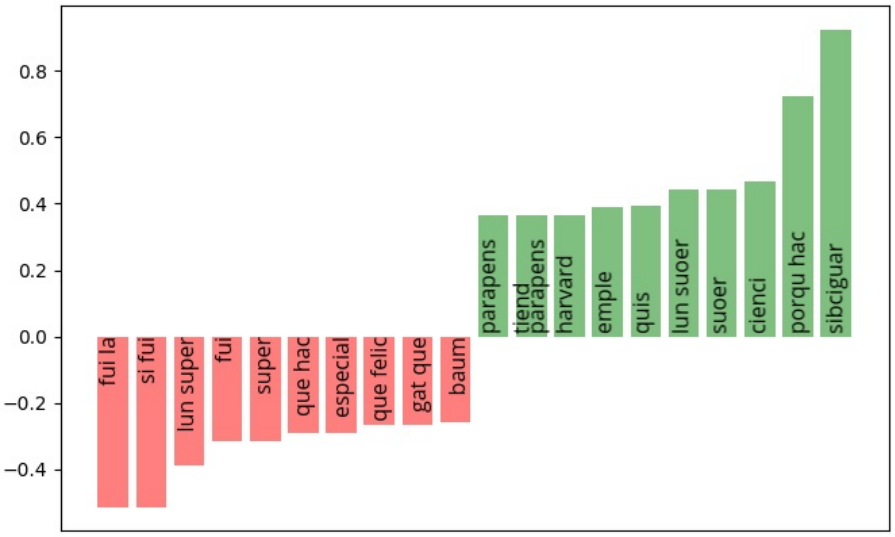

Figura 5: Top características de menor y mayor relevancia para la categoría N con SVM.

Por lo anterior, decidimos no realizar un análisis de las características del modelo *BERT multilingual uncased* sobre nuestro corpus. Sin embargo, este trabajo puede resultar interesante para entender mejor los modelos neuronales. Una posible ruta para este análisis es el modelo SELFEXPLAIN de Rajagopal et al. (2021), el cual incluye una capa interpretable globalmente que identifica los términos más relevantes dentro de un conjunto de entrenamiento, así como una capa interpretable local que permite calcular la contribución de cada input local respecto a una clase a predecir.

6. Conclusiones y trabajo futuro

Este artículo aporta una perspectiva adicional al estudio de las emociones en Twitter, en concreto centrándose en la felicidad. En general, los métodos para detectar la alegría basados en diccionarios han dado resultados muy consistentes. Nosotros nos preguntamos si siempre que se hace uso de este léxico el tuit denota realmente felicidad. Para la investigación nos hemos basado en un estudio anterior (Sidorov et al., 2016), donde se indican cinco categorías posibles que usan palabras relacionadas con el campo semántico que nos atañe: alegría, publicidad, felicitación, consejo y sarcasmo. Con este criterio se ha elabora-

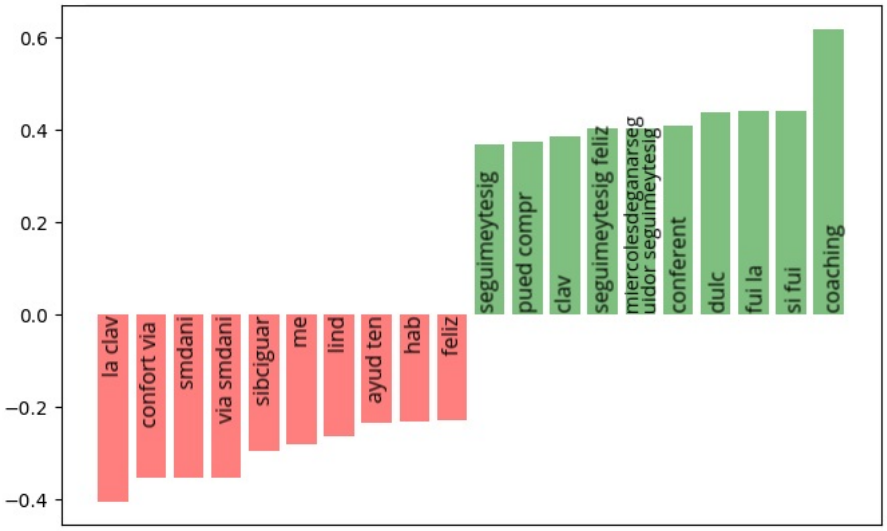

Figura 6: Top características de menor y mayor relevancia para la categoría P con SVM.

do un corpus obtenido con hashtags que remiten normalmente a la felicidad. Dicho corpus se ha etiquetado teniendo en cuenta estas categorías.

Los resultados explicitan que menos de la mitad de los tuits se refieren efectivamente a la emoción alegría, mientras un número elevado (21.5 %) se pueden considerar como consejos y reflexiones, e incluso casi el 14 % no suscita consenso entre los anotadores. El análisis léxico por categoría muestra que las variación en las palabras utilizadas en cada categoría es muy pequeña, de manera que el problema de distinguir entre unos significados o otros se convierte en una tarea complicada. Para abordarlo, se han implementado sistemas de aprendizaje automático con a) distintos tipos de pre-procesamiento y características; b) distintos clasificadores. Los mejores resultados se han obtenido con un pre-procesamiento que consiste en la inclusión de palabras funcionales, extracción de raíz y TF-IDF, con un sistema SVM. Pruebas adicionales han reportado que el uso de BERT produce una leve mejora, tanto en la exactitud como en F_1. Aún así, los mejores resultados de BERT no han sobrepasado el 78 % de exactitud.

Para el futuro, se propone el uso de redes neuronales para resolver este mismo problema, así como la extensión de la metodología al resto de emociones de Ekman.

El trabajo sobre identificación de emociones en redes sociales parece ser una tarea dura, ya que la denotación de los términos y el sentido general de los tuits no siempre se puede deducir directamente del significado connotativo del léxico utilizado. Este artículo es una primera aproximación a la clasificación de emociones según el sentido último que transmite el mensaje. Esta tarea se enmarca dentro del tratamiento computacional de procesos pragmáticos del lenguaje natural, entre los que se pueden contar el sentido figurado, la segunda intención o la necesidad de persuasión.

Reconocimientos

El presente trabajo se ha realizado con el apoyo de los proyectos CONACyT CB A1-S-27780, y DGAPA-UNAM PAPIIT números TA400121 y TA101722. Los autores agradecen al CONACYT por los recursos de cómputo brindados a través de la Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje del Laboratorio de Supercómputo del INAOE.

Referencias

Bakhtiyar, Syed, Indurthi Vijaysaradhi, Shah Kulin, Gupta Manish & Varma Vasudeva. 2019. Ingredients for happiness: Modeling constructs via semi-supervised content driven inductive transfer learning. Informe técnico. Centre for Search and Information Extraction Lab International Institute of Information Technology Hyderabad, India.

Banea, Carmen, Rada Mihalcea & Janyce Wiebe. 2011. Multilingual sentiment and subjectivity. En *Multilingual Natural Language Processing*, Prentice Hall.

Bansal, Mohit, Claire Cardie & Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. En *International Conference on Computational Linguistics (COLING)*, 15–18.

Camacho Vázquez, Vanessa Alejandra, Grigori Sidorov & Sofía Natalia Galicia Haro. 2018. Automatic detection of negative emotions within a balanced corpus of informal short texts. *Cyberpsychology, Behavior, and Social Networking* 21(12). 781–787. 10.1089/cyber.2018.0207.

Chiorrini, Andrea, Claudia Diamantini, Alex Mircoli & Domenico Potena. 2021. Emotion and sentiment analysis of tweets using BERT. En *EDBT/ICDT Workshops*, vol. 2841, online.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

Ekman, Paul. 1992. An argument for basic emotions. *Cognition and Eotion* 6. 169–200.

Ekman, Paul & Richard Davidson. 1994. *The nature of emotions: fundamental questions*. Oxford University Press.

Go, Alec, Richa Bhayani & Lei Huang. 2009. Twitter sentiment classification using distant supervision. Stanford University. https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

González, José Ángel, Lluís-F. Hurtado & Ferran Pla. 2021. TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. *Neurocomputing* 426. 58–69. 10.1016/j.neucom.2020.09.078.

Gruzd, Anatoliy, Sophie Doiron & Philip Mai. 2011. Is happiness contagious online? a case of twitter and the 2010 winter olympics. En *44th Hawaii International Conference on System Sciences*, 1–9. IEEE. 10.1109/HICSS.2011.259.

Kim, Soo-Min & Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. En *COLING/ACL Main Conference and Poster Session*, 483–490.

Kim, Soo-Min & Eduard Hovy. 2007. Crystal: Analyzing predictive opinions on the web. En *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1056–1064.

Koppel, Moshe & Itai Shtrimberg. 2004. Good news or bad news? let the market decide. En *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 86–88.

Kumar, Akshi, Prakhar Dogra & Vikrant Dabas. 2015. Emotion analysis of twitter using opinion mining. En *8th International Conference on Contemporary Computing (IC3)*, 285–290. 10.1109/IC3.2015.7346694.

Mogilner, Cassie, Sepandar D Kamvar & Jennifer Aaker. 2011. The shifting meaning of happiness. *Social Psychological and Personality Science* 2(4). 395–402. 10.1177/1948550610393987.

Mohammad, Saif. 2012. #emotional tweets. En **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, 246–255.

Mohammad, Saif & Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. *CoRR* abs/1708.03696. http://arxiv.org/abs/1708.03696.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh & Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. En *12th International Workshop on Semantic Evaluation*, 1–17. 10.18653/v1/S18-1001.

Mohammad, Saif & Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. En *11th International Conference on Language Resources and Evaluation (LREC 2018)*, s.pp.

Morency, Louis-Philippe, Rada Mihalcea & Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. En *13th International Conference on Multimodal Computing (ICMI)*, 10.1145/2070481.2070509.

Naderi, Habibeh, Behrouz Haji Soleimani, Saif Mohammad, Svetlana Kiritchenko & Stan Matwin. 2018. DeepMiner at SemEval-2018 task 1: Emotion intensity recognition using deep representation learning. En *12th International Workshop on Semantic Evaluation*, 305–312. 10.18653/v1/S18-1045.

Plutchik, Robert. 1980. *A general psychoevolutionary theory of emotion* 3–33. Academic Press. 10.1016/B978-0-12-558701-3.50007-7.

Pruthi, Danish, Mansi Gupta, Bhuwan Dhingra, Graham Neubig & Zachary Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913* .

Rajagopal, Dheeraj, Vidhisha Balachandran, Eduard Hovy & Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. *arXiv preprint arXiv:2103.12279* .

Rogers, Anna, Olga Kovaleva & Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* 8. 842–866. 10.1162/tacl_a_00349.

Rosá, Aiala & Luis Chiruzzo. 2021. Emotion classification in Spanish: Exploring the hard classes. *Information* 12(11). 438. 10.3390/info12110438.

Serrano, Sofia & Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731* .

Sidorov, Grigori, Sofía Natalia Galicia Haro & Vanessa Alejandra Camacho Vázquez. 2016. Construcción de un corpus marcado con emociones para el análisis de sentimientos en twitter en español. *Revista Escritos BUAP* 1. 1–33.

Thomas, Matt, Bo Pang & Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 327–335.

Turney, Peter. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. En *40th Annual Meeting of the Association for Computational Linguistics*, 417–424. 10.3115/1073083.1073153.

Vilares, David, Miguel A Alonso & Carlos Gómez-Rodríguez. 2013. Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento del Lenguaje Natural* 50. 13–20.

Wojatzki, Michael, Torsten Zesch, Saif Mohammad & Svetlana Kiritchenko. 2018. Agree or disagree: Predicting judgments on nuanced assertions. En *7th Joint Conference on Lexical and Computational Semantics*, 214–224. 10.18653/v1/S18-2026.

Yeh, Chih-Kuan, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister & Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems* 33. 20554–20565.

Zeng, Zeyuan, Shaowu Zhang, Lu Ren, Hongfei Lin & Liang Yang. 2021. Senti-bsas: A bert-based classification model with sentiment calculating for happiness research. En *7th International Conference on Computing and Artificial Intelligence*, 272—-277. 10.1145/3467707.3467748.