Análisis social aplicando técnicas de lenguaje natural a información extraída de Twitter

Social analysis applying natural language techniques to information extracted from twitter.

J. D. Diaz-Mendivelso ; M. J. Suarez-Baron

***Resumen*— Este artículo presenta el desarrollo de un modelo computacional que aplica técnicas de procesamiento de lenguaje natural PLN. Se elaboró revisión completa del estado de arte, encontrando la necesidad de elaborar herramientas y modelos que ayuden al análisis y explotación de la información encontrada en redes sociales. Adicionalmente, se implementó un algoritmo de crawling para extraer tweets y se aplicó PNL identificar la estructura sintáctica, léxica y semántica. El resultado final es un algoritmo que permite identificar patrones y tendencias en términos de I+D cuyo propósito final es ofrecer una alternativa de identificación tendencias y análisis social en centros de investigación y desarrollo tecnológico.**

***Palabras claves*— Análisis social, extracción de información, I+D+i, indexación semántica latente (ISL), procesamiento de lenguaje natural (PLN).**

***Abstract*— This article presents the development of a computational model that applies PLN natural language processing techniques. A complete review of the state of the art was developed, finding the need to develop tools and models that help the analysis and exploitation of the information found in social networks. Additionally, a crawling algorithm was implemented to extract tweets and NLP was applied to identify the syntactic, lexical and semantic structure. The final result is an algorithm that allows identifying patterns and trends in terms of R&D whose final purpose is to offer an alternative for identifying trends and social analysis in research and technological development centers.**

***Index Terms*— Information Extraction social analysis, latent semantic indexing (LSI), natural language processing (NLP), R&D.**

Este manuscrito fue enviado 15 de mayo de 2019 y aceptado el 26 de septiembre de 2019. Este artículo es producto de un proyecto de Maestría en Tecnología Informática de la Universidad Pedagógica y Tecnológica de Colombia - UPTC, titulado: "Sistema computacional para la gestión y análisis sintáctico de información no estructurada extraída de una red social de tipo investigativo".

J. D. Diaz-Mendivelso, Docente investigador, Escuela de ingeniería de sistemas y computación, Universidad Pedagógica y Tecnológica de Colombia, Colombia. Correo electrónico: johan.diaz@uptc.edu.co.

M. J. Suarez-Baron, Profesor Asociado Escuela de Ingeniería de sistemas y computación, Universidad Pedagógica y Tecnológica de Colombia. Correo electrónico: marco.suarez@uptc.edu.co.

I. INTRODUCCIÓN

La evolución de los medios de comunicación ha obligado a las personas y organizaciones a migrar al uso de ambientes tecnológicamente más avanzados; la difusión de la información se venía realizando por medios físicos como centros de discusión o conversaciones persona a persona y el progreso tecnológico ha generado que actualmente se realice en entornos virtuales [1]. Esta nueva tendencia ha convertido el texto en un componente clave para la comunicación en la sociedad, consolidando este método como adecuado para el intercambio de información gracias a su elevada usabilidad digital [2]. Partiendo que la concurrencia de usuarios en sitos como las redes sociales, han convertido estos espacios en las fuentes más grandes de datos debido a su gran flujo de información [3].

Se plantea la idea de aplicar técnicas que ayuden a cubrir la necesidad de herramientas y métodos de extracción de información a partir de múltiples fuentes heterogéneas [4]. Implementando el uso de técnicas de crawling [5] [6], las cuales ayudan al rastreo y extracción de información de páginas web o entornos virtuales. Se continúa con la manipulación de la información extraída aplicando técnicas de PLN e implementando métodos de aprendizaje a nivel de análisis morfológico, sintáctico y léxico. El proceso anterior indica el uso de tokenización, lematización, segmentación de palabras e indexación semántica latente, lo cual lograra detectar qué información es útil para ciertos perfiles de usuario [7] e identificara palabras o conectores gramaticales que no son necesarios [8].

El modelo puede ser implementado por cualquier tipo de organización, que tenga procesos de divulgación y uso de tecnologías como redes sociales, blogs o páginas que ayuden a una mejor interacción con las opiniones dadas por sus propios usuarios. Para lograr una divulgación y aprobación por parte de la comunidad I+D+i, el modelo aplica normas y políticas para la gestión de calidad, lo que genera un nivel de confianza para los usuarios. Un mecanismo viable y que ayuda a cumplir los requisitos mencionados son los marcos de referencia de investigación, desarrollo e innovación (I+D+i – R&D) [9] [10] [11].

II. METODOLOGÍA

La elaboración y ejecución del modelo se basa en la metodología o proceso descrito en la Fig. 1, en la cual muestra cuatro fases importantes para la conclusión del mismo, donde la fase de antecedes define el camino ya tomado por diferentes investigaciones garantizando la necesidad de aplicar proyectos que tenga fines relacionado con el análisis de texto o información no estructura; para continuar, la fase de extracción de información da inicio al desarrollo del modelo exponiendo el origen de los datos que serán analizados en la fase de procesamiento de información, para así terminar con la estructuración de resultados finales.

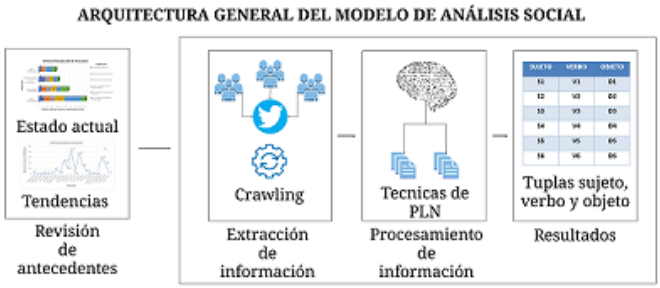

Fig. 1. Arquitectura aplicada sobre el modelo de análisis social.

A. *Antecedentes*

Se elaboró una revisión de antecedentes para caracterizar el estado actual de desarrollos e identificar tendencias que ayuden como insumo para la investigación; la búsqueda de documentación se realizó en bases de datos indexadas usando palabras claves relevantes a los temas centrales del modelo. Los términos principales fueron Redes sociales (social networks), web social, fuentes de información, crawling y extracción de información (extraction of information). También se aplicaron búsquedas relacionadas a: técnicas para la extracción de datos, minería de texto y minería web; análisis sintáctico (syntactic analysis), análisis léxico (lexical analysis), lematización sintáctica (syntactic lematization) e indexación semántica latente (latent semantic indexing). En cuanto a los temas enfocados a la estructura gramatical se exploró alrededor de análisis de cadenas textuales; lenguaje natural (natural language) y procesamiento de lenguaje natural (natural language processing). Y por último I+D+i, innovación (innovation), desarrollo (development), investigación (research) y R&D, para cubrir el enfoque a marcos de referencia para la creación de proyectos I+D+i.

Los resultados de las búsquedas se pueden analizar en la fig 2, donde se observa que varían dependiendo el tema y la base de datos en las que se encontraron los artículos usados para la conformación del estado de arte.

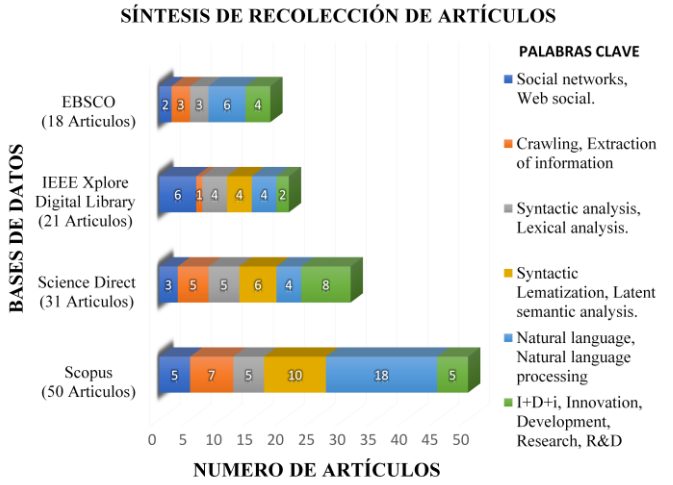

Fig. 2. Síntesis de recolección de artículos en bases de datos indexadas.

Las fuentes encontradas ratifican la necesidad de crear herramientas que ayuden a explorar mejor la información que se encuentra en las redes sociales o ambientes virtuales, afirmando que el análisis de los datos albergados en los servidores de estos sitios web son una buena estrategia de negocio, debido a que se generan escenarios como: lecciones aprendidas, casos de éxito y experiencias significativas [12]. Toda la información mencionada se puede sustentar con estadísticas reales dadas por eventos como: We are social que muestra en su última conferencia que, de los 7.593 millones de personas en el mundo, 4.021 millones están conectados a internet y de ellos 3.196 millones pertenecen a redes sociales, dando a conocer un aumento del 13% de usuarios nuevos en el transcurso de un año [13]. Todo este boom de nuevas tecnologías obliga las personas, sociedades y organizaciones a ser parte del cambio, demostrando que la información que se maneja es tan importante que muchas empresas en el mundo están dispuestas a adoptar estas tecnologías para ser parte de tal impacto [14]. Todo lo anterior se debe a que la evolución de las redes sociales fue mejorando desde una perspectiva estratégica como lo se demuestra el cambio de la línea de tiempo de la fig 3, donde se observa cada red social con el número de usuarios que posee hasta la fecha. Los cambios se ven reflejados dependiendo los avances de las redes sociales que se iban creando, donde las redes sociales con más números de usuarios son las cuales generaron un espacio que cuenta con el mayor número de servicios ofrecidos.

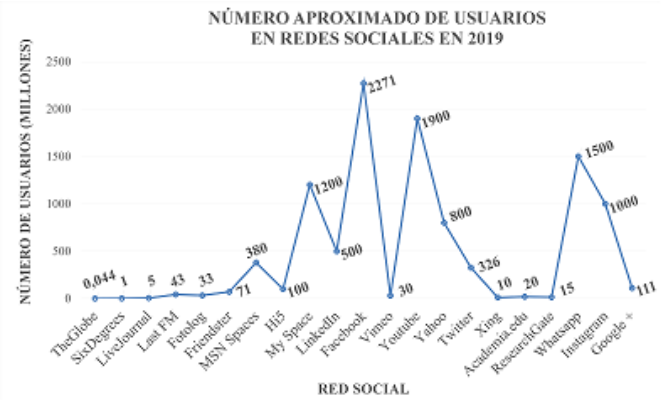

Fig. 3. Redes sociales con número de millones de usuarios hasta la fecha.

Con el crecimiento del uso de redes sociales y entornos virtuales el cambio o evolución tecnológica, ha generado que el análisis de datos históricos ya no sea una solución precisa; por ejemplo: si se tiene en cuenta el número de tweets, estados de Facebook, fotos en Instagram, videos en Youtube, etc, publicados en el transcurso de tiempo invertido en un análisis de un cluster con datos antiguos, ya se han generado nuevas fuentes de información, posiblemente con mejores datos que los que se están analizando. De tal manera, para lograr realizar un análisis detallado es necesario la implementación de herramientas de crawling, usando api's que permitan tal extracción en tiempo real, de esta manera realizar un análisis de toda la información posible.

Descubriendo la gran cantidad de información que se puede extraer usando api's especializadas es necesario implementar herramientas que ayuden al análisis de grandes cantidades de texto no estructurado; a partir de esto, el PLN comienza a hacer parte de la solución, siendo una de las técnicas computacionales que interactúan con el lenguaje humano a partir de software para ejecutar todo el proceso de análisis de texto de una manera automática [15], garantizando un análisis de la información con resultados provechosos, aplicando técnicas de análisis morfológico, sintáctico, léxico, tokenización, lematización, segmentación de palabras e indexación semántica latente.

Por último, se debe tener en cuenta que la creación y ejecución de proyectos relacionados con nuevas tecnologías deben ser implementados y publicados teniendo en cuenta normas que ayuden a su correcta divulgación dentro de un ambiente social, científico, político o industrial; todo se resalta en la aplicación de marcos de referencia I+D+i, los cuales al crear puentes de conexión entre empresa, estado y universidad o centros de investigación que apoyan a la externalización de conocimiento, garantiza a las organizaciones generar mejores avances tecnológicos en el campo de TI [16], ayudando a generar insumos que ayuden a emerger y crear innovación dentro de sus productos y procesos.

B. Extracción de información

Para la ejecución del crawling es necesario poseer los permisos por parte de twitter, red social elegida como primer origen de datos debido a su extenso contenido de opiniones derivadas de cualquier tema; al realizar todos los requisitos para el acceso de los datos por parte del api [17], se generar sus correspondientes credenciales de acceso. Con lo anterior, se selecciona Python como lenguaje de programación, en donde se ejecutarán cada una de las sentencias necesarias para la ejecución correcta de crawiling. Al finalizar la fase de exploración y extracción de información se obtendrá un archivo plano con los tweets que fueron seleccionados dependiendo el tema o palabras claves que fueron ingresadas, todo esto con el fin de generar un primer filtro, debido a que se publica continuamente mucha información y esto ayudara a balancear el nivel de procesamiento

C. Procesamiento de información

Al poseer un documento con tweets que analizar se recurre a la librería de PLN de Python NLTK, la cual funciona como una buena herramienta de análisis de lenguaje; al tener todo el entorno instalado se prosigue con el refinamiento de la información realizando una tokenización del tweet, para identificar stopwords o palabras vacías como lo pueden ser artículos, pronombres, preposiciones, etc y así garantizar la eliminación de estructuras que no es necesario de analizar. El siguiente paso es implementar algoritmos especializados que ayuden a realizar análisis morfológico, sintáctico y semántico y así generar categorización o tagged, el cual identifica cual es el origen léxico de cada una de las palabras del tweet, para esta tarea NLTK posee un corpus llamado CESS_ESP, en el cual se encuentran una gran cantidad de palabras en el idioma español con sus respectivas etiquetas.

Al elegir un corpus y etiquetador se procede a seleccionar el algoritmo que se desea emplear dentro del etiquetador, NLTK plantea que el análisis de texto y lenguaje se tiene que centrar en el contexto en que se llevan las palabras, de tal manera la palabra que se desea etiquetar tiene que verificar el contexto en que se etiquetaron las palabras anteriores, esto se llama el análisis en n-grams o etiquetador n-grams.

Con el PLN natural realizado, ayudara a generar las etiquetas de cada una de las palabras que componen los tweets para que con esto se puedan evaluar con herramientas especializadas para detectar tendencias o información provechosa que ayude a la toma de decisiones. Como resultado final se desea exportar un documento o archivo plano en el cual se encontrarán las tuplas SUJETO, VERBO, OBJETO de cada uno de los tweets analizados, lo cual es un insumo importante para detectar tendencias que se pueden presentar en la población seleccionada y al mismo tiempo ejecutar herramientas que ayuden a generar nuevos análisis de tal información.

III. MATERIALES Y MÉTODOS

Para desarrollo y ejecución del modelo se identificaron recursos necesarios los cuales fueron resultado de la exploración de información realizada, a continuación, se explica da uno de ellos:

A. Api twitter

El elemento principal para obtener acceso a la información, es el api ofrecida directamente por Twitter, la cual en su

última versión implementa diferentes métodos ya estructurados que ayudan a la manipulación de la información que es publicada dentro de la red social; cabe mencionar que el acceso a este api debe ser con claves de acceso suministradas directamente por parte del sistema de twitter, de otra manera no se podrá tener acceso a la información.

B. NLTK - Natural Language Toolkit

Es una plataforma que facilita la creación de algoritmos que ayuden al análisis de texto escrito o estructurado en lenguaje humano [18], esta herramienta ofrece recursos como: textos raíces o corpus, bibliotecas de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis, lematización y razonamiento semántico. Todo esto basado en métodos informáticos con argumentación matemática para garantizar los resultados obtenidos con la aplicación de la herramienta.

C. Etiquetas Eagle

Para la ejecución del análisis morfológico de texto es necesario el uso de etiquetadores los cuales generan un sistema de etiquetado que ayuda a conocer el valor léxico y morfológico de cada palabra. En el caso del etiquetamiento de palabras en idioma español el grupo Eagles [19], crea el modelo de etiquetamiento el cumple con la identificación de adjetivos, adverbios, artículos, determinantes, nombres, verbos, pronombres, conjunciones, numerales, interjecciones, abreviaturas, preposiciones y signos de puntuación. El modelo de etiquetamiento crea etiquetas basadas en las características de tiene cada palabra; por ejemplo, para etiquetar adjetivos los etiquetadores se basan en los atributos que se muestran en la tabla 1, donde la etiqueta comenzara con la letra A por pertenecer a la categoría de adjetivo y continuara con las características que se cumplan de tipo, grado, genero, numero, caso, función.

TABLA I
Atributos de un adjetivo basado en etiquetas Eagle [19].

Pos.	Atributo	Valor	Código
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
3	Grado	Apreciativo	A
4	Genero	Masculino	M
		Femenino	F
		Común	C
5	Numero	Singular	S
		Plural	P
		Invariable	N
6	Caso	-	0
7	Función	Participio	P

D. Corpus CESS_ESP

El análisis morfológico, sintáctico y semántico es parte esencial para generar la categorización o tagged, el cual identifica cual es el origen léxico de cada una de las palabras del tweet, para esta tarea NLTK posee un corpus llamado CESS_ESP, en el cual se encuentran una gran cantidad de palabras en el idioma español con sus respectivas etiquetas como lo muestra la fig 4, encontrando la palabra o lema, el tiempo y la categoría léxica en que se encuentra, por parte de la categorización tenemos que tener en cuenta que se llevara a cabo a partir de etiquetas Eagle.

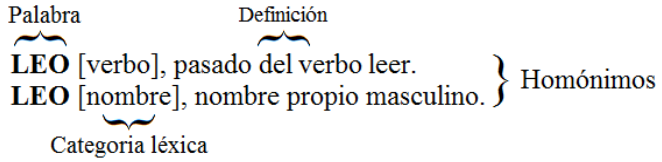

Fig. 4. Secciones claves de una palabra en un etiquetador. Fuente. Adaptado de NLTK Book.

E. Etiquetador n-gram

Al poseer el corpus y modelo de etiquetas que se van a manejar para la el etiquedado de cada una de las palabras que se van a analizar, se plante el uso de un etiquetador especializado basado en el modelo mostrado en la fig 5, donde se muestra como es el funcionamiento para etiquetar una palabra a partir de n-grams, partiendo de W_n siendo la palabra a etiquetar y W_{n-1}, W_{n-2}, W_{n-n} son los tokens anteriores que ya fueron etiquetados de esta manera podemos recorrer las etiquetas tn-1, tn-2, t_{n-n} para poder evaluar su contexto y así identificar la etiqueta de la palabra actual o t_n.

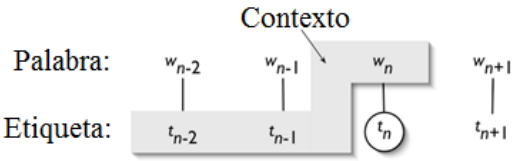

Fig. 5. Modelo de ejecución de un etiquetado N-gram. Fuente. Adaptado de NLTK Book.

Los etiquetados n-gram toman un nombre en específico cuando se analizan un numero de palabras anteriormente mencionadas y de esta manera lograr garantizar que cada una de ellas pertenezca al contexto de la palabra que se está analizando; en (1), se muestra la validación lógica que se toma a la hora de la ejecución de un n-gram, teniendo en cuenta que C es el contexto que se va alimentando desde el análisis de la primera palabra, T es una etiqueta y T_{wn} es la palabra a etiquetar.

$$\boldsymbol{Tw_n = T \in [C \in (Tw_{n-1}, Tw_{n-2}, Tw_{n-3}, \dots, Tw_{n-m})]} \quad (1)$$

Teniendo en cuenta que la elección de la etiqueta de la palabra a etiquetar pertenezca al contexto al que pertenezcan las etiquetas inmediatamente anteriores.

IV. Resultados y Discusión

Durante el desarrollo del modelo se identificaron y se ejecutaron varios escenarios, en los cuales se obtuvo resultados clave para la ejecución de la arquitectura general del modelo de análisis de texto (fig 1), cada uno de los escenarios muestra evidencias de la ejecución de cada una de

las fases.

A. *Escenario uno*

Se plantea la creación de un modelo en leguaje Python en el cual se programarán todos algoritmos necesarios. En la fase de extracción de información de Twitter se realizó un algoritmo de crawling el cual cumple con el diagrama de flujo mostrado en la fig 6, como primera medida el algoritmo desarrollado permite conectarse en tiempo real a la plataforma de twitter y extraer tweets completos que son publicados dentro del territorio colombiano y que se identifiquen con unas palabras claves que fueron agregadas con anterioridad.

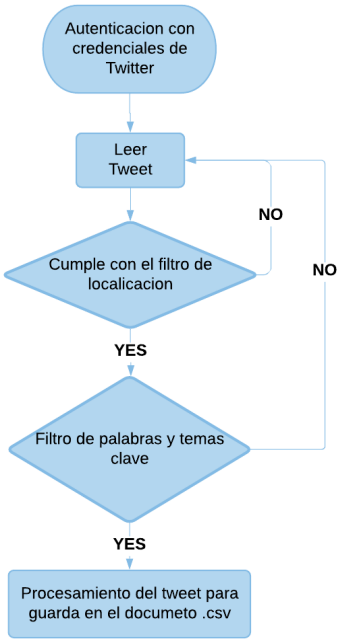

Fig. 6. Diagrama de flujo del algoritmo aplicado para la extracción de información.

Al realizar todo el proceso de crawling el objetivo es obtener un documento con los tweets que fueron extraídos exitosamente de la plataforma como lo muestra la fig 7, con el tweet extraído de una manera ordenada ya es más sencillo aplicar herramientas de PLN, lo que mejorara la identificación léxica y semántica que corresponde a tweet extraído, al mismo tiempo nos generara un formato de etiquetación para cada uno de los componentes del tweet, donde la sección azul es un identificador único asignado para cada tweet y la sección roja es el tweet extraído.

Identificador tweet

```
07/03/2019//10:03:59--!!--Los semilleros de investigación serán espacios lúdicos que buscarán promover la cultura emprendedora y empresarial de instituciones públicas y privadas de los grados 9, 10, y 11 del departamento.
```

Tweet extraido

Fig. 7. Secciones de un tweet extraído.

B. *Escenario dos*

Con el tweet extraído, se comienza con la aplicación de los métodos de PLN como primera medida se tokeniza el tweet para posteriormente aplicar una lematización para conocer el origen de cada una de las palabras y así identificar palabras que estén mal escritas o simplemente no exista dentro del lenguaje español y de esta manera aplicar una limpieza de stopwords o palabras vacías y con esto proceder con el etiquetado de cada una de las palabras resultantes.

La etiquetación o tagged de las palabras en el PLN es crucial para un correcto análisis del lenguaje, debido a eso es necesario aplicar corpus especializados para cada uno de los lenguajes, como se mencionó con anterioridad se aplica un corpus llamado CESS_ESP el cual se ejecuta como un etiquetador aplicando etiquetas EAGLE. La fig 8 muestra un tweet que ya fue refinado quedando como se muestra la línea número 1, ya en la línea número 2 cada palabra de la lista contiene su etiqueta EAGLE, estas etiquetas se elaboran a partir de características que posea la palabra.

```
1 ['Antonio', 'desarrolla', 'investigación']
2 [('Antonio', 'np00000'), ('desarrolla', 'vmip3s0'), ('investigación', 'ncfs000')]
```

Tweet tokenizado

Tweet etiquetado

Fig. 8. Resultado de un tweet etiquetado por medio de corpus CESS_ESP.

Para validar el resultado de la fig 8 se tiene en cuenta la sección ETIQUETAS EAGLE, donde se explica su funcionamiento; con lo anterior, se da inicio a la etiqueta de un nombre como lo es la primera palabra de la lista de la fig 8, de esta manera se verifica la etiqueta relacionada a la palabra, la cual es NP00000, donde cada uno de los caracteres representa una característica que posee por cada atributo: N: categoría - nombre, P: tipo - propio, 0: no posee género, 0: no posee un número, 0: no posee un caso, 0: no es de genero semántico, 0: no posee un grado. De esta manera a parir de las etiquetas EAGLE podemos etiquetar cada una de las palabras del idioma español.

C. *Escenario tres*

Al ejecutar un esquema de etiquetas se elige que algoritmo encargara de etiquetar cada palabra como mejor corresponda al contexto que se está llevando en el texto. Como se mencionó anteriormente se elige el algoritmon n-grams llamado de BigramTagger que funciona como lo muestra la fig 5, debido a que es un etiquetador que tiene en cuenta para su trabajo las dos etiquetas inmediatamente anteriores a la que se desea etiquetar, asegurando que se va mantener un contexto dentro de los resultados finales, para elegir este tipo de algoritmos es necesario tener en cuenta la calidad de información que se va a etiquetar debido a que puede llegar el caso de que las dos palabras inmediatamente anteriores posean un error ortográfico o no se encuentren dentro del corpus, esto conllevaría a que la etiqueta que se asigne no sea correspondiente al contexto o discurso que está llevando el texto. Para la ejecución de BigramTagger es necesario la creación de un corpus que contenga cada una de las palabras del idioma con sus respectivas etiquetas, esto se hace a partir de un corpus que ya posee la librería NLTK.

Tweet extraido

```
['Enviada', 'de', 'la', 'OIT', 'al', 'Seminario', 'de', 'Derecho', 'Laboral', 'en', 'Unilibre', 'Pereira', ',', 'habla', 'sobre', 'la', 'importancia', 'de', 'los', 'semilleros', 'de', 'investigación', 'y', 'el', 'impacto', 'en', 'la', 'vocación', 'de', 'los', 'estudiantes']
```

Tweet etquetado

```
[('enviada', 'aq0fsp'), ('oit', None), ('seminario', 'ncms000'), ('derecho', 'aq0ms0'), ('laboral', 'aq0cs0'), ('unilibre', None), ('pereira', None), (',', 'Fc'), ('habla', 'vmip3s0'), ('importancia', 'ncfs000'), ('semilleros', None), ('investigación', 'ncfs000'), ('impacto', 'ncms000'), ('vocación', 'ncfs000'), ('estudiantes', 'nccp000')]
```

Fig. 9. Resultado del etiquetado de un tweet.

Luego del correcto etiquetado (fig 9) se emplean algoritmos de decisión para poder seleccionar los sujetos, verbos y objetos encontrados en cada uno de los tweets que se analizaron, al obtener las listas mencionadas, se prosigue a generar tuplas sujeto, verbo, objeto con cada uno de los elementos encontrados y de esta manera extraerlos en un archivo plano para que funcione como insumo en el análisis de información dentro de otra plataforma. Cada tweet analizado da como resultado un numero de tuplas que permite ser usada sobre otras herramientas, generando así información o datos de alta calidad; de tal manera, se puede concluir que entre más tuplas se generen es posible llegar a obtener resultados más notables y de mayor contribución; la fig 10 muestra el número de tuplas generadas por 94 tweets extraídos, donde se encuentran tweets con 1 y 1501 tuplas, con una media de 64 tuplas por tweet.

Fig. 10. Número de tuplas encontradas en 94 tweet analizados.

D. *Escenario cuatro*

Se tiene que tener en cuenta que a la hora de realizar el etiquetado de las palabras que componen los tweets extraídos, se presentan varios retos sintácticos para el etiquetador, un ejemplo claro de esto es la captura de tweets como el que muestra la fig 11, donde se puede ver que se inicia con el verbo "Enviada", un verbo pero con su primera letra en mayúscula, esto genera que la palabra en cuestión sea etiquetada sin categoría teniendo en cuenta las reglas lingüísticas que se tienen dentro de la librería de NLTK.

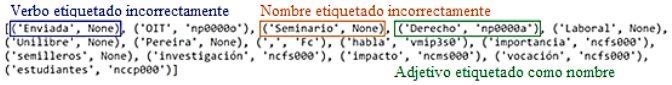

Fig. 11. Proceso de etiquetación con etiquetas incorrectas.

```
subjects = []
verbs = []
objects = []

for x in result_tagged:
    if x[1] is None:
            subject = x[0]
            subjects.append(subject)

for x in result_tagged_lower:
    if x[1] != None:
        validate = True
        i = 0
        for y in subjects:
            if y.lower() == x[0]:
                validate = False
                subjects.pop(i)
                break
            i=+1
        if validate is True:
            label = x[1]
            label_2 = label[0:2]
            label = label[0:1]
            if label is "v":
                verb = x[0]
                verbs.append(verb)
            elif etiqueta_2 is "np":
                subject = x[0]
                subjects.append(subject)
            elif label != 'F':
                object = x[0]
                objects.append(object)
```

Fig. 12. Fragmento de código algoritmo de filtrado objeto, verbo, objeto.

Estos problemas pueden ser comunes dentro de redes sociales o espacios virtuales donde sus usuarios no usen las reglas gramaticales con exigencia. Este problema o situación fue controlada programando un algoritmo (fig 12), que permite analizar primero los sujetos y convertir cada una de las letras en minúsculas y volver analizar y al mismo tiempo comparar si el verbo etiquetado se encuentra en el grupo de los sujetos, de ser así, es eliminado y etiquetado solamente como verbo, como se puede observar en la fig 13 cada token o palabra está en minúscula lo que genera que el etiquetado se pueda realizar de manera correcta.

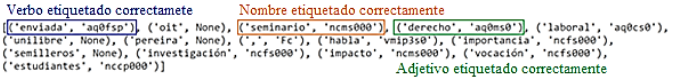

Fig. 13. Proceso de etiquetación con etiquetas correctas.

E. *Escenario cinco*

La ejecución del crawling en a la red social twitter fue todo un éxito, logrando extraer demasiada información que fue insumo para lograr elaborar un algoritmo de PLN que pudiera cumplir con los objetivos; algunos de los resultados obtenidos se pueden visualizar en la fig 14, cabe resaltar que el algoritmo ejecutado extrae tweet es tiempo real y de esta manera puede mantenerse conectado continuamente atento a capturar información que cumpla con las características requeridas, de esta manera se convierte en una herramienta de continuo análisis que funcionaria a toda cabalidad para cualquier organización.

```
1 03/04/19//12:41:54--!!--Enviada de la OIT al Seminario de Derecho Laboral en Unilibre Pereira, habla
  sobre la importancia de los semilleros de investigación y el impacto en la vocación de los
  estudiantes
2 03/04/19//12:53:52--!!--"Si todo avanza como está previsto, @hidroituango aportará su energía a
  partir del año 2021 para el desarrollo del país y el progreso de los colombianos. Este gran proyecto
  se constituye en la seguridad energética nacional para las próximas décadas" @EPMestamosahi
3 03/04/19//01:42:34--!!--#temporalesSENA somos el eje de la transferencia de conocimiento, del
  aprendizaje, la investigación, la innovación y enlace para que nuestros aprendices conquisten el
  mundo @EstradaCarlosM @MintrabajoCol @SENAComunica  @Farid_Figueroat  @wilsonariasc @baena
  @AliciaArango
4 03/04/19//02:52:21--!!--Aceptamos que los requisitos cambien, incluso en etapas tardías del
  desarrollo. Los procesos Ágiles aprovechan el cambio para proporcionar ventaja competitiva al
  cliente.
5 03/04/19//02:52:24--!!--Los procesos Ágiles promueven el desarrollo sostenible. Los promotores,
  desarrolladores y usuarios debemos ser capaces de mantener un ritmo constante de forma indefinida.
6 03/04/19//04:02:24--!!--Mañana damos inicio al #CongresoFIAPASOFONDOS  @fondosdepension . El equipo
  de #DobleP está orgulloso de formar parte de estas iniciativas que apuestan por el desarrollo del
  país.
7 03/04/19//04:04:16--!!--@unaroladice @Cloquis Correcto en el asunto cosmético no hay discusión. Pero
  el asunto central es la investigación farmacológica dirigida a la medicina veterinaria.
```

Fig. 14. Resultado de la extracción de tweets.

Por último, al crear un proceso de extracción de información que se ejecute correctamente, es posible visualizar todos los casos posibles de análisis que se tienen que realizar, de esta manera la aplicación de la librería NLTK fue una tarea donde se logró aplicar la mayoría de métodos de refinamiento de información, lematización, análisis léxico, semántico, morfológico y sintáctico, logrando la creación de un algoritmo el cual sigue un ciclo como lo muestra el flujo de la fig 15, siendo óptimo en etiquetamiento de palabras.

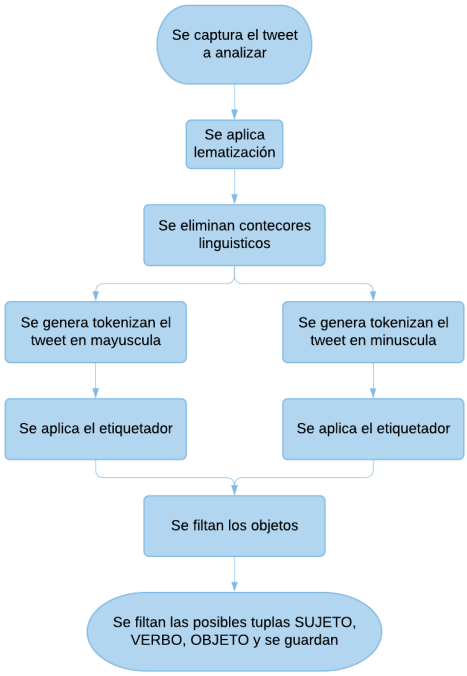

Fig. 15. Flujo para la extracción de tuplas sujeto, verbo y objeto.

Y así generando al final un documento estructurando como se pueden visualizar en la fig 16, mostrando cada una de las tuplas sujeto, verbo objeto que fueron extraídas de cada tweet analizado, generando finalmente información más detallada la cual es posible analizar y visualizar usando herramientas especializadas.

```
1  03/04/19//12:41:54//1--!--[Unilibre,habla,derecho]
2  03/04/19//12:41:54//2--!--[Unilibre,habla,importancia]
3  03/04/19//12:41:54//3--!--[Unilibre,habla,investigación]
4  03/04/19//12:41:54//4--!--[Unilibre,habla,impacto]
5  03/04/19//12:41:54//5--!--[Unilibre,habla,vocación]
6  03/04/19//12:41:54//6--!--[Unilibre,habla,estudiantes]
7  03/04/19//12:41:54//7--!--[Pereira,habla,derecho]
8  03/04/19//12:41:54//8--!--[Pereira,habla,importancia]
9  03/04/19//12:41:54//9--!--[Pereira,habla,investigación]
10 03/04/19//12:41:54//10--!--[Pereira,habla,impacto]
11 03/04/19//12:41:54//11--!--[Pereira,habla,vocación]
12 03/04/19//12:41:54//12--!--[Pereira,habla,estudiantes]
13 03/04/19//12:41:54//13--!--[semilleros,habla,derecho]
14 03/04/19//12:41:54//14--!--[semilleros,habla,importancia]
15 03/04/19//12:41:54//15--!--[semilleros,habla,investigación]
16 03/04/19//12:41:54//16--!--[semilleros,habla,impacto]
17 03/04/19//12:41:54//17--!--[semilleros,habla,vocación]
18 03/04/19//12:41:54//18--!--[semilleros,habla,estudiantes]
19 03/04/19//12:53:52//65--!--[hidroituango,avanza,si]
20 03/04/19//12:53:52//66--!--[hidroituango,avanza,previsto]
21 03/04/19//12:53:52//67--!--[hidroituango,avanza,energía]
22 03/04/19//12:53:52//68--!--[hidroituango,avanza,año]
23 03/04/19//12:53:52//69--!--[hidroituango,avanza,desarrollo]
24 03/04/19//12:53:52//70--!--[hidroituango,avanza,país]
25 03/04/19//12:53:52//71--!--[hidroituango,avanza,progreso]
26 03/04/19//12:53:52//72--!--[hidroituango,avanza,colombianos]
```

Fig. 16. Documento con las tuplas generadas con los tweets extraídos.

V. Conclusiones y trabajo futuro

Como se mencionó en el transcurso del trabajo el uso de las redes sociales para la extracción de información útil es un tema de gran impacto, por la cantidad de datos que se encuentra en ellas; desafortunadamente las redes sociales que se dedican a compartir conocimiento investigativo como Academia.edu y ResearchGate se han dejado a un lado por no dedicarse a temas de ocio, de esta manera se recomienda su uso por ser piezas clave para la innovación y el desarrollo de la sociedad, ya que la exploración de la información que se alberga en ellas asegura un porcentaje más alto de encontrar información que genere progreso y desarrollo dentro de las organizaciones; gracias a los diferentes casos de estudio y experiencias que se han investigado en los distintos sectores productivos.

Desde la perspectiva de gestión de la información se tiene que tener en cuenta que crear modelos o sistemas que se encarguen del manejo de información escrita en idiomas principales no es suficiente debido a que se cuenta con una gran diversidad de idiomas en el mundo y eso significa que puede existir información relevante que no ha sido utilizada, por no haber creado herramientas que cubran todos los modelos lingüísticos que existen; de tal manera es necesario seguir explorando en la investigación de nuevos métodos que ayuden con el manejo de la información en los diferentes idiomas del mundo.

En el desarrollo del crawling usando el api de Twitter, se logró notar que, por parte del estudio del análisis del leguaje, es necesario que otras plataformas entren en la tarea de desarrollar sus propias herramientas o permitan la conexión con sus procesos para capturar información y de esta manera lograr hacer ejercicios investigativos como el que se elaboró. Todo esto para lograr crear un enlace directo con la gran cantidad de información que se procesa o transmite en la red en el día de hoy.

La carencia de estudios para la creación de herramientas

computacionales que ayuden en el análisis del idioma español se convierte en un reto para centros de investigación, debido a que el español es declarado uno de los idiomas más usado en las redes sociales [20]. Esto significa que existe una gran cantidad de información sin analizar por no contar con las herramientas óptimas que logren un análisis óptimo de texto, de esta manera el desarrollo de un algoritmo basado en PLN y análisis léxico que permita el análisis de información en idioma español permitirá emerger en el estudio y análisis de una gran parte de la información que se alberga en internet. Generando la necesidad de crear corpus más completos, en donde se pueda contar con palabras que hagan parte de la jerga cultural y de esta manera lograr la ejecución de este tipo de algoritmos en información más informal.

REFERENCIAS

[1] O. Yagan, D. Qian, J. Zhang y D. Cochran, "Conjoining speeds up information diffusion in overlaying social-physical networks" IEEE Journal on Selected Areas in Communications, vol. 31, pp. 1038-1048, 2013. doi: 10.1109/jsac.2013.130606

[2] D. Mladenić y M. Grobelnik, "Automatic text analysis by artificial intelligence" Informática, Special Issue: 100 Years of Alan Turing and 20 Years of SLAIS Guest Editors, vol. 37, pp. 27-33, 2013.

[3] W. Tan, M. B. Blake, I. Saleh y S. Dustdar, "Social-network-sourced big data analytics" IEEE Internet Computing, vol. 17, pp. 62-69, 2013. doi: 10.1109/mic.2013.100

[4] H. A. Sleiman y R. Corchuelo, "A survey on region extractors from web documents" IEEE Transactions on Knowledge and Data Engineering, vol. 25, pp. 1960-1981, 2013. doi: 10.1109/tkde.2012.135

[5] M. J. Culnan, P. J. McHugh y J. I. Zubillaga, "How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value" MIS Quarterly Executive, vol. 9, pp. 243-259, 2010.

[6] C. Olston, M. Najork y others, "Web crawling" Information Retrieval, vol. 4, pp. 175-246, 2010. doi: 10.1561/1500000017

[7] M. A. Tayal, M. M. Raghuwanshi y L. G. Malik, «ATSSC: Development of an approach based on soft computing for text summarization,» Computer Speech \& Language, vol. 41, pp. 214-235, 2017. doi: 10.1016/j.csl.2016.07.002

[8] S. Sun, C. Luo y J. Chen, "A review of natural language processing techniques for opinion mining systems" Information Fusion, vol. 36, pp. 10-25, 2017. doi: 10.1016/j.inffus.2016.10.004

[9] I. Nonaka y H. Takeuchi, "Die Organisation des Wissens: Wie japanische Unternehmen eine brachliegende Ressource nutzbar machen", Campus Verlag, 2012.

[10] ICONTEC, "NORMA 5801, Gestion de la investigación, desarrollo e innovación (I+D+i). Requisitos del sistema de gestión I+D+i," Bogota D.C, 2008.

[11] C. tecnico AEN/CTN 166, «Gestion de la I+D+i: Requisitos del sistema de gestion de la I+D+i"[En línea] Disponible en: http://lidiagroup.org/images/descargas/citic/une_166002-2006.pdf

[12] M. J. S. Barón, "semantic analysis over lessons learned contained in social networks for generating organizational memory in centers R&D" Computer Science \& Information Technology (CS \& IT), 2016. doi: 10.5121/csit.2016.60621

[13] We_Are_Social, "Digital in 2018: global overview" [Online] Disponible: https://digitalreport.wearesocial.com/

[14] J. K. Sinclaire y C. E. Vogus, "Adoption of social networking sites: an exploratory adaptive structuration perspective for global organizations" Information Technology and Management, vol. 12, pp. 293-314, 2011. doi: 10.1007/s10799-011-0086-5

[15] E. Cambria y B. White, "Jumping NLP curves: a review of natural language processing research [review article]" IEEE Computational Intelligence Magazine, vol. 9, pp. 48-57, 2014. doi: 10.1109/mci.2014.2307227

[16] Manual de Oslo, "Guía para recogida e interpretación y datos sobre innovación" [En línea] Disponible en: http://www.itq.edu.mx/convocatorias/manualdeoslo.pdf

[17] "Twitter Developer Platform", Developer.twitter.com, 2019. [Online]. Disponible: https://developer.twitter.com/.

[18] Natural Language Toolkit — NLTK 3.4.1 documentation", Nltk.org, 2019. [Online]. Disponible: https://www.nltk.org/.

[19] "ETIQUETAS EAGLES", Lsi.upc.es, 2019. [Online]. Disponible: http://www.lsi.upc.es/~nlp/tools/parole-sp.html.

[20] Instituto Cervantes, Informe 2016 "El español: una lengua viva", [En línea] Disponible en: https://www.cervantes.es/imagenes/File/prensa/EspanolLenguaViva16.pdf

Johan David Diaz Mendivelso Sogamoso-Boyacá-Colombia un 27 de abril de 1993. Estado civil soltero y residente en Sogamoso-Boyacá-Colombia. Docente Investigador y actualmente profesor de tiempo completo en la escuela de ingeniería de sistemas y computación de la Universidad Pedagógica y Tecnológica de Colombia-UPTC. El área de desempeño laboral e investigativo se centra en temas de procesamiento de texto, análisis social, analítica de datos y aprendizaje de máquina. Su formación académica es el pregrado en Ingeniería de Sistemas.

Marco Javier Suarez Baron Duitama-Boyacá-Colombia un 18 de septiembre de 1974. Estado civil casado y residente en Tunja-Boyacá-Colombia. Investigador Asociado I de Colciencias y actualmente profesor de tiempo completo en la escuela de ingeniería de sistemas y computación de la Universidad pedagógica y tecnológica de Colombia-UPTC. El área de desempeño laboral e investigativo se centra en temas de Inteligencia artificial, aprendizaje de máquina y web semántica. Su formación académica es el pregrado en Ingeniería de Sistemas, master en Inteligencia artificial, magister en gestión de información, PhD en planeación y gestión de tecnología y estancias posdoctorales en Analítica de datos y análisis social. CvLAC: https://scienti.colciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000306851.