

TRABAJO DE FIN DE MÁSTER

«ANÁLISIS … TWITTER… RAE… TELEFÓNICA»

Autores: JOSE MARÍA HERNÁNDEZ DE LA CRUZ

BÁRBARA SAIZ …

Tutores: ADRIÁN RIESCO …

RAFAEL ….

MÁSTER EN LETRAS DIGITALES

Curso académico 2020-2021

Fecha de presentación XX/XX/2021



FACULTAD DE FILOLOGÍA Y FACULTAD DE INFORMÁTICA

TABLA DE CONTENIDOS

[INTRODUCCIÓN 4](#_Toc67398118)

[1. Objetivos 6](#_Toc67398119)

[2. Motivaciones 6](#_Toc67398120)

[3. Metodología 6](#_Toc67398121)

[4. Agradecimientos 6](#_Toc67398122)

[5. Nuestra contribución 6](#_Toc67398123)

[6. Organización del studio 6](#_Toc67398124)

[7. Reservas 6](#_Toc67398125)

[ESTADO DE LA CUESTIÓN 8](#_Toc67398126)

[I. LAS INSTITUCIONES EN EL MEDIO DIGITAL 8](#_Toc67398127)

[1. La RAE en el medio digital 8](#_Toc67398128)

[2. Twitter 15](#_Toc67398129)

[II. PROCESAMIENTO DEL LENGUAJE NATURAL (NLP) 16](#_Toc67398130)

[1. ¿Qué es? 16](#_Toc67398131)

[2. Historia 16](#_Toc67398132)

[3. Aplicaciones actuales 16](#_Toc67398133)

[DESARROLLO 17](#_Toc67398134)

[I. LDA 17](#_Toc67398135)

[1. Apartado 1 17](#_Toc67398136)

[2. Apartado 2 17](#_Toc67398137)

[II. LSA 17](#_Toc67398138)

[3. Apartado 1 17](#_Toc67398139)

[4. Apartado 2 17](#_Toc67398140)

[III. NUESTRO MÉTODO 17](#_Toc67398141)

[5. Apartado 1 18](#_Toc67398142)

[6. Apartado 2 18](#_Toc67398143)

[CONTRIBUCIONES 19](#_Toc67398144)

[CONCLUSIÓN 20](#_Toc67398145)

[BIBLIOGRAFÍA 21](#_Toc67398146)

**Resumen:**

**Palabras clave:**

**Abstract**: This Tfm….

**Keywords**: RAE; Topic Modeling; Twitter; LDA; LSA;

INTRODUCCIÓN

La llegada de las redes sociales a nuestras vidas a finales de la primera década del presente siglo, y su evolución hasta límites insospechados a lo largo del decenio posterior, ha supuesto un cambio carente de precedentes en la manera en las que nos comunicamos. Cuando McLuhan acuñó su famoso término “aldea global”, no podía imaginar la creación de unas herramientas que conectaran de una manera tan abrumadora como las redes sociales de hoy en día. Las redes sociales se dividen en diversos tipos, según el contenido que permiten intercambiar y el sesgo de edad al que van dirigidas: Facebook abarca todas las edades y permite una amplia variedad de soportes (textos, imágenes, videos), mientras que otras como Instagram se ciñen a contenidos de imagen para jóvenes, y para aún más jóvenes está pensada tiktok, por ejemplo. Todas estas son un buen reflejo de la virtualización de la vida humana, publicar contenido en estas redes es algo cotidiano y sin lo que muchos verían su vida un poco más vacía. Sin embargo, Twitter se erige como una red social diferente, cuya principal peculiaridad es la brevedad obligatoria de los mensajes publicados. Twitter es, además, la red social donde las instituciones tienen mayor seguimiento. Las cuentas de la policía, los representantes políticos, los principales referentes deportivos y artísticos hacen uso de esta red social de manera más efectiva de lo que lo pudiera hacer con otra.  Tal es la magnitud que ostentan, que no hay un ámbito de la vida que no tenga su reflejo en las redes sociales (política, deportes, medios de comunicación, grandes corporaciones…) Las instituciones creadas en el siglo XVIII para preservar la historia, la lengua, las artes de las principales metrópolis del globo tampoco han dejado pasar la oportunidad de subirse al carro de la digitalización. En este contexto encontramos la cuenta de la Real Academia Española, órgano que controla el uso del español en el mundo. La cuenta de la rae cuenta, en el momento que escribimos estas líneas con más de un millón ochocientos mil seguidores, lo que implica que cada vez que la institución hace uso de esta vía para transmitir un mensaje, este llega a un mínimo de 1.8 millones de cuentas, un número que se verá multiplicado con la difusión extra que pueda tener cada uno de esos tuits. La cuenta oficial de la RAE ofrece la posibilidad de resolver dudas a los usuarios de esta red social. Aquellos que se cuestionen algún aspecto del uso del español pueden hacérsela llegar a través de esta red social. Sabiendo el elevado seguimiento que tiene la cuenta por parte de usuarios de todas partes del mundo, y que pueden emplear un español distinto al de otras regiones, así como la multiplicidad de campos en los que pueden surgir dudas en cuanto al uso del español, es previsible que los mensajes recibidos por esta cuenta sean dispares y muy numerosos. Es en esta amplitud donde se hace necesaria una herramienta que ayude, al menos, a clasificar y cuantificar los tweets recibidos. Así, desde la institución, y con el apoyo de Telefónica, se nos pidió elaborar una herramienta que ayudara a tal menester. De esta manera hemos elaborado un Trabajo de Fin de Máster. Para la realización de este proyecto hemos tomado dos caminos distintos: uno de corte académico, no olvidando el ámbito académico al que no ceñimos; y otro más práctico, en el que hemos tratado de programar una herramienta que ayude en la clasificación y cuantificación de datos tanto a la Real Academia Española como a Telefónica. Para el primero de los itinerarios hemos investigado sobre estudios en redes sociales que se han llevado a cabo en años anteriores, prestando atención a los que se centraban en el análisis de aspectos que tuvieran que ver con la red social Twitter. Además, hemos tenido en cuenta las herramientas de análisis masivo de textos, especialmente las enfocadas a textos cortos, como lo son los tweets que nos ocupan. Para la elaboración de la herramienta, concluimos que usar el lenguaje de programación Python sería la opción más adecuada, debido al uso extendido de dicho lenguaje y ……Como punto de inicio para la parte práctica del proyecto, tomamos unos treinta y dos mil tweets publicados entre diciembre de 2020 y febrero de 2021, todos ellos dirigidos al perfil de la RAE en Twitter (@RAEinforma). A grandes rasgos, nuestro trabajo se centró en sus estadios iniciales en limpiar y clarificar cuáles eran los tipos de dudas más usuales entre los usuarios que exponían sus dudas. Es pasos posteriores, buscamos poder clasificar cualquier nuevo tweet añadido a alguna de las categorías previamente creadas.

## Objetivos

Bla bla bla…

## Motivaciones

Bla bla bla…

## Metodología

Bla bla bla…

## Agradecimientos

Bla bla bla…

## Nuestra contribución

Bla bla bla…

## Organización del studio

Bla bla bla…

## Reservas

Bla bla bla…

ESTADO DE LA CUESTIÓN

# LAS INSTITUCIONES EN EL MEDIO DIGITAL

## La RAE en el medio digital

La RAE es la institución que regula el uso del español en el mundo.

los ideales de la institución española se han instalado como pauta de referencia normativa en ámbitos diversos (Rizzo 426)

A principios del siglo XXI, acompañando los tiempos de la globalización y en el marco de una política orientada al desarrollo del estatuto simbólico de la lengua española (del Valle, 2007) impulsada por el Estado español, la RAE inició una nueva etapa, (Rizzo 426)

En los últimos años, los avances en las tecnologías de la información y de la comunicación favorecieron la inclusión de herramientas normativas en el mundo digital, afectando al modo en el que los hispanohablantes se relacionan con la *autoridad* reguladora de su lengua. Más allá de las plataformas institucionalmente respaladas, surgen otras plataformas no institucionales con el propósito de crear colaborativamente diccionarios en linea—como *Wordreference—* sirviendo estos como espacios de resolución de dudas “descentrados de normatividad.” (Rizzo 426)

La apertura del perfil de la RAE en Twitter se enmarca dentro De una serie acciones que, como parte de una política de Estado financiada por un conjunto de empresas privadas y fundaciones, generando un “despliegue de una enérgica política de área idiomática que se designó oficialmente como «panhispánica»” (Rizzo 428). La entrada de la RAE en esta plataforma no supuso de manera única un cambio de soporte, ya que Twitter encorseta el contenido a una dinámica propia que “plantea un modo distinto de vinculación entre la voz institucional, los usuarios de la lengua y el saber lingüístico” (Rizzo 427).

El nuevo proyecto planteaba, por un lado, el desplazamiento de un modelo normativo monocéntrico —anclado en la variedad culta de Castilla— por uno pretendidamente pluricéntrico a partir de la incorporación de aspectos de la variación lingüística desplegada a ambos lados del Atlántico

dible la actuación de agentes tradicionales de regulación idiomática —en especial, de la RAE como entidad de más larga trayectoria en el ámbito hispánico— en estos nuevos contextos, principalmente en las redes sociales, dada la relevancia creciente de estos ámbitos como espacios de intervención normativa en el presente. De este modo, en 2012 la RAE incorporó al tradicional servicio de consultas lingüísticas, que presta desde hace muchos años, un nuevo canal de comunicación: Twitter. Ahora bien, la actuación en esta red social no significó un simple cambio de soporte o formato ya que se trata de una plataforma novedosa con una dinámica propia que plantea un modo distinto de vinculación entre la voz institucional, los usuarios de la lengua y el saber lingüístico. (Rizzo 427)

En la misma dirección se sitúa el aumento de su participación en eventos con repercusión pública, en medios de comunicación y en Internet; en este último caso, mediante la elaboración y actualización de la página electrónica oficial de la RAE y la ASALE, la digitalización de archivos, el acceso libre a recursos lingüísticos y la apertura de cuentas en las principales redes sociales. (Rizzo 429)

Finalmente, han sido objeto de estudios recientes nuevos espacios de consulta sobre usos normativos del español en contextos digitales no institucionales, especialmente, los foros de discusión en diccionarios *on line* que producen un efecto de descentralización normativa —es decir, de apartamiento de una norma única, monolítica y estable— (Rizzo 429)

la intervención de la RAE en Internet y en las redes sociales, además de contribuir en la construcción de una imagen renovada, más «cercana» a los hablantes, que desde hacía varios años impulsaba la institución. El espacio de normatividad que analizamos, esto es, la sección de consultas lingüísticas de la RAE en Twitter, se inscribe en un marco institucional; en este sentido, comparte ciertos rasgos con los instrumentos lingüísticos normativos, en parte, porque son las fuentes a partir de las cuales se construyen los enunciados que buscan proporcionar una recomendación lingüística y dar respuesta a las consultas realizadas. A esto se agrega una nueva dimensión: la interacción entre la academia y los usuarios de la lengua en un contexto digital y el carácter público de estos intercambios, lo que significa que cualquier sujeto puede leer y comentar enunciados producidos por una institución que, en otro momento, parecía inaccesible. Al mismo tiempo, la incorporación de la RAE a la plataforma Twitter supone la posibilidad de llegar a una cantidad ilimitada de hablantes distribuidos a lo largo de todo el mundo; si bien se trata de un público sumamente heterogéneo desde múltiples puntos de vista —geográfico, social, educativo, etario, etc.— tiene en común —al menos— cierto interés por los asuntos normativos. (Rizzo 430)

La cuenta oficial de la RAE en Twitter fue abierta en 2011 y al año siguiente comenzó con el «servicio de consultas lingüísticas», tal como fue nombrado oficialmente, a cargo del Departamento de *Español al día.* Este equipo de trabajo fue creado en 1998 con el propósito de modernizar la atención de consultas lingüísticas a través de medios electrónicos, que hasta esa fecha se había realizado exclusivamente por correo postal o por fax. (Rizzo 431)

En octubre de 2012, la cuenta de la RAE en Twitter comenzó la actividad referida a recomendaciones y consultas lingüísticas, identificadas desde un comienzo con la etiqueta *#RAEconsultas.* Los mensajes producidos en una primera etapa provienen mayormente de la institución, quien realiza alrededor de una decena de publicaciones por día, mientras que las preguntas de los usuarios no superan esta cantidad, de modo que la función de la sección se inclina más hacia la difusión de normas que hacia la resolución de dudas. En efecto, la interacción con los usuarios es mínima, en comparación con lo que sucede en la actualidad. Ante consultas específicas, en muchas ocasiones el usuario de la institución española —@RAEinforma— en lugar de brindar la información requerida, sugiere redirigir la consulta al Departamento de *Español al Día* a través de un formulario en la página web de la RAE (Rizzo 431)

Como se puede observar, en una primera etapa, el servicio que presta la RAE a través de su cuenta de Twitter se limita a difundir tópicos normativos que remiten a información adicional, o bien reglas dogmáticas sobre cuestiones —casi en su totalidad— ortográficas: uso de mayúsculas y minúsculas, prefijos, tildes, puntuación, abreviaturas, escritura de extranjerismos, de expresiones numéricas, homófonos; también algunos temas morfológicos como conjugación de verbos irregulares o formación del femenino de profesiones y cargos. (Rizzo 432)

La importante diferencia entre la cantidad de seguidores y los usuarios a los que sigue da cuenta de la considerable influencia y el gran alcance que tiene la institución en esta red social, de modo que activa la relación uno-a-muchos propia de los medios masivos de comunicación tradicionales (Rizzo 432)

Sin embargo, como veremos, la interacción entre la RAE y los usuarios adopta la dinámica uno-a-uno más cercana a la comunicación personal. (Rizzo 432)

un lugar de prestigio como autoridad normativa en el mundo hispanohablante a partir de la producción de los códigos normativos canónicos —el diccionario, la gramática y la ortografía— y de la formulación de un discurso apoyado en ciertos mecanismos de autorrepresentación (del Valle, 2014; Lauria, 2019) así como en un imaginario de lengua como patrimonio compartido (Rizzo 426)

La cuenta de la ASALE tiene gran actividad en la red social, aunque está muy lejos de acercarse a la RAE en cantidad de seguidores y de tuits publicados. Por lo general, su función se centra en compartir información proveniente de las otras cuentas de Twitter de las academias, también del Instituto Cervantes y de la Fundéu, (Rizzo 433)

Con respecto a la existencia o no de una sección de resolución de dudas lingüísticas, ninguna otra institución académica presta este servicio en Twitter (Rizzo 433)

Estas publicaciones se realizan de tres modos distintos: (a) la mayor parte de las interacciones son iniciadas por los usuarios, quienes acuden a la institución normativa para resolver una duda lingüística; por lo general, se la interpela en tanto discurso autorizado pero también hay algunos casos de provocación o burla; (b) otro grupo de publicaciones, cada vez más esporádicas, consiste en recomendaciones publicadas por la RAE sobre cuestiones normativas que aluden a dudas o preguntas frecuentes; (c) por último, en otras ocasiones, durante un intercambio entre dos o más

2 Datos correspondientes a septiembre de 2017.

*Revista de Investigación Lingüística,* 22 (2019), pp. 425-450 ISSN:1139-1146

433

*El discurso normativo de la RAE en Twitter*

usuarios, se menciona a la corporación (en plataformas digitales la arroba delante del nombre de usuario cumple esta función), es decir, se apela a ella y se la invita a intervenir en el diálogo para resolver alguna duda o disputa lingüística. (Rizzo 433–34)

(Rizzo 434)

(Rizzo 434)

discurso normativo que produce la RAE en Twitter. En líneas generales, este discurso puede ser caracterizado como divulgativo en la medida en que está orientado principalmente a un lector no especializado en temas lingüísticos. (Rizzo 435)

las comunicaciones iniciadas por la RAE se fueron reduciendo cada vez más al tiempo que se multiplicaron las consultas de los usuarios. (Rizzo 435)

predomina el uso de tercera persona y de construcciones impersonales. Es decir, se presenta como discurso autorizado cuya legitimidad se apoya en su enunciación misma. En ciertos casos se genera un contacto más directo mediante el uso de la segunda persona del singular (con *usted* como fórmula de tratamiento): (Rizzo 435)

@RAEinforma se excusa de responder un pedido de información porque no es pertinente para la sección de consultas lingüísticas (Rizzo 436)

los enunciados producidos por la RAE en la actualidad contienen mayor información normativa, es decir, formulan sucintamente la norma o la recomendación; a veces, también se agregan ejemplos o breves explicaciones, siempre ceñidas al límite de caracteres que permite la plataforma; por último se proporcionan, en ocasiones, hipervínculos cuya función es brindar la explicación de la regla (Rizzo 438)

Es decir, se presenta como discurso autorizado cuya legitimidad se apoya en su enunciación misma. En ciertos casos se genera un contacto más directo mediante el uso de la segunda persona del singular (con *usted* como fórmula de tratamiento) (Rizzo 435)

formato de pregunta-respuesta, a partir de consultas lingüísticas que realizan los usuarios a la institución española, con el propósito de mostrar que, si bien en el proceso de adaptación del texto fuente —donde está formulada la norma «original»— al dispositivo digital son inevitables los recortes y la selección textual, los criterios adoptados en los distintos casos no son uniformes. (Rizzo 438)

el discurso normativo formulado en Twitter que es objeto de este trabajo, la voluntad institucional de privilegiar un «español general» acorde con las demandas de expansión de la lengua se enfrenta, de algún modo, a las consultas y dudas lingüísticas que recibe la misma institución como autoridad normativa, producidas por hablantes singulares de diferentes puntos geográficos que requieren información sobre sus repertorios lingüísticos o, por el contrario, sobre formas que no les resultan familiares, propias. Por lo tanto, la difusión de un «español general» a partir de la formulación de normas que puedan alcanzar a un público más amplio, más factible en una comunicación basada en la relación de unoa-muchos, se articula con otro tipo de discurso normativo, donde el vínculo entre los participantes se basa en una dinámica uno-a-uno más cercana a la comunicación personal, que evidencie adecuación a las consultas recibidas, prestando mayor atención a la variación lingüística. (Rizzo 442)

Las discusiones sobre temas normativos, que antes quedaban mayormente dentro de los límites de las instituciones o tomaban estado público en palabras de figuras con notoriedad, últimamente se han abierto también a espacios digitales —institucionales y no institucionales— donde los hablantes desde su condición de usuarios de la lengua, es decir, a partir de las experiencias que resultan de sus prácticas lingüísticas, formulan consultas sobre los usos normativos del español. (Rizzo 445)

las particularidades del dispositivo digital de Twitter —especialmente, la brevedad en la extensión de los mensajes y la dinámica de interacción que propone donde, al tiempo que se responde la duda concreta de un hablante específico, el mensaje es leído por un público sumamente amplio y heterogéneo— condicionan en parte los rasgos del discurso normativo que allí se formula. La limitación de caracteres en los mensajes que se publican en esta red social exige un gran poder de síntesis que inevitablemente conlleva recortes: cierta información se privilegia y otra se suprime. (Rizzo 445)

no son siempre los mismos, de modo que se privilegian, según los casos, unas formas lingüísticas sobre otras. (Rizzo 445)

el resultado es la omisión de formas alternativas válidas que podrían ser consideradas más o menos recomendables pero integran, sin duda, el repertorio lingüístico de gran parte de los hispanohablantes, con lo cual se privilegia la difusión de un «español general» (Rizzo 445)

tendencia a privilegiar ciertas opciones –hegemónicas, de quienes toman mayormente las decisiones en la

*Revista de Investigación Lingüística,* 22 (2019), pp. 425-450 ISSN:1139-1146

445

*El discurso normativo de la RAE en Twitter*

regulación del español para el conjunto del área hispánica– frente a otras –consideradas “periféricas”– (Rizzo 445–46)

la sección de consultas lingüísticas de la RAE en Twitter constituye un espacio no tradicional de (re)formulación de la norma que pone en evidencia el modo como la institución académica busca resolver la tensión entre la tendencia a privilegiar un «español general» que pueda alcanzar a un público lo más amplio y diverso posible —a partir de la presentación de ciertas formas lingüísticas como compartidas por el mundo hispanohablante— y la necesidad de dar respuestas que se ajusten a las consultas o dudas lingüísticas singulares de sujetos específicos pertenecientes a diferentes zonas geográficas que interpelan a la RAE en tanto autoridad normativa (Rizzo 446)

Evitar connotaciones negativas, lenguaje coloquial (Soraya Salazar Puerta, Pablo R., Prieto Dávila 19)

un lugar de prestigio como autoridad normativa en el mundo hispanohablante a partir de la producción de los códigos normativos canónicos —el diccionario, la gramática y la ortografía— y de la formulación de un discurso apoyado en ciertos mecanismos de autorrepresentación (Rizzo 426)

Retuitear y aportar respuestas sobre los tuits de otros usuarios. (Soraya Salazar Puerta, Pablo R., Prieto Dávila 19)

Expandir tus ideas, propiciando que los seguidores hablen de la institución. (Soraya Salazar Puerta, Pablo R., Prieto Dávila 22)

Procurar remitir siempre a la fuente oficial de la información y confirmar la veracidad (Soraya Salazar Puerta, Pablo R., Prieto Dávila 19)

## Twitter

Bla bla bla…

# PROCESAMIENTO DEL LENGUAJE NATURAL (NLP)

Topic modeling is a text mining technique which provides methods for identifying co-occurring keywords to summarize large collections of textual information. It helps in discovering hidden topics in the document, annotate the documents with these topics, and organize a large amount of unstructured data.

Text classification is a supervised machine learning problem, where a text document or article classified into a pre-defined set of classes. Topic modeling is the process of discovering groups of co-occurring words in text documents. These group co-occurring related words makes "topics". It is a form of unsupervised learning, so the set of possible topics are unknown. Topic modeling can be used to solve the text classification problem. Topic modeling will identify the topics presents in a document" while text classification classifies the text into a single class.

## ¿Qué es?

## Historia

## Aplicaciones actuales

DESARROLLO

Nadeem Aslam's second novel…

# LDA

The main…

## Apartado 1

When reading the novel, readers have complete access to

## Apartado 2

In the…

# LSA

LSA uses bag of word(BoW) model, which results in a term-document matrix(occurrence of terms in a document). Rows represent terms and columns represent documents. LSA learns latent topics by performing a matrix decomposition on the document-term matrix using Singular value decomposition. LSA is typically used as a dimension reduction or noise reducing technique.

**Topic Coherence**

## Apartado 1

When reading the novel, readers have complete access to

## Apartado 2

In the…

# NUESTRO MÉTODO

The main LSA…

## Apartado 1

When reading the novel, readers have complete access to

## Apartado 2

In the…

CONTRIBUCIONES

CONCLUSIÓN

Previous researches…

BIBLIOGRAFÍA

Obras citadas

Rizzo, María Florencia. “Discurso normativo de la RAE en Twitter.” *Revista de Investigación Lingüística,* vol. 22, 2019, pp. 425–50. ​revistas.um.es​/​ril/​article/​view/​386881/​278181.

Soraya Salazar Puerta, Pablo R., Prieto Dávila. “Gestión y administración de la comunicación institucional en Twitter.” *Disertaciones: Anuario electrónico de estudios en Comunicación Social,,* vol. 8, no. 1, 2015,