# Language Detection on Twitter

Yudivián Almeida-Cruz
yudy@matcom.uh.cu
*Universidad de La Habana*

Suilan Estévez-Velarde
sestevez@matcom.uh.cu
*Universidad de La Habana*

Alejandro Piad-Morffis
apiad@matcom.uh.cu
*Universidad de La Habana*

**ABSTRACT**

The paper presents an alternative to identify languages on Twitter without having to use training sets or aggregated information. Such alternative is based on trigram recognition algorithms and small words techniques. The use of these algorithms is evaluated both on their own and in a model of composition. Also, the incidence of pre-processing of tweets in the accuracy of identifying the language is discussed. Finally, after a process of experimentation, the best alternative, out of those studied, is determined.

**KEYWORDS:**
Small words; Twitter; Language detection; n-grams; trigrams

# Detección de Idioma en Twitter

Yudivián Almeida-Cruz
yudy@matcom.uh.cu
*Universidad de La Habana*

Suilan Estévez-Velarde
sestevez@matcom.uh.cu
*Universidad de La Habana*

Alejandro Piad-Morffis
apiad@matcom.uh.cu
*Universidad de La Habana*

**RESUMEN**

El trabajo presenta una alternativa para identificar idiomas en Twitter sin que sea necesario utilizar conjuntos de entrenamiento o información agregada. En dicha alternativa se utilizan técnicas basadas en los algoritmos de reconocimiento de trigramas y small words. Se valora la utilización de estos algoritmos por sí solos y en un modelo de composición. Asimismo, se analiza la incidencia del pre-procesamiento de los tweets en la precisión de la identificación de los idiomas. Finalmente, después de un proceso de experimentación, se determina la mejor alternativa de las estudiadas.

**PALABRAS CLAVE:** Detección de Idiomas, n-gramas, trigramas, small words, twitter.

## INTRODUCCIÓN

Internet es la principal plataforma para el intercambio de datos. En ella los usuarios, de distintas geografías y bajo diferentes idiomas, comparten e intercambian información. Con el tiempo el número de usuarios, así como el volumen de información existente, se incrementa considerablemente dificultando el análisis de la misma. La facilidad del acceso a Internet y la evolución de la dinámica de creación de contenidos provocan que cada vez más personas para expresar sus ideas recurran a la web para expresar sus ideas.

La web, el espacio más visible de Internet, evolucionó de una plataforma fundamentalmente informativa donde los usuarios consumen la información generada por los suministradores (instituciones, agencias de noticias, etc.) a un entorno colaborativo donde los propios usuarios producen y comparten la información. Este modelo colaborativo es conocido como la web 2.0. Este modo de interactuar ha crecido aceleradamente en los últimos años debido a la cantidad de información producida. Cualquier individuo con acceso a la red es un potencial generador de información.

Una de las expresiones características de la web 2.0 son las redes sociales. En estos espacios de comunicación los usuarios expresan sus experiencias u opiniones personales sobre diversos temas, siendo ellos mismos los generadores de contenido y gestores de la información (Bollen, et al., 2011). Dentro de las redes sociales, el microblogging ha tenido un espacio preponderante propulsado por el auge de los dispositivos móviles. Esta forma de comunicación es más sencilla de producir, distribuir y consumir. El microblogging se basa en la difusión o publicación de textos breves (usualmente sobre los 140 caracteres) que pueden ser distribuidos, generados y consumidos desde dispositivos móviles, emails, mensajería instantánea o directamente en la web. Este fenómeno es muy utilizado por los usuarios para expresar una opinión, informar de una noticia o hacer referencia a otros sitios de la web (Bollen, et al., 2011).

Entre las redes sociales más conocidas se encuentra Twitter (http://twitter.com), fundada en el año 2006, que introduce en Internet el concepto de microblogging (Álvarez, 2010). Twitter cuenta con millones de usuarios de todo el planeta, que comparten información en forma de mensajes cortos denominados tweets, en múltiples idiomas. Esta red social ha cobrado tal importancia que se realizan continuamente estudios y desarrollos para analizar la información proveniente de ella.

Muchos de estos estudios y desarrollos tienen al texto como elemento primario de análisis. Generalmente, estos procesos de análisis son dependientes del idioma particular de los textos que se analizan. Es así que en muchos desarrollos es necesario determinar primero el idioma original del tweet. La incorrecta determinación del lenguaje puede provocar errores en la posterior interpretación del contenido.

A pesar del que el problema de identificación de idioma se considera generalmente resuelto (MacNamee, 2005), existen contextos particulares que dificultan considerablemente esta tarea. Los tweets se caracterizan por tener textos muy cortos, una elevada presencia de ruido (etiquetas propias, enlaces, etc.), mezcla de idiomas, uso de expresiones de jerga y emoticones, y fuertes modificaciones de la ortografía. Todos estos factores influyen negativamente en la calidad de los algoritmos de detección de idiomas de propósito general (Lui, et al., 2014; Baldwin y Lui, 2010, Hughes et al., 2006). Entonces es necesario diseñar estrategias específicamente concebidas para lidiar con las características particulares de Twitter en la solución del problema de identificación de idioma.

Este trabajo presenta una propuesta compuesta a partir de métodos clásicos en la identificación de idioma (n-gramas (Cavnar, et al., 1994) y small words (Johnson, 1993)). Se analiza la precisión de cada método por separado así como dos alternativas diferentes de composición. Igualmente, se estudia la incidencia en la precisión de la identificación de idioma dedos posibles preprocesamientos que permitan eliminar el posible ruido en los tweets.

## DETECCIÓN DE IDIOMA

El problema de detección de idioma fue formalizado inicialmente por Gold (Gold, 1967). Varios investigadores han trabajado desde entonces con diferentes puntos de aproximación. Se han utilizado métodos deterministas (n-gramas (Cavnar, et al., 1994), small words (Johnson, 1993)) y métodos de aprendizaje (e.g. Naive Bayes (Gottron, et al., 2010)). La detección de idioma tiene dos vertientes fundamentales: textos monolingües (Lui, et al., 2011) y textos multilingües (Lui, et al., 2014). Durante los años 90 se enfoca como un problema de clasificación, donde el idioma determina la categoría. La formulación formal para textos monolingües, cómo los tweets que son esencialmente monolingües, consiste en asignar una categoría c_i de un conjunto de lenguajes C a un documento d_ide un conjunto de documentos D.

En la identificación de idioma de textos cortos los métodos tradicionales por si solos aún no resuelven totalmente el problema debido, en gran medida, a la falta de información o el posible ruido existente en la misma (Baldwin y Lui, 2010, Hughes et al., 2006).

En particular, los mensajes de Twitter, escritos en una gran cantidad de idiomas, puede considerarse que unitariamente se escriben en un único idioma aunque a menudo se utilizan expresiones de jerga o palabras importadas de otro idioma. Asimismo, el texto de los tweets también está sujeto a fuertes modificaciones de la ortografía que dificultan su clasificación (Bergsma, et al., 2012).

Trabajos previos en la identificación de idioma en Twitter (Lui et al., 2014; Carter et al., 2013; Tromp y Pechenizkiy, 2011) se basan en conjuntos de entrenamiento y dependen, para su rendimiento, del tamaño de estos conjuntos. Igualmente, en la eficiencia de estos métodos incide la longitud de los tweets e información agredada a partir de datos de los usuarios o del propio contenido de los tweets. Asimismo, las particularidades de los usuarios y sus propias características comunitarias pueden incidir negativamente en la utilización de algoritmos de aprendizaje.

Los métodos que se basan en el cálculo de atributos del idioma, como los n-gramas (Cavnar, et al., 1994) o las small words (Johnson, 1993) no son dependientes de conjuntos de entrenamientos o información agregada. Sin embargo, la brevedad de los textos si puede incidir directamente en el rendimiento de estas aproximaciones. La utilización compuesta de estos métodos podría elevar el rendimiento en la identificación del idioma de los tweets sin que sea necesaria la utilización de utilización de entrenamientos previos o información agregada, disminuyendo así la complejidad de la clasificación en tanto no es necesario obtener esos datos y se evitan problemas de información local que puedan distorsionar la clasificación misma.

## METODOLOGÍA

La aproximación que se propone al problema de detección de idioma en Twitter se basa en los métodos de trigramas (Schmitt, 1991) y small words (Johnson, 1993). Estos métodos son de sencilla implementación y no requieren de conjuntos de entrenamiento para su funcionamiento. En su forma estándar estos algoritmos por sí solos no pueden lidiar efectivamente con mensajes muy cortos (Los tweets tienen un máximo de 140 caracteres), por lo que se proponen estrategias compuestas, basadas en una combinación de ambos métodos.

Debido a la elevada presencia de ruido que regularmente contienen los tweets, se analiza el impacto que provoca una normalización previa de los mensajes en la efectividad de la clasificación. Esta normalización consiste en pasos de preprocesamiento que intentan eliminar las fuentes fundamentales de ruido presentes en los mensajes: etiquetas de Twitter, URLs, emoticones y jerga.

## MÉTODO DE LOS TRIGRAMAS

Un trigrama es un n-grama con n=3, es decir es una subcadena del texto de longitud 3, incluyendo los espacios entre palabras (Golding, et al., 1996). Por ejemplo, la frase Hola mundo contiene los trigramas: Hol, ola, la_, a_m, _mu, mun, und y ndo.

El algoritmo de identificación de idioma basado en trigramas consiste en contar para un documento la fracción de trigramas de cada idioma que aparecen en el documento. Se asigna al documento el idioma que maximice este valor (Schmitt, 1991). Una variante consiste en asignar a cada trigrama un peso distinto en función de su frecuencia de aparición en el idioma correspondiente, y usar una suma ponderada para clasificar un documento.

La técnica de los trigramas presenta cierta desventaja para su uso en textos cortos. Debido a que el espacio de los posibles trigramas es pequeño, es muy probable que incluso en textos cortos se encuentren trigramas asociados a uno o más idiomas. Sin embargo, para idiomas que tienen una raíz común, es posible que los trigramas más frecuentes se encuentren compartidos en ambos idiomas, lo que dificulta la clasificación. Una forma de aliviar este fenómeno consiste en asociar a los trigramas un peso no solo basado en la frecuencia de aparición en un idioma, sino además penalizando la aparición en más de un idioma.

## MÉTODO DE *SMALL WORDS*

Un *small word* es toda palabra con menos de 5 caracteres que no contiene números ni signos de puntuación. Los artículos, las preposiciones, las conjunciones y algunos adverbios son los predominantes en la lista de los small words. Al igual que para los trigramas, la clasificación basada en small words (Johnson, 1993) consiste en asignar a un documento el idioma que maximice la fracción de small words contenida en el mismo. Se aplican las mismas ideas sobre la asignación de pesos a los small words.

Esta técnica es más efectiva en textos más largos, ya que a medida que aumentan la cantidad de palabras, aumenta la probabilidad de encontrar small words. En textos cortos es posible que no se encuentre ninguno, sobre todo en mensajes con grandes

deformaciones gramaticales, como los tweets, donde se suele prescindir de los artículos y preposiciones en virtud de la limitada longitud de los mensajes. Sin embargo, cuando aparecen, los small words ofrecen una información más confiable que los trigramas, debido a que consisten en palabras completas, y que generalmente son muy características de un idioma particular (e.g. el artículo the en idioma inglés).

## COMPOSICIÓN

Como se ha visto, ambos métodos, trigramas y small words, presentan características ventajosas y desventajosas para lidiar con textos cortos y ruidosos. Una propuesta que utilice ambos método podría aumentar la calidad de los resultados. Así se decidió valorar una propuesta compuesta entre los métodos de trigramas y small words.

Para estudiar la combinación de los resultados de ambos métodos, se propuso utilizar el promedio y el máximo de ambos resultados, respectivamente. La variante de promediar los resultados de ambos métodos se basa en la hipótesis de que un consenso entre los dos métodos puede suplir las deficiencias que tengan por separado. La variante de aplicar el máximo se basa en la hipótesis de que ambos métodos mejoran su precisión mientras mayor sea la cantidad de trigramas o small words presente, por lo que cuando un método da un valor mayor, generalmente se debe a que tiene más información. De esta forma se le da mayor peso al método que más seguro se encuentre.

## EXPERIMENTACIÓN

Para utilizar los métodos descritos en las secciones anteriores se confeccionaron 2 corpus, uno compuesto por un conjunto de small words y el otro por un conjunto de trigramas, en cada caso los más utilizados por idiomas.

En la construcción de estos corpus se utilizó el conjunto de documentos representativos de un idioma de la biblioteca NLTK (Bird, 2006). Cada idioma cuenta con 10 documentos representativos, que en total contienen alrededor de 30 millones de palabras. Fueron creados corpus, tanto para trigramas como small words, para 11 idiomas diferentes (Se crearon para los idiomas danés, alemán, griego, inglés, español, finés, francés, italiano, holandés, portugués y sueco).

El conjunto de tweets utilizado para establecer la precisión de los métodos se confeccionó a partir de una muestra de extraída de Twitter en octubre de 2012. Los mensajes se dividieron previamente a partir del idioma asociado al perfil del usuario que escribió cada mensaje.

Luego los mensajes fueron depurados por un especialista para determinar el idioma real. Finalmente, se creó un conjunto de 1529 tweets clasificados por idioma, con una media de 76 caracteres por mensaje. La tabla 1 muestra la composición del corpus.

**Tabla 1.** Cantidad de documentos (tweets) en el corpus por idioma.

| Idioma | Tweets | Media de caracteres |
|---|---|---|
| Español | 620 | 74 |
| Inglés | 518 | 73 |
| Alemán | 169 | 89 |
| Francés | 197 | 78 |
| Griego | 25 | 79 |
| Total | 1529 | 76 |

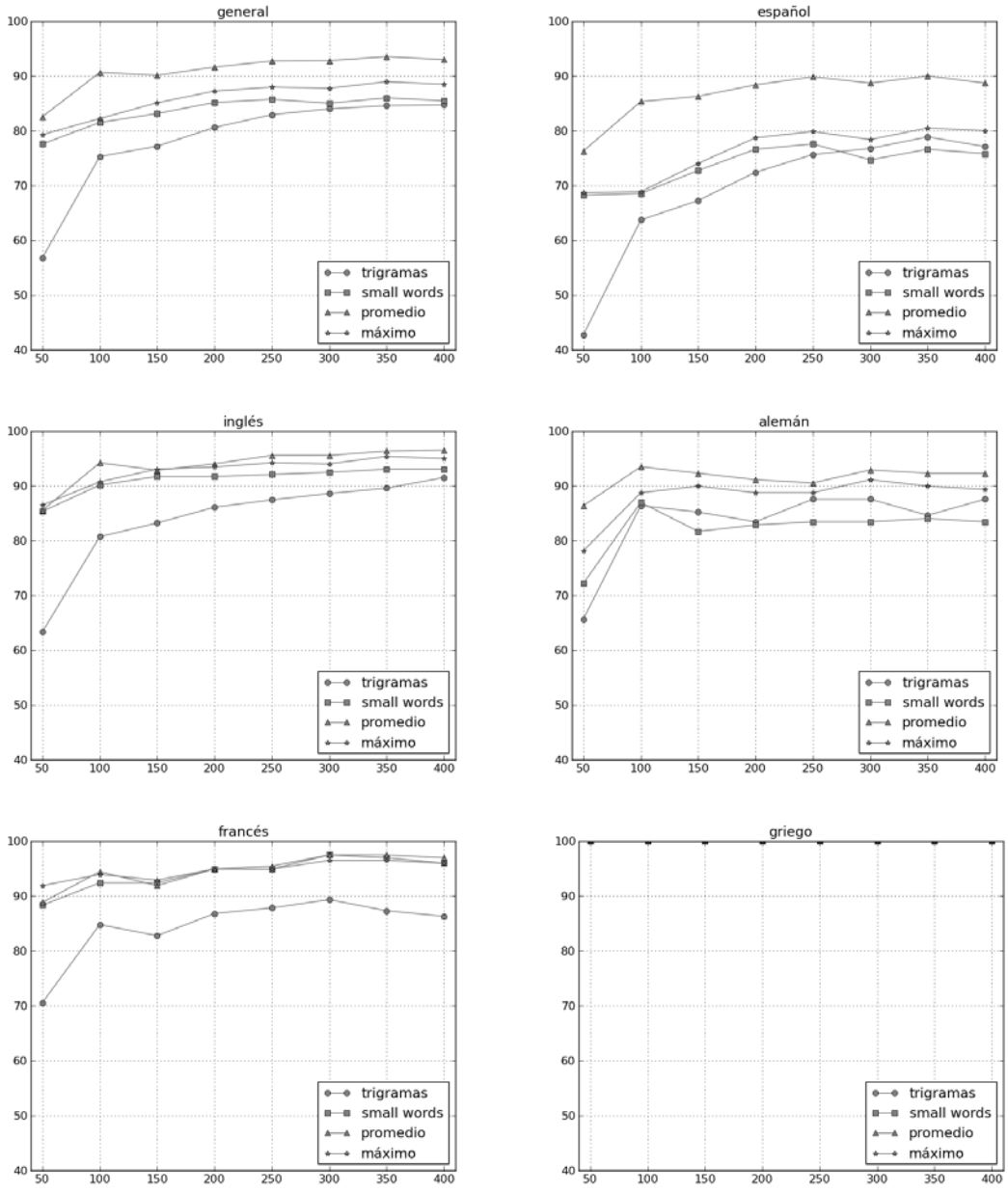
Para analizar la influencia del ruido en la efectividad de los algoritmos se realiza un proceso de normalización que elimina algunas de las fuentes más comunes de ruido, comparando la precisión de cada método sobre los mensajes originales, y sobre los mensajes normalizados.

Se consideran como fuentes de ruido las etiquetas de Twitter que sirven para identificar usuarios y tópicos (hashtags) así como los signos de puntuación. Además, también fueron tenidas en cuenta las URLs embebidas en los mensajes, que pueden influir negativamente en la clasificación, al estar conformadas por palabras de un idioma distinto al idioma del tweet. Un caso particular lo constituyen las etiquetas de tópico (hashtags), que aunque a menudo son consideradas como ruido, pueden influir positivamente en la determinación del idioma, debido a que se constituyen generalmente de palabras propias del idioma en cuestión (e.g. #CopaMundial). Por este motivo se realiza además una comparación eliminando todas las etiquetas mencionadas, excepto los hashtags.

Así se realizaron experimentos que permitieran clasificar el conjunto de tweets en 11 posibles idiomas diferentes. Estos experimentos se realizaron para corpus de trigramas y small words de diferentes tamaños (desde tamaño 50 hasta 400). Igualmente, los experimentos fueron repetidos para el texto normal y para las dos variantes de normalización de los textos.

Como resultado de la experimentación, y según recoge la literatura (Cavnar, et al., 1994), se pudo observar que la calidad de la precisión comienza a aumentar hasta una vecindad de tamaño 300 de los corpus de trigramas y small words. La figura 1 muestra la evolución según el tamaño de los corpus de la precisión de la identificación para los tweets sin normalizar.

**Figura 1.** Evolución de la precisión (en %) según el tamaño del corpus en tweets sin procesar.




De manera general, para los corpus de 350 términos se obtuvieron los mejores valores de precisión que son presentados en las tablas 2, 3 y 4. En la tabla 2 se muestra la precisión alcanzada por los dos métodos originales y los dos métodos compuestos sobre los mensajes originales sin normalizar. La tabla 3 muestra la precisión cuando se emplea una fase de normalización que elimina todas las etiquetas de Twitter. Finalmente, la tabla 4 muestra la precisión cuando en la normalización se excluyen los hashtags.

**Tabla 2.** Precisión (en %) con los mensajes sin procesar, con corpus de 350 términos.

| Idioma | Trigramas | *Small Words* | Promedio | Máximo |
|---|---|---|---|---|
| General | 84.57 | 86.00 | **93.53** | 88.95 |
| Español | 78.87 | 76.61 | **90.00** | 80.48 |
| Inglés | 89.58 | 93.05 | **96.33** | 95.37 |
| Alemán | 84.62 | 84.02 | **92.31** | 89.94 |
| Francés | 87.31 | 96.95 | **97.46** | 96.45 |
| Griego | 100.0 | 100.0 | **100.0** | 100.0 |

**Tabla 3.** Precisión (en %) eliminando las etiquetas propias de Twitter, con corpus de 350 términos.

| Idioma | Trigramas | *Small Words* | Promedio | Máximo |
|---|---|---|---|---|
| General | 85.05 | 85.51 | **93.11** | 89.25 |
| Español | 78.16 | 76.86 | **89.97** | 81.55 |
| Inglés | 89.56 | 90.91 | **94.39** | 94.20 |
| Alemán | 87.50 | 85.12 | **94.05** | 91.07 |
| Francés | 90.86 | 96.95 | **97.97** | 97.46 |
| Griego | 100.0 | 100.0 | **100.0** | 100.0 |

**Tabla 4.** Precisión (en %) eliminando las etiquetas, excepto hashtags, con corpus de 350 términos.

| Idioma | Trigramas | *Small Words* | Promedio | Máximo |
|---|---|---|---|---|
| General | 85.18 | 85.44 | **93.25** | 89.11 |
| Español | 77.99 | 76.86 | **89.81** | 81.55 |
| Inglés | 90.33 | 90.91 | **94.97** | 94.20 |
| Alemán | 86.90 | 84.52 | **94.05** | 89.88 |
| Francés | 90.86 | 96.95 | **97.97** | 97.46 |
| Griego | 100.0 | 100.0 | **100.0** | 100.0 |

También, durante el proceso de experimentación se determinó una medida de la clasificación errónea por idioma. Del mismo modo que ocurrió con la precisión de la clasificación, los mejores valores se alcanzaron para corpus de trigramas y small words de 350 términos. La tabla 5 muestra la medida de clasificación errónea alcanzada por los dos métodos originales y los dos métodos compuestos sobre los mensajes originales sin normalizar.

**Tabla 5.** Clasificación errónea (en %) con los mensajes sin procesar, con corpus de 350 términos.

| **Idioma** | **Trigramas** | ***Small Words*** | **Promedio** | **Máximo** |
|---|---|---|---|---|
| Español | 0.77 | 0.11 | **0.11** | 0.11 |
| Inglés | 2.08 | 0.99 | **0.69** | 1.29 |
| Alemán | 0.81 | **0.37** | 0.66 | 0.66 |
| Francés | 2.33 | 4.50 | **0.83** | 3.83 |
| Griego | 0.00 | 4.85 | **0.00** | 0.60 |

## DISCUSIÓN

Los resultados de los experimentos muestran un aumento de la precisión al aumentar el número de trigramas o *small words* tenidos en cuenta, teniendo un punto máximo a partir de 350 términos. Esa misma tendencia se observa para la clasificación errónea donde esta va disminuyendo paulatinamente.

Se pudo comprobar que la composición de los métodos, tanto trigramas como *small words*, permiten una mejor identificación del idioma de los *tweets* que los métodos por separado. De los métodos compuestos utilizados el que mejor precisión ofrece es el del promedio. En ello incide, sobre todo, que siendo el texto muy breve, con una media de 76 caracteres de manera general, el número de *small words* en el texto es pequeño y la utilización de alguna de ellas en un idioma diferente al del texto puede llevar a un erróneo valor elevado de pertenencia a cierto idioma.

Fue alcanzada una mayor precisión general, para la variante del promedio con corpus de 350 términos, de un 93.53% y por idiomas valores siempre superior o iguales a un 90%. Para la misma variante, la medida de clasificación errónea fue siempre, en cada idioma, inferior al 1%.

La mayor precisión general fue alcanzada para *tweets* sin normalizar y bajo la variante compuesta del promedio aunque los valores para los distintos procesamientos son muy cercanos. Sin embargo, este comportamiento no es homogéneo ni para los otros métodos (trigramas, *small words* y máximos), ni por idiomas.

Este resultado hace necesario que se estudie con mayor detenimiento el procesamiento a realizar pues se evidencia que existe un conjunto de etiquetas propias que contribuyen a la clasificación del idioma bajo estos métodos. Sin embargo, por otra parte, se encontraron documentos particulares donde los emoticones, las letras repetidas y la jerga son causantes de que el *tweet* fuera mal clasificado. Elementos estos que no fueron considerados en las variantes de procesamiento.

Por idiomas el que mejor precisión de clasificación obtuvo fue el griego con un 100%. Esta exactitud se explica por el alfabeto diferente que utiliza dicho idioma y los métodos analizados al basarse en análisis de caracteres logran una total discriminación en este sentido.

El idioma que más problemas presentó fue el español. En ello incide la utilización en términos en otros idiomas, dado sobre todo por los patrones migratorios de los hispano hablantes, así como la cercanía idiomática con idiomas romances utilizados para clasificar, como por ejemplo el portugués.

## CONCLUSIONES

Este trabajo presenta una alternativa para determinar idiomas en *Twitter* sin que sea necesario utilizar conjuntos de entrenamiento o información agregada obtenida ya sea de los *tweets* o los usuarios. En dicha alternativa se utilizan técnicas sencillas y simples de implementar, basadas en los algoritmos de reconocimiento de *small words* y trigramas. Igualmente, se corroboró que la utilización de variantes de composición de estos algoritmos ofrece resultados superiores a la aplicación de los algoritmos por separado.

Los mejores resultados, acorde a la precisión en la clasificación del idioma, se alcanzaron para un tamaño del corpus de 350 *small words* y trigramas. La creación de corpus en 11 idiomas diferentes permitió la clasificación de *tweets* en ese mismo número de lenguas.

La experimentación en la clasificación de un conjunto de 1529 *tweets* de 5 idiomas diferentes de 11 posibles ofreció una mejor precisión general de 93.53% para corpus de 350 trigramas y *small words* en mensajes sin normalizar y con método compuesto basado en el promedio. En esta misma variante, la medida de clasificación errónea para cada idioma del conjunto de *tweets* fue siempre inferior al 0.9%. La precisión por idioma fue igual o superior al 90%.

Así, considerando esta variante como la alternativa que se propone, se muestra que el método es poco sensible a las variantes de preprocesamiento analizadas y que resulta efectivo para la identificación de idiomas en *Twitter*.

## REFERENCIAS

Álvarez, R. M. (2010). Análisis de opiniones en Internet a partir de la red social Twitter [Report]. - [s.l.] : Anales de Mecánica y Electricidad.

Baldwin T. and Lui M. (2010). Language identification: The long and the short of the matter [Conference] //In Proc. HLT-NAACL, pages 229–237.

Bird, S. (2006). NLTK: the natural language toolkit [Conference]. Association for Computational Linguistics. Proceedings of the COLING/ACL on Interactive presentation sessions, pp. 69-72.

Bollen, J., Mao H. and Pepe A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. [Conference] // ICWSM.

Carter, S., Weerkamp Wouter, and Tsagkias Manos (2013). Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic Text. *Language Resources and Evaluation Journal* (forthcoming).

Cavnar W. B. and Trenkle J. (1994). M N-gram-based text categorization. Ann Arbor MI. - [s.l.] : Citeseer, 1994. - 2 : Vol. 48113, pp. 161--175.

Gold E. (1967). Mark Language identification in the limit. Information and control. - [s.l.] : Elsevier, 1967. - 5 : Vol. 10: 447-474.

Golding, A. R. and Schabes, Y. (1996). Combining trigram-based and feature-based methods for context-sensitive spelling correction. Proceedings of the 34th annual meeting on Association for Computational Linguistics. pp. 71--78.

Gottron, T. and Lipka N. (2010). A comparison of language identification approaches on short, query-style texts. Advances in information retrieval. Springer.

Hughes, B., Baldwin, T., Bird, S., Nicholson, J., and Mackinlay, A. (2006). Reconsidering language identification for written language resources [Conference] // In Proc. LREC, pages 485–488.

Johnson, S. (1993). Solving the problem of language recognition [Report] / Technical report, School of Computer Studies, University of Leeds.

Lui, M. and Baldwin, T. (2011). Cross-domain feature selection for language identification. In Proceedings of 5th International Joint Conference on Natural Language Processing.

Lui, M., Lau, J. H. and Baldwin, T. (2014). Automatic detection and language identification of multilingual documents [Journal] // Transactions of the Association for Computational Linguistics.

McNamee, P. (2005). Language identification: a solved problem suitable for undergraduate instruction. *Comput. Sci. Coll.*, 20(3):94–101.

Schmitt, J. C. (1991). Trigram-based method of language identification. Google Patents, US Patent 5,062,143.

Tromp, E. and Pechenizkiy, M. (2011). Graph-based n-gram language identication on short texts, In Proc. 20th Machine Learning conference of Belgium and The Netherlands, pages 27–34.