

Construcción de un corpus marcado con emociones para el análisis de sentimientos en *Twitter* en español

*Grigori Sidorov, Sofía Natalia Galicia Haro
y Vanessa Alejandra Camacho Vázquez*

El análisis de sentimientos (AS) trata de manera computacional opiniones, sentimientos y subjetividad en textos, así como crece exponencialmente en redes sociales y sobre varios problemas. Presentamos el desarrollo de un corpus emocional basado en tweets para el español analizado a mano con la finalidad de crear automáticamente recursos de mayor tamaño y calidad. Se muestra que Twitter se usa tanto para expresar emociones positivas que negativas, al menos en español. También nuestra investigación explora los estudios más relevantes del área en Twitter para brindar un panorama a investigaciones futuras exponiendo carencias y posibles direcciones.

INTRODUCCIÓN

Un *microblog* es un blog con la limitación de ciento cuarenta caracteres en cada entrada, donde una entrada puede ser un texto corto que a menudo expresa una emoción. Estos textos cortos publicados por el usuario son mostrados en su página de perfil y enviados inmediatamente a otros usuarios que han elegido la opción de recibirlos.

La recuperación de información (RI) en *microblogs* siendo *Twitter* el más popular, ha sido expuesta de forma general en la investigación (Efron, Information search and retrieval in microblogs) donde se presentan los principales tipos de búsqueda utilizados, así como los problemas por enfrentar siguientes:

El análisis de sentimientos (AS) porque los *tweets* son informales y seguido expresan tanto emociones como opiniones [(Alexander Pak), (Nicholas A. Diakopoulos), (Andranik Tumasjan), (Brendan O'Connor)].

La popularidad de los metadatos por las múltiples convenciones que amplían la expresividad del texto. Por ejemplo, los *hashtags* son cadenas de caracteres precedidas por el signo “#”, usados para detectar emociones y sentimientos, incluso su recuperación es un tipo de búsqueda de entidad que puede extenderse (Efron, Hashtag Retrieval in a Microblogging Environment).

Además, fue señalada la necesidad de desarrollar sistemas específicos y se concluyó que satisfacer los requisitos del usuario en *microblogs* requerirá gran creatividad en diversas investigaciones.

En este artículo trataremos de cubrir los dos problemas mencionados a través de un análisis que incluirá las investigaciones más relevantes y actuales existentes tanto para el inglés como para el español.

El trabajo presenta conceptos relevantes, líneas base, metodologías, fuentes de datos, recursos utilizados, resultados y conclusiones de cada estudio para brindar una vista global a las futuras investigaciones.

Además, lo anterior ha permitido crear un corpus emocional con *hashtags* en el idioma español abarcando todas sus variedades, es decir, sin limitaciones geográficas. También el análisis manual de dicho recurso arroja resultados favorables que enriquecen nuestra investigación, incluso es propuesta una metodología semiautomática y diversos experimentos ejecutados podrían contribuir en el desarrollo automático de recursos en español.

INVESTIGACIONES DE EXPRESIÓN DE SENTIMIENTOS EN **MICROBLOGS** PARA EL INGLÉS

La investigación (K. B. Mike Thelwall, Sentiment in Twitter Events) mencionó el concepto evento como algo no relacionado

a *Twitter* que engancha un incremento en la frecuencia de una o más palabras dentro de *Twitter*, también el concepto tema general similar a una palabra clave de búsqueda o contenido relacionado con el evento pero no limitado por el tiempo. El objetivo fue encontrar por qué eventos particulares se vuelven generales y comprobar si éstos se relacionan más con incrementos en la fuerza del sentimiento positivo o negativo. La fuente de datos era un corpus de 2,749,840 cuentas diferentes y 34,770,790 *tweets* descargados del 9 de febrero al 9 de marzo del 2010. En el preprocesamiento se convirtieron las palabras de plural a singular (Porter). Además, el recurso utilizado fue *SentiStrength* (K. B. Mike Thelwall).

El estudio (K. B. Mike Thelwall, Sentiment strength detection for the social web) evaluó *SentiStrength 2* para detectar la fuerza del sentimiento en textos cortos de redes sociales y comprobar si era viable en general. El corpus contenía varios conjuntos de datos (*MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums* y un conjunto mezclado por todos) y hubo una confiabilidad de evaluadores con acuerdo moderado usando el coeficiente *Krippendorff's alpha*. Para el preprocesamiento, las palabras vacías o stopwords consideradas comunes como preposiciones y conjunciones, no se removieron por su posible relación con expresiones de sentimientos. La línea base fueron los algoritmos de aprendizaje automático. Las características incluyeron unigramas, bigramas y trigramas, una lista de emoticones y la codificación de puntuación repetida relacionada al sentimiento.

Ambas investigaciones aplicaron diferentes versiones de *SentiStrength* detalladas enseguida:

SentiStrength contiene una lista de 693 palabras de sentimiento con polaridad y fuerza dada por juicios humanos. *SentiStrength 2* agrega a dicha lista 1,617 palabras con términos negativos de *General Inquirer*, luego otras 179 palabras con exclusiones como palabras raras y más largas equivalentes a palabras de sentimientos.

SentiStrength posee otra lista de palabras para la negación que invierte palabras de emoción usando opcionalmente *booster words*. *SentiStrength 2* incluye la negación de términos negativos

para cambiarlos a neutrales más que a positivos, por ejemplo, *I do not hate him* no es positivo.

SentiStrength tiene una lista de idioma que identifica el sentimiento de pocas frases comunes sobrescribiendo la fuerza del sentimiento de cada palabra. *SentiStrength 2* aporta una extensión de frases para palabras de sentimiento comunes, un ejemplo, *is like* tiene valor “uno” para fuerza ya que es un comparador más que un término positivo.

SentiStrength añade una regla que ignora los sentimientos negativos en las preguntas y *SentiStrength 2* remueve dicha regla.

El estudio (Youngguae Bae) describió el concepto audiencia del usuario popular como los seguidores activos o personas que responden, mencionan o *retweetean* acerca del usuario popular. No existen trabajos previos que implementen la medida de influencia propuesta llamada *PN Influence*. La investigación determinó si la audiencia es favorable a los usuarios populares mediante un análisis de sentimientos para usuarios populares y sus audiencias. Mediante consultas hechas cada treinta minutos del 13 de mayo al 8 de julio del 2011, se recolectaron 3,321,387 *tweets* de los usuarios más populares y sus audiencias, éstos últimos contenían las palabras clave @nombre_usuariopopular. Durante el preprocesamiento se removieron *URLs* y palabras específicas (@nombre_usuario). Además, el recurso aplicado fue el diccionario *Linguistic Inquiry and Word Count* (LIWC). LIWC basado en (Andranik Tumasjan), mide emociones negativas, positivas y específicas como ansiedad, incluso contiene 4,500 palabras y raíces de palabras dentro de setenta categorías, además agrega palabras de truncamiento, emoticones [(Alexander Pak), (K. B. Mike Thelwall)] y mide la frecuencia de palabras relacionadas a cada dimensión emocional dentro del *tweet*.

Las metodologías resumidas, experimentos y conclusiones se presentan a continuación:

La investigación (K. B. Mike Thelwall, *Sentiment in Twitter Events*) eligió treinta eventos de una lista de palabras con los más grandes picos de interés, dada por el método de digitalización de series de tiempo (Mario Cataldi). Luego, generó búsquedas boo-

leanas mapeando *tweets* relevantes para cada tema. Después, clasificó la fuerza del sentimiento de cada *tweet* usando *SentiStrength* (K. B. Mike Thelwall) y calificó cada *tweet* sobre una escala de uno a cinco, es decir, sentimiento no presente a sentimiento muy fuerte positivo/negativo, respectivamente.

En los resultados, las pruebas estadísticas dividieron los *tweets* en tres pares de categorías para cada tema (H1n, H2n, H3n, H1p, H2p y H3p) arrojando que los sentimientos negativos juegan un rol importante dentro de los principales *spiking events* en *Twitter* (H1n y H3n aceptados), mientras los sentimientos positivos dan sólo alguna evidencia (H3p aceptado).

En conclusión, el sentimiento negativo es seguido la llave para los eventos populares en *Twitter*. Se comprobó que la cantidad de *tweets* identifica más eventos populares que la fuerza de la emoción expresada. Fue descubierto que generalmente los usuarios no dan su opinión del evento sino su punto de vista humorístico o analítico para su beneficio. Como trabajos futuros se obtendrían resultados para períodos de tiempo mayores y menores a un mes.

La evaluación del estudio (K. B. Mike Thelwall, *Sentiment strength detection for the social web*) probó treinta veces cada algoritmo de aprendizaje automático con WEKA (Mark Hall) aplicando la validación cruzada con el valor diez para la variable *k*, incluso diferentes tamaños de conjunto de características (de cien en cien hasta mil). Entonces se obtuvo el mejor algoritmo y tamaño de conjunto de características para cada conjunto de datos.

Los resultados determinaron que para correlaciones positivas, SLOG 200 superó en desempeño en todos los conjuntos y aún mejor en *Twitter* al comprender que la mayoría de las *URLs* son sugerencias positivas aunque para correlaciones negativas, el aprendizaje automático sólo aventajó en tres conjuntos incluido *Twitter*. Por último, *SentiStrength* ganó en el conjunto final que mezcla múltiples dominios y géneros.

La conclusión, el AS basado en *SentiStrength* es robusto y de manera relativa independiente del dominio (Maite Taboada), (Georgios Paltoglou). *SentiStrength* debe aplicarse donde es importante explotar sólo los términos afectivos directos y evitarse en

aplicaciones comerciales, debates de noticias o corpus con más de mil textos marcados porque algoritmos como SLOG lo superan. Para trabajos futuros, serían examinadas más redes sociales, así como la detección de bromas, sarcasmo e ironía.

El estudio (Youngguae Bae) eligió los trece usuarios más populares de *Twitter* con más de un millón de seguidores (excepto *Donald Trump*) durante mayo del 2011. Posteriormente, detectó audiencias positivas y negativas de cada usuario popular calculando la puntuación del sentimiento y polaridad para cada *tweet* y usuario, así como las medidas *PN Tweet Ratio (i)*, *PN User Ratio (i)* y *PN Influence*. *PN Influence* incorpora medidas de *retwiteos*, respuestas y menciones al tamaño de la audiencia de un usuario popular, también el *PN user ratio* para un tiempo específico porque la influencia del usuario popular cambia con el tiempo. Después, generó el coeficiente de relación *Pearson* entre cada cambio de sentimiento de un usuario popular y su audiencia diariamente.

Los resultados, en *PN User Ratio* las puntuaciones más alta y baja fueron *Lady Gaga* 4.22 y *CNN Breaking News* 0.57. También se comprobó si la medida *PN Influence* de usuarios populares se relaciona a los sentimientos del mundo real mediante dos experimentos.

El primer experimento del 13 de mayo al 7 de julio del 2011, seleccionó a *Barack Obama* para evaluar la aprobación de su trabajo diario mediante la obtención de opiniones en *Twitter* y el *Gallup Daily*. El análisis de correlación aplicado fue el análisis de causalidad *Granger*. Los resultados mostraron que cambios de sentimiento negativos en *Twitter* influyen más sobre los cambios de sentimiento negativos en el mundo real que los cambios de sentimiento positivos. Entonces *Twitter* podría usarse para predecir panoramas del mundo real.

El segundo experimento del 21 de mayo al 9 de julio del 2011, eligió a *Lady Gaga* y *Britney Spears* para determinar si cambios de sentimiento en *Twitter* se relacionan con el *Billboard Weekly Chart* en la posición de sus canciones semanalmente. El análisis de correlación de rangos *Spearman* fue aplicado. Se comprobó

que cambios dentro del total de la audiencia y la audiencia positiva de *Twitter* tienen correlaciones significativas con los resultados del *Billboard Weekly*.

En resumen, los seguidores no son siempre favorables a los usuarios populares. Además, fue comprobado que sentimientos en *tweets* de usuarios populares influenciaron su audiencia, de hecho, *Lady Gaga* y *Barack Obama* tuvieron más influencia positiva sobre sus audiencias y se descubrió que éstas son más largas. En cuanto a trabajos futuros se aplicaría *PN influence* en un algoritmo de influencia global y tal medida sería comparada con otras medidas de influencia en *Twitter*.

La investigación (Antonio Reyes) analizó humor e ironía en *tweets* mediante un modelo basado en las siguientes características textuales con sus respectivas notas iniciales:

Ambigüedad. A mayor ambigüedad estructural, mayor probabilidad de ser un texto humorístico. Incluso, la ambigüedad morfosintáctica o dependencias sintácticas determinan la complejidad de textos humorísticos y no humorísticos.

Polaridad. Hay más elementos positivos etiquetados mediante el *Lexicon Macquarie Semantic Orientation* (MSOL) con 76,400 entradas (Saif Mohammad).

Desequilibrio contextual. A menor semántica, mayor desequilibrio contextual en textos divertidos e irónicos.

Escenarios emocionales. La ironía se caracteriza por palabras abstractas y emocionales para definir un esquema de contextos emocionales favorables y desfavorables.

La fuente de datos constó de cincuenta mil *tweets* divididos en cinco conjuntos de igual tamaño. Un conjunto sin requisitos pero los otros con un *hashtag* *#humor*, *#irony*, *#politics* y *#technology*. Como características fueron empleados trigramas con interpolación y *Kneser-Ney* descuento, un método de suavizado. En cuanto al preprocesamiento de características polaridad y desequilibrio contextual, los *tweets* fueron *stemmed*, es decir, se obtuvieron las raíces de cada palabra, luego fueron descartadas las *stopwords*. Además, los recursos utilizados fueron el analizador sintáctico, el *Whissell's Dictionary of Affect in Language*, *Google N-gram* y el

conjunto de herramientas *Stanford Research Institute Language Modeling* (SRILM).

Cada característica textual se evaluó de distinta forma. Para la ambigüedad, la complejidad de la sentencia etiquetó *tweets* con etiquetas-POS reduciendo la ambigüedad producida con un módulo de *desambiguación* POS y ejecutó un analizador sintáctico. En la polaridad, fue usada la base de datos MSOL con la fórmula tres en el AS, dicha fórmula es la adición de sumatorias de los conjuntos de palabras positivas y negativas, dividido entre la longitud. El desequilibrio contextual se calculó con la medida de *Resnik* (Ted Pedersen) y midió la similitud semántica entre palabras, el contexto abarcó tres palabras y un método *backoff* o de retroceso que asignó el sentido más frecuente. En cuanto a los escenarios emocionales, se calcularon puntuaciones para casi nueve mil palabras en inglés mediante el diccionario (Whissell), con un rango de uno (pasivo, difícil formar imagen mental, desagradable) a tres (activo, fácil formar imagen mental, agradable), por ejemplo, flor es pasiva con valor uno, fácilmente representable de valor tres y tiende a ser agradable, por lo tanto, su puntuación final es 2.75.

Prácticamente, el modelo se evaluó en las dimensiones representatividad y relevancia. Para la representatividad fueron creados cinco modelos lingüísticos (uno para cada conjunto de datos) y uno más representativo usando *Google N-gram*. También se calculó la perplejidad para cada conjunto (PC) con SRILM y su proporción correspondiente a cada *tweet* (PT), la cual comparó los lenguajes del modelo entre el conjunto de datos al que pertenece y *Google*. PT es la longitud del *tweet* por PC, entre el conjunto de datos igual a diez mil.

La relevancia determinó la capacidad de clasificar automáticamente textos dentro del conjunto de datos al que pertenecían. Cada *tweet* de los cincuenta mil se convirtió en un vector de palabras ponderado por frecuencia y basado en su proporción de perplejidad obtenida anteriormente. Entonces se usó un árbol de decisión para clasificar los textos y cuatro clasificadores se entrenaron considerando un escenario binario, es decir, un conjunto

positivo siempre se comparó contra uno negativo. Todos los clasificadores consideraron un 70% para entrenamiento y un 30% para pruebas.

En resultados, los clasificadores tres y cuatro evaluaron todas las características ambigüedad, polaridad, inesperado y escenarios emocionales. El clasificador tres consideró los conjuntos de datos (humor contra ironía, política, tecnología y general) y el clasificador cuatro (ironía contra humor, política, tecnología y general), al final obtuvieron los mejores valores de evaluación $F1 = 0.93$ en humor contra ironía y $F1 = 0.92$ en ironía contra política correspondientemente.

Las observaciones finales para las características textuales fueron los textos humorísticos e irónicos representados por estructuras bien formadas de poca complejidad sintáctica dentro de la ambigüedad morfosintáctica. En la ambigüedad semántica aparecieron relevantes estrategias semánticas que presentan mínimos dos interpretaciones posibles. Respecto a la polaridad, los textos humorísticos tienen palabras positivas más representativas aunque en textos irónicos se concentran más palabras negativas. Incluso, el desequilibrio contextual es relevante para clasificar textos irónicos y no humorísticos. Además, los escenarios emocionales fueron útiles para clasificar textos humorísticos e irónicos, también se aumentó su efectividad involucrando las categorías activación, imágenes y simpatía.

Como conclusión, las características para representar patrones recurrentes en el lenguaje figurativo humor e ironía arrojaron resultados satisfactorios y aclararon que el humor es más adecuado que la ironía. Los trabajos futuros mejorarían y detectarían nuevas características para ironía, también evaluarían el alcance de las características al calcular su rendimiento con otros tipos de conjuntos de datos y sistemas figurativos.

Otro estudio relevante que podría aplicarse a *Twitter* para profundizar las investigaciones del área (H. Andrew Schwartz) está enfocado a *Facebook*, analizó un enfoque de vocabulario abierto para encontrar variaciones relevantes del lenguaje con respecto al género, edad y personalidad. Es decir, los mismos datos guía-

ron para examinar y comprender el lenguaje que distingue gente. Además, descubrió conexiones extras a las obtenidas por un enfoque de vocabulario cerrado o tradicional palabra con categoría. La fuente de datos se compuso por casi 75,000 voluntarios, quince millones de actualizaciones de estado entre enero del 2009 y octubre del 2011, en total trescientos nueve millones de palabras, así como setecientos millones instancias de características de palabras-frases y temas.

Fue explorado el enfoque conteo de palabras y el enfoque *Differential Language Analysis* (DLA). El Enfoque DLA tuvo como infraestructura la extracción de características lingüísticas (palabras-frases y temas), el análisis de correlación y la visualización. La línea base aplicó el lexicón LIWC al corpus, cuestionarios de personalidad con valores a través de un continuo normalizado y temas generados automáticamente con las quince palabras más comunes para el año 2000. Las variables estudiadas fueron género, edad y personalidad, ésta última incluyó ser extrovertido, agradable, escrupuloso, neurótico y sincero. Los resultados progresaron de una prueba variable de concepto género a variables múltiples como grupos de edades y las cinco dimensiones de personalidad. Entonces el lenguaje de vocabulario abierto producido para cada variable fue el siguiente:

Género. En masculino, fuertemente correlacionan *my_wife* y *my_girlfriend*. Para femenino, sólo usan *husband* o *boyfriend*, incluso los preceden con *amazing* o *her*.

Edad. Entre trece y dieciocho años usan más seguido emoticones y palabras como *school*, *homework* y *tomorrow*. El grupo de diecinueve a veintidós años contiene el tema *college* donde resaltan las palabras *semester*, *classes*, *summer*, *registered* y *schedule*, incluso el tema *drunk* que destaca las palabras *hangover*, *sober* y *drunken*. Entre veintitrés y veintinueve años el tema *work* enmarca las palabras *job*, *position*, *company*, *manager*, *interview experience* y *office*, también el tema *beer* remarca las palabras *drinking*, *cold*, *pong* y *ale*. Por último, de treinta a sesenta y cinco años apareció el tema *family* con las palabras destacadas *wonderful*, *daughter*, *son*, *proud*, *friends* y *prayer*.

Personalidad. Ser extrovertido implicó usar palabras sociales como *party, love, you, boys* y *ladies*. Por otro lado, los introvertidos aplicaron palabras de actividades solitarias como *internet, computer, anime, pokemon* y *japanese*. Respecto a ser agradable aparecieron palabras como *music, art* y *writing*. Pero ser neurótico involucró la frase *sick_of* y las palabras *fucking, pissed, depressed* y *stressed*, mientras usuarios estables emocionalmente escribieron sobre actividades disfrutables y sociales como *sports, vacation, beach, church, team* y *family_time*.

Para clasificar la variable binaria de género se aplicó la máquina de vectores de soporte lineal (SVM) y *ridge regression* para predecir la edad y cada factor de personalidad. Las características de vocabulario abierto superaron significativamente a las características de vocabulario cerrado basadas en LIWC. Además, los clasificadores del estudio actual lograron una precisión de 91.9% superando la investigación previa basada enteramente sobre el lenguaje que obtuvo la más alta precisión de 88.0% para predecir el género en *Twitter*.

En resumen, los resultados mostraron validez aparente porque usuarios en bajas elevaciones escribieron de playas, se asociaron con otras investigaciones ya que usuarios neuróticos desproporcionadamente usaron la palabra *depressed* y la frase *sick_of*, sugirieron nuevas hipótesis como una vida activa implica estabilidad emocional y aportaron casos específicos, por ejemplo, los hombres preceden palabras como *wife* o *girlfriend* con el artículo posesivo *my* en mayor grado notoriamente que las mujeres con las palabras *husband* o *boyfriend*.

El enfoque vocabulario abierto permitió descubrir categorías extras mediante conjuntos agrupados de palabras automáticamente dentro de temas coherentes, siendo ejemplos los equipos deportivos, tipos de ejercicio al aire libre o caricaturas japonesas. Finalmente, incluyó un ahorro en mano de obra para crear categorías al considerar todas las palabras encontradas y adaptación a la evolución del lenguaje en redes sociales u otros géneros.

Otro estudio (Alberto Acerbi) analizó emociones en libros para reflejar tendencias más largas de años o décadas. El corpus usó

Google's Ngram con aproximadamente cinco millones de volúmenes digitalmente escaneados, esto representa el 4% de todos los libros publicados hasta el 2008. Dicho corpus contiene textos en inglés británico y americano, además informa cuántas veces un n-grama es usado durante un año dado, es decir, incluye frecuencias de palabras.

Fue concluido que un decremento emocional general podría deberse al cambio de las palabras emocionales porque se usaron listas de palabras contemporáneas en esta investigación. Por otro lado, de 1980 al 2007, un incremento de palabras emocionales para el inglés americano coincide con el aumento de sentimientos antisociales y narcisistas, también un incremento en el porcentaje de pronombres en primera persona singular se relaciona con la disminución en palabras que indican interacciones sociales como “amigo”, “platicar”, “hijo”.

Además se planteó como cuestión interesante, si el uso de las palabras representa el comportamiento real de una población o posiblemente una ausencia de ese comportamiento externada mediante la ficción literaria o el discurso en línea, por ejemplo, en Inglaterra, la supresión del deseo en la era *Elizabethan* aumentó la demanda de escritura obsesionada con el romance y el sexo.

Los trabajos futuros explotarán la dinámica de retroalimentación entre autores y público en general que involucren el conjunto de datos *Google's Ngram*, porque los autores no representan la población general. Explorarían a profundidad la diversidad y analizarían los conjuntos más complejos de rasgos culturales que podrían ser más de diagnóstico que las palabras emocionales o las palabras de contenido libre.

Finalmente, fue expuesto que la gran oportunidad de la validación empírica en estudios dinámicos de la cultura humana se llevará a cabo con la actual disponibilidad y abundancia de datos cuantitativos.

En conclusión, las fuentes de datos construidas para el inglés recopilaron entre cincuenta mil y treinta millones de *tweets* durante un promedio de treinta días, entonces el tamaño del corpus es fundamental para las investigaciones, sin embargo, un estudio

solamente descargó 4,218 *tweets* aunque incluyó otros conjuntos de datos y profundizó el análisis del corpus mediante evaluadores, por lo tanto, agregar calidad al corpus es algo que puede compensar el tamaño.

Investigaciones de expresión de sentimientos en *microblogs* para el español

El incremento de investigaciones en otros idiomas diferentes al inglés podría ser fundamental, debido a que la información no se encuentra en un determinado idioma. Por mencionar, estudios para procesar el lenguaje natural en textos del idioma español han aumentado [(Julian Brooke), (Eugenio Martínez-Cámara), (S. M.-J.-J.-S.-R.-G. Grigori Sidorov), (Sidorov)].

De hecho, investigaciones recientes en español se enfocaron en la normalización de *tweets* (Alejandro Mosquera López), el objetivo fue describir el taller correspondiente del congreso SEPLN 2013 que estandarizó *tweets* usando la herramienta multilingüe TENOR y técnicas para el reconocimiento automático del habla (RAH).

Se expusieron las principales tendencias para normalizar *tweets*, la primera emplea técnicas de traducción automática y requiere un corpus relativamente grande previamente normalizado y alineado para obtener buenos resultados. La segunda se basa en la corrección ortográfica pero en este caso se dificulta la tarea porque el contexto no juega un papel tan relevante. La última tendencia aplica técnicas RAH basadas en que la mayoría de las variantes léxicas no estándar tienen una equivalencia homófona estándar como *ksa* y *casa*, además los sistemas de normalización no supervisada que la han utilizado obtuvieron los mejores resultados.

Para etiquetar las palabras dentro de tres grupos (variantes léxicas, palabras correctas o palabras de otro idioma) y obtener su versión canónica, el sistema ejecutó una metodología que consistió en detectar las variantes léxicas o palabras fuera del vocabulario con TENOR y la herramienta de análisis de lenguaje *Freeling* (Lluís Padró). Luego, TENOR sustituyó las variantes detectadas en el paso anterior mediante el algoritmo de indexado fonético

que arroja listas de palabras con pronunciaciones equivalentes. Los pasos para la sustitución de variantes léxicas fueron:

- Filtrado. Eliminó caracteres no imprimibles y símbolos de puntuación, excepto emoticones.
- Abreviaturas/transliteraciones. Suprimió repeticiones de vocales o consonantes. Después usó tabla de equivalencias para sustituir por la transliteración más apropiada, ejemplos: “separa2”, “ning1”, “c4s4”, “qta1”, “xfa”.
- Indexado fonético. Usó GNU *Aspell* con conjugaciones en diferentes tiempos verbales y entidades nombradas. También empleó el algoritmo del *metaphone* adaptado al español.
- Similitud léxica. Calculada entre la palabra a normalizar y cada candidata fonética del paso anterior mediante el algoritmo *Gestalt* que proporciona un índice de similitud entre cero y cien. Descartó índices menores de sesenta.
- Modelos del lenguaje. Usado cuando hay varias palabras candidatas con la misma similitud léxica. Basado en trigramas y entrenado sobre el corpus CESS-ESP.

La evaluación usó un corpus de 550 *tweets* aproximadamente, descargados entre el día 1 y 2 de abril del 2013, con la restricción de que se localizan en el área geográfica de la península ibérica. Finalmente, el corpus DEV100 con la primera ejecución DLSI-ALICANTE-1 empleó *Freeling* para extraer las palabras fuera del vocabulario y obtuvo la mejor precisión 68.03. En conclusión, TENOR arrojó resultados aceptables que servirán de referencia para mejorar el rendimiento del sistema propuesto en un trabajo futuro.

Otro estudio (David Pinto) para la normalización en SMS, adaptó el algoritmo *Soundex Phonetic* en la representación de dichos textos y aportó una mejora que permitió alcanzar el nivel de similitud entre los textos SMS y su correspondiente texto en inglés o español.

También la investigación (Julio Villena-Román) describió la segunda edición del taller TASS 2013 en el mismo congreso SE-

PLN, aunque con el objetivo de investigar técnicas y algoritmos avanzados para el análisis de sentimientos y la clasificación de opiniones en *tweets*. Las tareas evaluadas fueron el AS global a nivel sentencia, la clasificación de temas, el AS a nivel entidad y la identificación de tendencias políticas con sus mejores valores de evaluación F1 obtenidos 0.616, 0.804, 0.394 y 0.734 respectivamente. Los grupos participantes aumentaron a 31 siendo quince grupos el año anterior, incluso la calidad de los trabajos superó las expectativas y se encuentran disponibles tanto reportes como recursos en español probablemente útiles para otras investigaciones. El corpus general de TASS 2012 ha sido descargado por cincuenta grupos de investigación y se espera el nuevo corpus tenga el mismo impacto.

La investigación (Antonio Fernández Anta) analizó múltiples métodos para clasificar el AS en *Twitter* (inglés), señalando los mejores para *tweets* en español. Empleó como características unigramas, bigramas y trigramas donde cada característica apareció mínimo cinco veces en el corpus y se eliminaron características ambiguas con menos del 85% de sus ocurrencias relacionadas a un sentimiento específico. El preprocesamiento aplicó *lematización* con *Freeling* (Padró). Un algoritmo corrigió la ortografía de forma opcional en el preprocesamiento, antes de extraer atributos y vectores, usando el diccionario *Hunspell*. Otra opción fue transformar emoticones, notaciones cortas y *slang* mediante el diccionario SMS en palabras comprensibles para el diccionario principal *Hunspell*, por ejemplo, :-) se cambió por “feliz”.

Como nota relevante, fue descubierto que la información principal de una sentencia se concentra en pocas palabras clave, siendo adjetivos y verbos en el AS y usó *Freeling* para etiquetar tipos de palabras. También se aplicó un diccionario afectivo para agregar sus palabras como atributos a una lista reduciendo drásticamente su tamaño si los datos de entrada son filtrados. Para las palabras afectivas no encontradas en el corpus se construyeron vectores artificiales por categoría del AS (positivo +, positivo, negativo, negativo+, ninguna).

En la clasificación, los experimentos usaron cinco mil *tweets* para el entrenamiento y dos mil para su evaluación. Se aplicó WEKA similar a (Justin T) para los cinco algoritmos más populares *Ibk*, *Complement Naive Bayes*, *Naive Bayes Multinomial*, *Random Comité* y *SMO*. Cada experimento tuvo su configuración similar a la presentada en el taller TASS 2012, la cual dice al algoritmo los atributos a elegir y cómo crear vectores de atributos junto con su archivo de salida. A partir de la configuración básica, se probaron configuraciones derivadas cambiando solo un parámetro para explorar su efecto en el rendimiento, también para cada algoritmo de clasificación se creó una nueva configuración probada ajustando los parámetros que maximizaron la precisión.

En cuanto a los resultados para la estimación de sentimientos. La combinación de parámetros tuvo algún impacto porque la configuración trece con *Naive Bayes Multinomial* calculó la mejor precisión de 42.38% y el valor promedio de evaluación F1 de las clases fue 0.4 al obtener otros parámetros de rendimiento. Sin embargo, SMO expuso un conjunto más uniforme en valores de precisión siendo los mejores entre 41.38% y 41.93%.

Se concluyó que el conjunto de experimentos expuestos para clasificar *tweets* en español según el sentimiento evaluó el uso de *stemmers* y lematizadores, n-gramas, tipos de palabra, negaciones que invirtieron los signos de las tres palabras consecutivas, cambiadores de valencia positivos que multiplican el peso por tres y negativos por 0.5, procesamiento de enlace, motores de búsqueda, atributos *Twitter* como *hashtags* y cinco métodos de clasificación.

Aunque los *tweets* son muy difíciles de tratar por su brevedad y falta de contexto, los métodos clásicos pudieron aplicarse al español ya que la máxima precisión obtenida de 42% se aproximó a otros valores reportados en el taller TASS.

Por lo tanto, ningún método explorado es el mejor para la clasificación de *tweets* en español y aún se requieren esfuerzos lejanos para progresar la solución de estos problemas. Además, algunas tareas para trabajos futuros deberán incluir temas de normalización y detección de ironía, sentimientos mezclados, subjetividad, etc. Finalmente, la construcción de nuevos corpus con diferentes

variedades del español, un mayor tamaño y calidad podría ser una clave fundamental en el avance del área.

CONSTRUCCIÓN DEL CORPUS DE *TWEETS* EMOCIONALES EN ESPAÑOL

Fue probado el sistema en línea *Archivist* (Archivist) sin resultados satisfactorios porque no permitió implementar filtros de idioma ya que casi la mitad de los *tweets* recuperados estaban en portugués, ni ejecutar varias búsquedas al mismo tiempo teniendo como máximo tres palabras o *hashtags* diferentes, además el servicio no es gratuito. Entonces el uso de la biblioteca *Tweepy* con *Python* (Python) solucionó la implementación **en español sin restricciones de lugares geográficos** y la eliminación de *re-tweets*, también se aplicó una lista de *hashtags* emocionales para varias búsquedas.

El conjunto de datos recuperados de los *tweets* fue conformado por el id, texto, fecha de creación y *URLs* opcionales. Incluso, cada *tweet* debe tener mínimo cinco palabras detectadas como entradas en el diccionario morfológico para el español con más de 555,000 formas que corresponden a más de 76,000 combinaciones lema-POS proporcionado por *Freeling* (Lluís Padró), la finalidad fue asegurar un contexto suficiente para la expresión de emociones excluyendo las *stopwords* de NLTK (Steven Bird), menciones, *URLs*, fechas, etc.

El 7 de mayo del 2014 se creó el corpus aleatorio en español descargando 231,048 *tweets* al azar durante tres horas usando como filtro la preposición más común “de”. Un dato relevante es que dichos *tweets* no contienen los *hashtags* emocionales usados para formar el corpus emocional descrito a continuación.

En principio, el corpus emocional estaba compuesto por 20,942 *tweets* con *hashtags* en español recuperados entre el 6 de mayo y 1 de junio del 2014. Como filtro, cada *tweet* debe tener un *hashtag* emocional en cualquier parte, es decir, no importa su posición ya que su propósito es representar una emoción y esto agregará más casos válidos que las últimas investigaciones ya que solo incluyeron *hashtags* al final. Además, un *tweet* no puede contener otros *hashtags* pertenecientes a emociones contradictorias, este punto tampoco se aplicó en estudios previos. Ejemplos:

- *Tweet* válido: “464562341724844033 gracias dios por un triunfo más #feliz #grado #radiología #universidad #profesión amor a mi trabajo <3 #fb 2014-05-09 00:27:48”.
- *Tweet* descartado: “467802400724905984 Última noche en casa de mis padres! #feliz pero también última noche que duermo con mis niños #triste 2014-05-17 23:02:38”.

La lista de *hashtags* emocionales abarcó palabras que expresan una emoción básica de *Ekman* (Ekman) junto con sinónimos y adjetivos elegidos en base a investigaciones previas como el estudio (Gonzalo Blázquez Gil), el cual usó tres *hashtags* para cada emoción siendo en total dieciocho *hashtags*, también se utilizaron diccionarios en línea [(Diccionario), (WordReference), (Mohammad), (Saif M. Mohammad)]. A continuación se presentan los *hashtags* elegidos para representar cada emoción:

1. Alegría tuvo #felicidad, #feliz, #alegría, #alegre y #contento(a).
2. Tristeza incluyó #tristeza, #triste, #infeliz, #desconsuelo, #desconsolado(a), #nostalgia, #nostalgico(a), #desdicha y #desdichado(a).
3. Ira abarcó #ira, #enojo, #enojado(a), #enfado, #enfadado(a), #furioso(a), #rabia, #colera, #cabreado(a) y #encabronado(a).
4. Asco presentó #asco, #asqueroso(a), #desagradable, #repugnante, #repulsion, #desagrado y #repulsivo(a).
5. Miedo comprendió #miedo, #temor, #terror, #atemorizado(a), #aterrorizado(a), #susto, #asustado(a) y #espantado(a).
6. Sorpresa englobó #sorpresa, #sorprendido(a), #asombro, #asombrado(a), #impactado(a) e impresionado(a).

En esta parte de la investigación, el análisis de género en los *hashtags* posibles fue considerado relevante y el experimento arrojó resultados satisfactorios para cuatro emociones, sin embargo los *hashtags* no mencionados aparecieron cero veces en el corpus emocional. La cantidad de *tweets* recuperados y los mejores

hashtags para cada emoción se detallan enseguida:

- Los *hashtags* más eficientes para la emoción alegría fueron #feliz con 5,040 *tweets* y #felicidad con 3,303 *tweets*. En contraste, #alegre y #contento(a) tuvieron menos de setenta *tweets* cada uno, también #alegría (#alegría) y #alegre no fueron efectivos.
- El mejor *hashtag* para la emoción asco fue #asco con 4,373 *tweets*. En menos de cien *tweets* estuvieron #asqueroso(a), #desagradable y #repugnante cada uno.
- Los mejores *hashtags* para la emoción tristeza fueron #triste con 3,349 *tweets* y #tristeza con 724 *tweets*. Sin embargo, #infeliz apareció en cien *tweets* aproximadamente.
- El *hashtag* más eficiente para la emoción miedo fue #miedo con 2,418 *tweets*. El *hashtag* #terror recolectó casi doscientos *tweets* más enfocados en anuncios de películas, mientras #asustado y #temor aparecieron en menos de cincuenta *tweets* cada uno.

En *Twitter*, la recuperación de textos para las emociones sorpresa e ira ha sido complicada porque se recuperaron menos de cuatrocientos *tweets* para cada una, encontrándose entre los mejores *hashtags* #sorpresa y #sorprendido, así como #enojada, #rabia e #ira. Estos resultados no permitieron confirmar que el género es relevante en los *hashtags* emocionales puesto que tanto #sorprendido como #enojada recuperaron menos de cien *tweets* durante el período completo de descarga.

CREACIÓN DE CONJUNTOS DE **HASHTAGS** QUE ASEGURAN LA EXPRESIÓN DE CADA EMOCIÓN

En base a que tal vez el *hashtag* #dios se relaciona con la emoción alegría y encontrar relaciones entre ciertos *hashtags* podría asegurar el contenido de una emoción. Además, la obtención de nuevos *hashtags* eficientes para detectar emociones requirió la ejecución de otro experimento para determinar conjuntos de *hashtags* correspondientes a cada emoción. Entonces se contabilizaron las veces que aparecieron todos los *hashtags* dentro de los *tweets* co-

respondientes a cada emoción y fueron seleccionados los más comunes, no importó si el *hashtag* contenía una palabra emocional.

Los resultados para ira fueron #mal, #bronca, #frustrada, #odio, #ganasdematar y #rompotodo. Sorpresa obtuvo #cumple, #fiesta, #birthday, #cumpleaños y #regalo. Tristeza arrojó #nostalgia, #sad, #soledad, #decepcionada, #realidad, #dolor y #llo-ro. Asco señaló #vomito, #vergüenza, #vergüenza y #odio. Miedo presentó #panico, #trauma, #temblor, #ansiedad y #nervios. Por último, alegría expuso #bendecida, #teamo, #viernes, #vida, #domingo, #iglesia, #happy, #alegria, #paz, #familia y #love. Las conclusiones son:

- Se comprobó que el *hashtag* #amor apareció 876 ocasiones presentando ambigüedad ya que estaba en la emoción negativa tristeza y la emoción positiva alegría, por lo que #amor fue descartado.
- Aunque #cumple y #love se encontraron en la emoción positiva alegría y la emoción neutra sorpresa, solamente #cumple podría ser un buen candidato para detectar sorpresa. Al final, #odio apareció en asco e ira, ambas emociones negativas y será añadido como un posible candidato eficiente para ira.
- Estos resultados permitirán encontrar patrones y facilitar la construcción automática de recursos en español. Para la búsqueda de patrones también ayudará el análisis exhaustivo hecho a mano del corpus expuesto más adelante.

DESCARGA DE **TWEETS** CON TODOS LOS **HASHTAGS** UTILIZADOS EN ESPAÑOL

Después, fue requerida la construcción del *corpus_htags7dias* con cerca de 1,855,000 *tweets* que contenían *hashtags* aleatorios y generales en español, descargados del 19 al 26 de junio del 2014. Este corpus permitió analizar los *hashtags* emocionales aplicados y sugeridos anteriormente con sus correspondientes plurales y acentos en ciertos casos para profundizar la investigación. También ayudó a detectar *hashtags* compuestos o específicos del español para rastrear primordialmente las emociones más escasas sorpresa e ira en *Twitter*.

ANÁLISIS DE **HASHTAGS** SINGULARES, PLURALES O CON ACENTOS PARA DETECTAR EMOCIONES

Se verificaron casi sesenta y cinco *hashtags* emocionales en singular junto a sus *hashtags* en plural, los resultados confirmaron cuáles son más eficientes para recuperar *tweets* emocionales en español. Los datos constan del *hashtag* y su número de apariciones en el *corpus_tags7dias*, si el *hashtag* apareció cero veces entonces no fue registrado y los resultados para cada emoción son los siguientes:

- Alegría. En singular #feliz y #felicidad con mil cuatrocientos cuarenta y setecientos cincuenta apariciones de manera respectiva, #alegría con sesenta y cinco apariciones, luego #contento(a) y #alegre en menos de veinte apariciones cada uno. En plural #felicidades con ciento noventa apariciones, #felices con setenta y #alegres con una sola aparición.
- Tristeza. Singulares #triste y #tristeza con ciento diez y setenta apariciones de forma correspondiente. El plural #tristes con tan solo diez apariciones.
- Asco. Singular #asco con trescientas cuarenta apariciones mientras los plurales #asquerosos(as), #repugnantes y #desagradables se encontraron menos de diez veces cada uno.
- Miedo. En singular #miedo con doscientas veinte apariciones, #terror con setenta, así como #susto, #panico y #temor en menos de veinte apariciones cada uno. En plural #miedos estuvo en solamente diez ocasiones.
- Sorpresa. Singulares #sorpresa con setenta apariciones, #sorprendida y #asombroso se encontraron en menos de diez ocasiones. El plural #sorpresas se contabilizó en veinticinco ocasiones.
- Ira. Singulares #enojado(a) con menos de diez apariciones cada uno, #agresivo con una sola aparición. El plural #enojados(as) fue rastreado una sola vez.

Por lo tanto, ningún *hashtag* emocional en plural superó los resultados de su correspondiente *hashtag* en singular. Tampoco el

uso de acentos mejoró la recuperación, algunos ejemplos probados fueron nostálgico(a), repulsión y alegría.

DESCUBRIMIENTO DE **HASHTAGS** COMPUESTOS O ESPECÍFICOS PARA AMPLIAR EL CORPUS EMOCIONAL

Aunque el estudio detectó nuevos *hashtags* emocionales más específicos y compuestos con varias palabras, mencionados a continuación:

- Para la emoción alegría se identificaron *hashtags* aparecidos entre 1,100 y 3,100 veces. Estos fueron #happy, #mecontraenamoraque, #enestemomentoquiero y #buenosdias y #buendia. También #graciasati setecientos setenta veces, #gracias y #teamo más o menos seiscientos setenta veces cada uno. Por último, #unabuena noche para, #me encanta y #notieneprecio doscientas veces cada uno.
- Para las emociones negativas entre mil trescientas y tres mil cuatrocientas ocasiones estuvieron #nadamas feo que, #me cae mal y #me hace falta cada uno. Luego, entre cuatrocientas cincuenta a ochocientos cincuenta veces aparecieron #mesaca, #noesdechicasbien, #quisiera que, #5cosas que odio y #cosas que odio cada uno. Por último, #fuck, #me gusta pero y #formas de empezar una discusión menos de trescientas veces cada uno.
- Para la emoción sorpresa estuvieron #aun sigoa sombreado y #aun no puedo creerlo con trescientas treinta apariciones juntos y enfocados en noticias.

Además, el descubrimiento de nuevos *hashtags* brindará la posibilidad de construir recursos en español con un mayor tamaño y calidad.

Finalmente, el corpus *corpus_hhtags7dias* se filtró para **aumentar el tamaño del corpus emocional**, es decir, cada *tweet* tuvo mínimo cinco palabras encontradas en el diccionario morfológico para el español proporcionado por *Freeling*, etc.

En la emoción ira fueron agregados casi tres mil *tweets* con los

hashtags ordenados de manera descendiente en cuanto a su eficiencia #nadamasfeo, #mal, #mecaemal, #fuck, #mesaca, #odio, #bronca, #formasdeempezarunadiscusion y #cosasqueodio.

Para la emoción sorpresa fueron añadidos cerca de mil ochocientos *tweets* con los *hashtags* #regalo, #increible, #interesante, #impresionante, #sorprendente, #aunsigoasombrado y #aunno-puedocreerlo.

En cuanto a la emoción miedo, sus *hashtags* sugeridos al inicio permitieron agregar ciento setenta *tweets* de forma aproximada.

Otros quinientos *tweets* con los *hashtags* #avecesesmejor, #aprendiqueaveces y #todoesmasfacilcuando no se añadieron porque expresan conformidad, similar a un me da igual, ni una emoción positiva ni negativa. También los *hashtags* #frases y #humor aparecieron en casi ochocientas ocasiones cada uno y en menos de doscientas cincuenta ocasiones cada uno se detectaron #frasedeldía, #frasedeldia, #frase, #ironiaon, #ironía, #ironia, #frasesdemaraca, #frasesdecallejeros, #broma y #sarcasmo. Así que serán guardados para futuras pruebas.

ANÁLISIS EXHAUSTIVO A MANO DEL CORPUS EMOCIONAL

Como ha sido mencionado, estudios previos usaron *tweets* con un *hashtag* al final para cada emoción y su objetivo era realizar el mínimo trabajo manual. Sin embargo, comprobamos que tomaron en cuenta una cantidad considerable de *tweets* incorrectos ya que aunque pasaron los filtros, no expresan la emoción requerida.

Para esto, se seleccionó la emoción alegría, la cual obtuvo los mejores resultados en la recuperación de *tweets* con el *hashtag* #feliz. Además, estos *tweets* se referían a diversos contextos y no solamente al emocional. Ya analizados manualmente los cinco mil cuarenta *tweets* fueron detectados ochocientos *tweets* que no expresaron dicha emoción y han sido clasificados en cuatro grupos llamados anuncios publicitarios, frases, felicitaciones y no_feliz.

Antes de proceder con el análisis manual del corpus emocional completo, se ejecutó la siguiente metodología semiautomática con la finalidad de eliminar la mayor cantidad de *tweets* incorrectos. La metodología consiste en detectar los *tweets* con *hashtags* que

tuvieron la menor carga de la emoción requerida, también rastrear los *tweets* con palabras de mayor carga emocional contraria, por ejemplo, para alegría se usaron “triste”, “mal”, etc. Después, encontrar los *tweets* con palabras como chiste(s), broma(s), sarcasmo e ironía (ironía). Al final, una revisión parcial de los *tweets* recuperados en los pasos anteriores permite eliminar aquellos que son incorrectos.

Durante el análisis manual de la emoción asco se comprobó que está estrechamente vinculada con la emoción ira, debido a esto fue difícil encontrar *tweets* para la emoción ira aunque se solucionó descubriendo nuevos *hashtags* efectivos para dicha emoción. Entonces la única emoción escasa en *Twitter* es sorpresa.

Tabla 1. Descripción y resumen de datos del corpus emocional compuesto por *tweets* en español.

Emoción	Cantidad final de <i>tweets</i>	Observaciones relevantes
Alegría	4,245	El 16%, es decir, 795 <i>tweets</i> #feliz no expresaron alegría.
Asco	4,244	Casi 3% de <i>tweets</i> #asco no expresaron dicha emoción.
Tristeza	4,053	El corpus emocional contiene <i>hashtags</i> eficientes simples y compuestos/específicos, los últimos aplicados en las emociones ira y sorpresa. Para la alegría, el vocabulario más utilizado es “dios”, “más”, “mejor”, “vida”, “lindo”, “día” y “gracias”. Mientras que las emociones negativas contienen palabras como “nada”, “quiero”, “mal”, “asco” y “gente”.
Ira	3,510	
Miedo	3,462	
Sorpresa	2,452	
Total	21,966	

La emoción que más *tweets* incorrectos tuvo fue alegría, por lo tanto, una mayor ambigüedad. Hasta ahora parecía ser la emoción

positiva más común expresada en *Twitter* pero ha sido comprobado que tiene múltiples contextos con diversos fines como consejos para alcanzarla, etc. Por otro lado, las emociones con menos *tweets* incorrectos y menor ambigüedad fueron asco y tristeza.

Después del análisis a mano, el corpus final registró aproximadamente para asco y alegría la misma cantidad de cuatro mil doscientos *tweets* cada una (véase Tabla 1). En sorpresa siguen descargándose *tweets* hasta alcanzar los tres mil y pasar al siguiente punto dentro de nuestra investigación.

Pudo comprobarse que la emoción asco es expresada en *Twitter* la misma cantidad de veces que alegría, ya que la emoción positiva contiene mayor ambigüedad, incluso los *tweets* de asco e ira que son emociones negativas entrelazadas superarían a la emoción alegría.

CONCLUSIONES Y TRABAJOS FUTUROS

La información analizada en el estado del arte permitió concluir que no ha sido encontrado el mejor método de clasificación para *tweets* en español dentro del AS y se necesitan múltiples investigaciones para mejorar la solución de estos problemas. También el desarrollo de nuevos recursos con diferentes variedades del español, un mayor tamaño y calidad podrían ser una clave fundamental en el avance del área.

La construcción del corpus emocional comprobó que el género en los *hashtags* no fue relevante para la recuperación de *tweets*. Incluso, el estudio de los correspondientes *hashtags* en plural y con acento para detectar emociones, no arrojaron resultados satisfactorios. Esto fue incluido por ser considerado relevante conforme nuestra investigación avanzó.

Además, fueron descubiertos *hashtags* compuestos o específicos del español para rastrear las emociones en *Twitter* y recolectar una mayor cantidad de *tweets*, dicho experimento no ha sido propuesto ni llevado a cabo en estudios anteriores.

Incluso, el estudio de los siguientes puntos desarrollados dentro de nuestra investigación podría permitir la creación automática de recursos en español:

- Los grupos de *hashtags* obtenidos para cada emoción.
- Los conjuntos de *tweets* con *hashtags* que expresan conformidad, humor e ironía.
- La metodología semiautomática propuesta para detectar posibles *tweets* incorrectos en cada emoción, antes de ejecutar un análisis manual exhaustivo.
- Durante el análisis manual de la emoción alegría, el conjunto extraído de *tweets* que ha sido separado en cuatro grupos porque no expresan dicha emoción.

El análisis manual del corpus emocional corroboró que la única emoción escasa en *Twitter* es sorpresa puesto que la mayoría son textos de noticias y anuncios publicitarios. Luego, fue detectada la misma cantidad de *tweets* con la emoción positiva alegría y la emoción negativa asco, esto permite asegurar que al igual que la emoción positiva, las emociones negativas también son expresadas ampliamente dentro de *Twitter* en español.

Por último, la investigación aporta un corpus emocional de *tweets* en español, sin restricciones de lugares geográficos, es decir, incluye diversas variedades del español y analizado a mano detalladamente. Además, los trabajos futuros serán:

- Recolectar un corpus emocional durante un mayor tiempo para aumentar tamaño y actualizar grupos de *hashtags* emocionales.
- Descubrir patrones para el futuro desarrollo automático de recursos en español.
- Como los *hashtags* utilizados para las emociones negativas contienen menos ambigüedad y aparecen en mayor cantidad de acuerdo a los experimentos realizados, un posterior estudio enfocado solamente a este tipo de emociones podría servir para evitar múltiples problemas sociales como el suicidio y el *bullying* [(Hui Yang), (Kirk Roberts)].
- Detectar frases, humor e ironía en español con los *hashtags* rastreados durante la investigación.
- Detectar emociones mezcladas.

BIBLIOGRAFÍA

Alberto Acerbi, Vasileios Lampos, Philip Garnett, R. Alexander Bentley

2013 «The Expression of Emotions in 20th Century Books.» *PLOS ONE* 8.3, pp. 1-6.

Alejandro Mosquera López, Paloma Moreda

2013 «DLSI en Tweet-Norm 2013: Normalización de Tweets en Español.» *Tweet Normalization Workshop at SEPLN*, pp. 25-29.

Alexander Pak, Patrick Paroubek

2010 «Twitter as a Corpus for Sentiment Analysis and Opinion Mining.» *Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association, pp. 1320-1326.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabella M. Welp

2010 «Predicting elections with twitter: What 140 characters reveal about political sentiment.» *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Washington DC: Association for the Advancement of Artificial Intelligence, pp. 178-185.

Antonio Fernández Anta, Luis Núñez Chiroque, Philippe Morere, Agustín Santos

2013 «Sentiment analysis and topic detection of Spanish Tweets: A comparative study of NLP techniques.» *Procesamiento de Lenguaje Natural* 50, pp. 45-52.

Antonio Reyes, Paolo Rosso, Davide Buscaldi

2012 «From humor recognition to irony detection: The figurative language of social media.» *Data & Knowledge Engineering. Applications of Natural Language to Information Systems* 74, pp. 1-12.

Archivist, Tweet

2013 *Tweet Archivist Simple Poderoso Accesible Análisis de Twitter*. 2013-2014. 01 de mayo de 2014. <<https://es.tweetarchivist.com/>>

Brendan O'Connor, Ramnath Balasubramanian, Bryan R. Routledge, Noah A. Smith

2010 «From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.» *Fourth International AAAI Conference on Weblogs and Social Media*. Washington, D.C., pp. 1-8.

David Pinto, Darnes Vilariño Ayala, Yuridiana Alemán, Helena Gómez-Adorno, Nahun Loya, Héctor Jiménez-Salazar

2012 «The Soundex Phonetic Algorithm Revisited for SMS Text Representation.» *Text, Speech and Dialogue* 7499, pp. 47-55.

Diccionario. *SINONIMOS*

s.f. *Diccionario de sinónimos en español*. 05 de mayo de 2014. <<http://www.sinonimos.com/>>

Efron, Miles

2010 «Hashtag Retrieval in a Microblogging Environment.» *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, pp. 787-788.

2011 «Information search and retrieval in microblogs.» *Journal of the American Society for Information Science and Technology* 62(6), pp. 996-1008.

Ekman, Paul

1992 «An Argument for Basic Emotions.» *Cognition and Emotion* 6.3-4, pp. 169-200.

Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López

2011 «Opinion classification techniques applied to a spanish corpus.» *Natural Language Processing and Information Systems* 6716, pp. 169-176.

Georgios Paltoglou, Mike Thelwall

2012 «Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media.» *ACM Transactions on Intelligent Systems and Technology* 3.4, 66:1, 66:19.

Gonzalo Blázquez Gil, Antonio Berlanga de Jesús, José M. Molina López

- 2013 «Combining Machine Learning Techniques and Natural Language Processing to Infer Emotions Using Spanish Twitter Corpus.» *Highlights on Practical Applications of Agents and Multi-Agent Systems* 365, pp. 149-157.
- Grigori Sidorov, Francisco Velasquez, Alexander Gelbukh, Efstathios Stamatatos, & Liliana Chanona-Hernández
- 2013 «Syntactic Dependency-Based N-grams: More Evidence of Usefulness in Classification.» *LNCS*. Heidelberg: Springer, pp. 13-24.
- Grigori Sidorov, Francisco Velasquez, Alexander Gelbukh, Efstathios Stamatatos, Liliana Chanona-Hernández
- 2013 «Syntactic Dependency-Based N-grams as Classification Features.» *MICAI 2012, Part II LNAI*. Heidelberg: Springer, pp. 1-11.
- Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, Juan Gordon
- 2013 «Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets.» *Advances in Artificial Intelligence* 7629, pp. 1-14.
- H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, Lyle H. Ungar
- 2013 «Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.» *PLOS ONE* 8.9, pp. 1-16.
- Hui Yang, Alistair Willis, Anne de Roeck, Bashar Nuseibeh
- 2012 «A Hybrid Model for Automatic Emotion Recognition in Suicide Notes.» *Biomedical Informatics Insights* 5.1 pp. 1-14.
- Julian Brooke, Milan Tofiloski, Maite Taboada
- 2009 «Cross-Linguistic Sentiment Analysis: From English to Spanish.» *International Conference RANLP*. Borovets, Bulgaria: Association for Computational Linguistics, pp. 50-54.

Julio Villena-Román, Janine García-Morera, Sara Lana-Serrano, José Carlos González-Cristóbal

2014 «TASS 2013 - A Second Step in Reputation Analysis in Spanish.» *Procesamiento del Lenguaje Natural* 52, pp. 37-44.

Justin T, Gajšek R, Štruc V, Dobrišek S.

2010 «Comparison of different classification methods for emotion recognition.» *In MIPRO, 2010 Proceedings of the 33rd International Convention*. Opatija, Croatia : IEEE , pp. 700-703.

Kirk Roberts, Sanda M. Harabagiu

2012 «Statistical and Similarity Methods for Classifying Emotion in Suicide Notes.» *Biomedical Informatics Insights* 5.1 pp. 195-204.

Lluís Padró, Evgeny Stanilovsky

2012 «FreeLing 3.0: Towards Wider Multilinguality.» *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey, pp. 2473-2479.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede

2011 «Lexicon-based methods for sentiment analysis.» *Computational Linguistics* 37.2, pp. 267-307.

Mario Cataldi, Luigi Di Caro, Claudio Schifanella

2010 «Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation.» *In Proceedings of the Tenth International Workshop on Multimedia Data Mining*. New York: ACM Press, pp. 1-10.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten

2009 «The WEKA Data Mining Software: An Update.» *SI-GKDD Explorations* 11.1, pp. 10-18.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou

2011 «Sentiment in Twitter Events.» *Journal of the American Society for Information Science and Technology* 62.2, pp. 406-418.

- 2012 «Sentiment strength detection for the social web.» *Journal of the American Society for Information Science and Technology* 63.1, pp. 163-173.
- Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai
 2010 «Sentiment Strength Detection in Short Informal Text.» *Journal of the American Society for Information Science and Technology* 61.12, pp. 2544-2558.
- Mohammad, Saif M.
 2012 «#Emotional tweets.» *SemEval '12 Proceedings of the First Joint Conference on Lexical and Computational Semantics. Volume 1: Proceedings of the main conference and the shared task and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Stroudsburg, USA: Association for Computational Linguistics (ACM), pp. 246-255.
- Nicholas A. Diakopoulos, David A. Shamma
 2010 «Characterizing Debate Performance via Aggregated Twitter Sentiment.» *In Proceedings of the 28th International Conference on Human Factors in Computing Systems*. New York, USA: ACM Press, pp. 1195-1198.
- Padró, Lluís, Reese S., Agirre E., y Soroa A.
 2010 «Semantic services in freeling 2.1: Wordnet and ukb.» *In Global Wordnet Conference 2010*, pp. 99-105.
- Porter, M. F.
 1997 *Readings in information retrieval. An algorithm for suffix stripping*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Python, Central
 2011-2014 *Introduction to tweepy, Twitter for Python*. 02 de mayo de 2014. <<http://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python/>>.
- Saif M. Mohammad, Svetlana Kiritchenko
 2014 «Using Hashtags to Capture Fine Emotion Categories from Tweets.» *Semantic Analysis in Social Media, Computational Intelligence* 30.3, pp. 1-22.

Saif Mohammad, Cody Dunne, Bonnie Dorr

2009 «Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus.» *Proceedings of the 2009 Conference on EMNLP*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 599-608.

Sidorov, Grigori

2013 *Non-linear construction of n-grams in computational linguistics: syntactic, filtered, and generalized n-grams*. primera. D.F.: Sociedad Mexicana de Inteligencia Artificial.

Steven Bird, Ewan Klein, Edward Lope

2009 *Natural Language Processing with Python*. Sebastopol: O'Reilly.

Ted Pedersen, Siddharth Patwardhan, Jason Michelizzi

2004 «Wordnet::similarity - measuring the relatedness of concepts.» *Proceeding HLT-NAACL- Demonstration Papers 2004*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 38-41 .

Thelwall M., Buckley K., Paltoglou G.

2012 «Sentiment strength detection for the social web.» *Journal of the American Society for Information Science and Technology* 63.1, pp. 163-173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A.

2010 «Sentiment strength detection in short informal text.» *Journal of the American Society for Information Science and Technology* 61.12, pp. 2544-2558.

Whissell, Cynthia

2009 «Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language.» *Psychological Reports* 105.2, pp. 509-521.

WordReference

2014 *Diccionario de sinónimos y antónimos - WordReference*. <<http://www.wordreference.com/sinonimos/>>

Yao Lu, Xiangfei Kong, Xiaojun Quan, Wenyin Liu, Yinlong Xu

2010 «Exploring the Sentiment Strenght of User Reviews.» *Web-Age Information Management. Lecture Notes in Computer Science* 6184, pp. 471-482.

Youngguae Bae, Hongchul Lee

2012 «Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers.»
Journal of the American Society for Information Science and Technology 63.12, pp. 2521-2535.

PALABRAS CLAVE DEL ARTÍCULO Y DATOS DE LOS AUTORES

Corpus de tweets, análisis de sentimientos (AS), Twitter, detección de emociones, algoritmos de aprendizaje automático, recuperación de información (RI) en microblogs

Grigori Sidorov

Instituto Politécnico Nacional

Centro de Investigación en Computación (CIC)

Av. Juan de Dios Bátiz, Esq. Miguel Othón de Mendizábal,

Col. Nueva Industrial Vallejo, (G.A.M.)

CP 07738

Teléfono: 57296000 ext. 56518

e-mail: sidorov@cic.ipn.mx

Sofía Natalia Galicia Haro

Universidad Nacional Autónoma de México

Facultad de Ciencias

Circuito Exterior S/N, Ciudad Universitaria

Col. Copilco El Bajo, Delegación Coyoacán

CP 04510

Teléfono: +52 (55) 56223899 ext. 45789

e-mail: sngh@fciencias.unam.mx

Vanessa Alejandra Camacho Vázquez

Estudiante del Doctorado en Ciencias de la Computación

Instituto Politécnico Nacional

Centro de Investigación en Computación (CIC)

Av. Juan de Dios Bátiz, Esq. Miguel Othón de Mendizábal,

Col. Nueva Industrial Vallejo, (G.A.M.)

CP 07738

Teléfono: 5759443

e-mail: vcamachob10@sagitario.cic.ipn.mx