Determinación de niveles de agresividad en comentarios de la red social Facebook por medio de Minería de Texto

Martel Wilfredo, Carranco Diego, Cevallos Daniel
Universidad de las Fuerzas Armadas – ESPE
Quito, Ecuador
{wmartel, dcarranco, dcevallos}@espe.edu.ec

***Resumen*—La presente investigación está basada en la utilización de técnicas de Text Mining para análisis de comentarios realizados por los usuarios de la red social Facebook. Además, se utilizan diccionarios, conjunto de palabras ofensivas con pesos asignados, y algoritmos como el de Levenshtein que permiten encontrar la similitud entre dos palabras. Posteriormente se procede a la clasificación de los niveles de agresividad que van desde bajo, medio y alto, con el fin de clasificar las personas o entes cibernéticos que son potencialmente una amenaza al resto de la población virtual. Finalmente, como resultado de la investigación se obtiene un listado de personas que deberían ser investigados, expulsados o bloqueados de las redes sociales para prevenir futuros incidentes.**

Palabras Clave*—*Text Mining, Distancia de Levenshtein, API de consultas de Facebook.

I. INTRODUCCIÓN

En la actualidad debido al auge de las tecnologías de la información y las comunicaciones, se genera una auto dependencia de la personas a estas técnicas y herramientas, por tal motivo es inminente el quedar expuestos a grandes amenazas[12] entre las que se puede mencionar, el hurto y divulgación de datos confidenciales, ataques a sitios web para denegación de servicios, hurto de contraseñas e información, suplantación de identidad, y la intimidación o el acoso en sitios de opinión en la web, siendo los más comunes la redes sociales, blogs y foros.

Refiriéndonos específicamente a la red social Facebook® con un poco más de 1390 millones de usuarios agrupados en su mayoría entre adolescentes y adultos jóvenes, los cuales se encuentran expuestos y vulnerables constantemente a los bajos escrúpulos de entes cibernéticos los cuales pueden o no representar personas reales, con el fin de ofender, acosar, discriminar, maltratar, estafar, persuadir, o reclutar gente con fines delictivos[13], utilizando el foro o set de intercambio de notas de las llamadas páginas de fan de Facebook.

En respuesta a este peligro, es imprescindible el reconocimiento de los niveles de agresividad de los comentantes, de esta manera, basándose en algoritmos de distancia de Levenshtein [14] para reconocer las frases y palabras clave de acuerdo a un diccionario, conjunto de palabras ofensivas, establecido a nuestra zona geográfica, correspondiente en este caso a los ubicados en Ecuador y que realizan aportes al Fan Page de las instituciones gubernamentales elegidos como muestra, de esta parte significativa de la población de comentarios se pudo determinar patrones de comportamiento y así clasificar en niveles de agresividad [29][28] a los individuos cibernéticos que muchas veces se esconden en falsos perfiles para hacer daño [13].

El trabajo de investigación por lo antes expuesto se basa en agrupar a los individuos que exponen sus comentarios en niveles de agresividad para poder clasificarlos, a diferencia del Webcrowling y Minería de Opinión, que busca determinar agrupaciones sociales con metodología de clustering para fines de publicidad. Esto significa que al identificar posibles o potenciales amenazas se puede separar a estos grupos con el fin de darle el uso adecuado a las redes sociales que son parte de la web 2.0, enfocándolo incluso con fines educativos [17].

El resto del artículo ha sido organizado por secciones en las que se expone un marco teórico con los conceptos básicos y su uso en la investigación, posteriormente en el numeral tres y cuatro nos adentramos en la metodología y la topología de investigación para la obtención de resultados confiables y consistentes, a continuación se expone los trabajos relacionados de otros autores como referencia en el numeral cinco y finalmente las conclusiones y trabajos futuros que se pretenden por los investigadores de esta publicación.

II. MARCO TEÓRICO

A continuación, se detalla una serie de conceptos acerca de técnicas, algoritmos y métodos utilizados para dar marcha, ejecución y obtención de resultados del proyecto de investigación.

A. Minería de Texto

Técnica utilizada para extraer desde un texto plano datos que puedan generar información relevante, en nuestro caso se utiliza para limpieza, reconstrucción de datos y procesamiento de los comentarios que posteriormente serán analizados por el algoritmo de Levenshtein, con la finalidad de encontrar relación entre las palabras.

B. Distancias de Levenshtein

Es una técnica matemática[14] desarrollada para determinar el número de operaciones en que una cadena puede transformarse en otra. Su campo de aplicación va desde aplicativos de correctores ortográficos, sistemas de reconocimiento de voz hasta sistemas de detección de plagios.

Con esta herramienta se puede determinar la probabilidad de similitud entre dos palabras o frases. En nuestra investigación se utiliza esta técnica para comparar cada palabra de un comentario con un diccionario definido en una base de datos.

A continuación, se exponemos un ejemplo del algoritmo de Levenshtein:

La distancia entre “hola” y “brola” es 2, porque hay que hacer 2 operaciones sobre la palabra “hola” para obtener” brola”:

substituir la “h” por una “r” e insertar una “b”.

Consideraciones:

- Cuanto más corta es la distancia entre las dos cadenas, más parecidas son. Si la distancia es 0, las dos palabras son iguales.
- El algoritmo de Levenshtein no tiene en cuenta consideraciones fonéticas. Por ejemplo, en español, dos apellidos como Hernández y Fernández.

TABLA I. ALGORITMO DE LEVENSHTEIN

```
static int Levenshtein(string s1, string s2) {
  int coste = 0;
  int n1 = s1.Length;
  int n2 = s2.Length;
  int[,] m=new int[n1+1,n2+1];

  for (int i = 0; i <= n1; i++) {
    m[i,0] = i;
  }
  for (int i = 1; i <= n2; i++) {
    m[0,i] = i;
  }
  for (int i1 = 1; i1 <= n1; i1++) {
    for (int i2 = 1; i2 <= n2; i2++) {
      coste = (s1[i1 - 1] == s2[i2 - 1]) ? 0 : 1;
      m[i1, i2] = Math.Min(
        Math.Min(
          m[i1 - 1, i2] + 1,
          m[i1, i2 - 1] + 1
        ),
        m[i1 - 1, i2 - 1] + coste
      );
    }
  }
  return m[n1, n2];
}
```

C. *Red Social Facebook*

Es un aplicativo web creado por David Zuckerberg, destinado a la interrelación virtual de personas, y representadas por usuarios o avatares en la que se permite el intercambio de información multimedia, comentarios y opiniones. Además, esta red social cuenta con una gran concurrencia de usuarios la cual es perfecta para realizar la investigación. Se debe hacer énfasis que la red social cuenta con una herramienta Graph API [14] que permite extraer información de las publicaciones realizadas.

D. *Fan page*

Es la sección o funcionalidad de la red social Facebook en donde se permite la creación de un perfil público en el que se permite publicar contenido multimedia y de opinión y los usuarios que gusten (Like) de la misma puedan dar sus comentarios e interactuar con otros internautas.

E. *Población*

En estadística corresponde al universo de datos o anotaciones posibles para el caso de la investigación sería todo el universo de comentarios públicos expuestos en las fans pages de las instituciones gubernamentales.

F. *Muestra*

Es la representación significativa de la población en este caso se ha escogido a los comentarios de los perfiles de las paginas gubernamentales del Ecuador.

G. *Patrón de comportamiento de agresividad*

En psicología estos patrones de conducta violenta varían según la frecuencia y asociación entre conductas, los tipos de conducta agresiva que manifiestan y la situación en la que se presentan dichas conductas, apareciendo los gestos de ira y la agresión verbal como conducta más frecuente [4]. A continuación se expone brevemente una clasificación conceptual de la agresividad basada en [5]:

- Agresión física: Ataque a un organismo mediante armas o elementos corporales, con conductas motoras y acciones físicas, el cual implica daños corporales.
- Agresión verbal: Respuesta oral que resuelta nociva para el otro, a través de insultos o comentarios de amenaza o rechazo.
- Agresión social: Acción dirigida a dañar la autoestima de los otros, su estatus social u ambos, a través de expresiones faciales, desdén, rumores sobre otros o la manipulación de las relaciones interpersonales.

Existen muchos más tipos de agresiones, pero de acuerdo a la clasificación clínica se ha establecido diferencias para poder distinguirlas, entre ellas:

- Agresividad reactiva: Motivada por la emoción de la ira.
- Agresividad instrumental: Busca evitar un obstáculo o lograr una meta sin que exista una motivación de la emoción de la ira.

Esta investigación, se centra en la agresividad reactiva porque es la explosión de ira del individuo bajo ciertas condiciones. Es decir, que una persona con una conducta agresiva, en su momento puede agredir físicamente o insultar a otro solo por el hecho de tener ideologías distintas. Justamente las agresiones verbales son de interés para la investigación porque se persigue obtener perfiles de agresividad de acuerdo al vocabulario o forma de expresarse de una persona.

H. *Nivel de agresividad*

De acuerdo a lo expuesto en el punto g, para poder determinar perfiles que nos indiquen el nivel de agresividad de una persona en base a su vocabulario o forma de expresarse, se debe contar con una clasificación de acuerdo a la expresión utilizada, esto significa que se debe tener un diccionario de palabras nocivas [12] y clasificarlas de acuerdo a su ofensa. En esta investigación se determina tres niveles:

- Nivel bajo: se refiere a un perfil pasivo, muy poco sociable, negativo, siempre llevando la contraria y con mucha ansiedad de ser tomado en cuenta. Las palabras utilizadas por este perfil no son tan ofensivas porque trata de encontrar un equilibrio. El tipo de individuos de este perfil, después de ofender casi siempre terminan con gesto amable.
- Nivel medio: se refiere a un comportamiento medio explosivo, negativo y solo en ciertos casos dependiendo del tipo de interacción. Se consideran accesible, y pueden manejar la situación.
- Nivel alto: se refiere a un comportamiento amenazante, dominante en todos los sentidos, quiere siempre tener la razón, aunque carezca de ella, utiliza palabras ofensivas e hirientes para tratar de que su idea sea aceptada. Además, se debe presentar mucha atención a las ofensas porque puede tornarse en realidad.

I. *Amenaza*

Se refiere a la intención o al peligro inminente de ocurrir algún tipo de actividad que ponga en riesgo la integridad física, psicológica o social de una persona o grupo de personas.

J. Zonificación

Constituye a dar un espacio geográfico, agruparlo socialmente o por ciertas características que permita el tratamiento de la información generada de forma homogénea y sin influencia de datos aislados.

K. Variabilidad Lingüística

Corresponde a los diferentes matices que tiene el lenguaje incluso en el mismo idioma que hace de una zona o grupo social diferente. Es decir, no es lo mismo decir rata en un tema de biología que rata en un entorno político. Por tal motivo es importante tomar en cuenta este aspecto al momento de realizar la limpieza de los datos y su análisis.

L. R Studio

Herramienta informática, incorpora el lenguaje R, para extracción y procesamiento de información para darle un tratamiento, matemático, estadístico basado en modelos previamente establecidos. Esta herramienta cuenta con varios paquetes de análisis de texto tal como el de Levenshtein que a su vez tiene un compendio completo de algoritmos para encontrar similitudes entre cadenas, las cuales son útiles para la investigación.

III. MATERIALES Y MÉTODOS

Para llevar a cabo la investigación planteada, se tuvo que pasar por un proceso riguroso de extracción de información, depuración y posteriormente a su análisis en las herramientas de minería de datos. A continuación, se expone la topología (ver Fig., 1) utilizada que describe de manera general, el proceso para proseguir a la explicación del paso a paso. La experimentación se centra en tener un muestreo de comentario de la red social utilizando el Graph API de Facebook, para posteriormente realizar una limpieza de los datos, que consisten en eliminar tildes palabras no deseadas, emoticones y reconstrucciones de palabras por conflictos de codificación [15]. Una vez, que la información se encuentre limpia para el análisis, se procede a extraerla utilizando la herramienta R, la cual se conecta a la base de datos de PostgreSQL y procede a clasificarla, ordenarla y guardarla por medio de un algoritmo escrito en R [13], el cual, a su vez, incorpora otros algoritmos embebidos. Hay que mencionar, que los algoritmos interactúan directamente con el diccionario de la base para identificar si el contenido del comentario es agresivo y si es, identifica la probabilidad de similitud del contenido con la del diccionario. De acuerdo a las pruebas realizadas, la probabilidad de aceptación para encontrar la similitud más aproximada entre dos palabras es de 90% [13]. Finalmente, se obtiene un listado de resultados de todo el proceso realizado, los mismos que se exponen en una hoja de cálculo para su comprobación y presentación [14]. A continuación, se describe el proceso de extracción de información y análisis de datos.

Fig. 1. Topología de Extracción y Análisis de datos

A. Perfil de agresividad:

Determinar el perfil de agresividad de una persona en base a sus publicaciones, no es un trabajo sencillo, porque involucra analizar aspectos psicológicos del individuo. En esta investigación se utilizó un diccionario [12], conjunto de palabras ofensivas, las cuales fueron ponderadas de acuerdo a su agresividad. A continuación, se muestra un ejemplo.

TABLA II. DICCIONARIO DE PALABRAS CON PONDERACIÓN

Palabras	Peso	Palabras
Maldito	1	Maldito
Lárgate	0.5	Lárgate
Estúpido	0.4	Estúpido

La Tabla I, es solo una muestra de archivo original, esto se debe a su gran magnitud de palabras que son involucradas, pero parte de esta sección esta explicada en el video [16] donde se explica pasó a paso el proceso de la experimentación.

B. Niveles de agresividad:

Permite clasificar el nivel en que el usuario se encuentra. Por tal motivo se ha definido una matriz, tal y como se muestra en la Tabla 3.

TABLA III. NIVELES DE AGRESIVIDAD POR RANGO

Nivel de agresividad	Identificador	Rango
BAJO	B	0 .1- 0.4
MEDIO	M	0.5 – 0.7
ALTO	A	0.8 - 1

- Bajo: indica todas aquellas palabras que contengan peso entre 0.1 y 0.4.
- Medio: indica todas aquellas palabras que contengan peso entre 0.5 y 0.7

- Alto: indica todas aquellas palabras cuyo peso este entre los 0.8 y 1.

C. *Facebook fan page*

Para poder analizar los comentarios de las publicaciones, se necesitaba que sean necesariamente públicos por dos cosas: la primera, porque Facebook no presenta restricciones; la segunda, porque las personas no se limitan a decir lo que piensan, y eso es justamente lo que se buscaba en la investigación, que las personas sean libres de expresarse, para poder determinar su nivel de agresividad a través de sus comentarios. Así que, se escogió las fans page del gobierno por las cantidades enormes de comentarios realizados.

D. *Facebook Graph API*

Es una API de Facebook [14] que permite interactuar a los desarrolladores con Facebook y de esta formar tener acceso a los datos, pero solo funciona con páginas públicas de Facebook. Cabe mencionar que requieren de token, la URL y un identificador del post. Los identificadores del post, es una numeración que facebook asigna a cada comentario que un usuario realiza. A continuación, se describe un ejemplo de lo expuesto anteriormente.

Se establece la URL del a fan page de la cual se desea extraer todos los comentarios. URL a ser analizada: https://www.facebook.com/MashiRafael/photos/a.339953076034199.95492.248984428464398/1080747815288051/?type=1&theater

De esta URL lo que interesa es el identificador [id_post]:1080747815288051. Este identificador permite obtener todos los comentarios de dicha publicación, además devuelven datos de las personas que comentaron incluido la fecha de publicación.

Ahora se hace uso del API de Facebook para obtener los comentarios.

http://graph.facebook.com/[id_post]:/comments?limit=numComentarios&token=

- id_post: es el identificador
- numComentarios: cantidad de comentarios a mostrar.
- token: código generado por facebook, la cual te permite tener acceso a consultas via query por un tiempo determinado.

Al reemplazar los valores, usted obtendrá todos los comentarios realizados en el post.

E. *Transformación de JSON a filas y columnas*

Los datos obtenidos del API de Facebook, retornan datos en json y estos necesitan ser transformados a una estructura estática definida por nosotros en la base de datos y para posteriormente ser leídos en la base de datos postgreSQL y analizarlos con la herramienta R.

Para el proceso de transformación se utilizó la propia herramienta de postgreSql, es decir la base misma soporta almacenar datos en json y así mismo permite recorrerlos como si fuera una Tabla normal, ver Tabla 4.

F. *Inserción de Json a Base de Datos postgreSQL:*

Para este proceso se utilizaron sentencias SQL que permiten el trasformado de los nodos Json a registros de texto plano. A continuación, se muestra un ejemplo de la conversión de Json a texto plano.

TABLA IV. CONVERSIÓN DE JSON A TEXTO PLANO

```
SELECT (o->'from'->>'name') as Nombre,
       (o->'from'->>'id') as Id,
       (o->>'message') as Mensaje,
to timestamp((o->>'created time'),'YYYY-MM-DD"T"HH24:MI:SS"Z"') as FechaPublicacion,
       (o->'comments'->'data') as Comentario,
  FROM sitio s
  , json_array_elements(data->'data') o
```

Lo expresado en la Tabla 4, es la forma de transformar json a texto plano, sin necesidad de requerir a herramientas de terceros.

G. *Depuración de datos la base PostgreSQL*

Para poder analizar la información en R, es necesario que haya palabras completas y estén casi correctamente escritas, de caso contrario habrá problemas al momento de analizarlas. Por tal razón es necesario aplicar algoritmos [22] de corrección como los que fueron utilizados en esta investigación. A continuación se expone un ejemplo de la limpieza de datos.

Ejemplo: "Lucy es corrupt@ y tiene ….siempre….será…asi..:) ".

Comentarios como el que se acaba de ejemplificar son muy comunes en los usuarios. Debido a esto, se deben hacer limpieza de datos. En este caso, para el análisis del comentario es necesario quitar los puntos, paréntesis, y reemplazar las tildes para que los algoritmos puedan trabajar correctamente.

H. *Análisis de los datos en R*

Esta herramienta integra algunos algoritmos para análisis de texto, tales como el de Levenshtein cuyo nombre de paquete es "stringdist". Este paquete tiene muchos algoritmos entre ellos el "stringsim" de la misma familia. Este último algoritmo, retorna la probabilidad de similitud entre dos palabras. Los parámetros de "stringdist" se describen a continuación. La Tabla 4 muestra su forma de uso:

- Palabra objetivo: es la palabra ideal que se desea encontrar, en nuestro caso proviene de nuestro diccionario.
- Palabra origen: son los comentarios localizados en nuestra base de datos.
- Método: el método utilizado es el de "jw", el cual brinda un mejor rendimiento en comparación respecto a los otros.
- Probabilidad de acierto: es la probabilidad de acierto de similitud entre dos palabras. Entre más cercano este a cero mucho mejor será la comparación pero para eso deben ser extremadamente idénticas por tal razón, se mantiene un valor prudente para nuestro propósito.

TABLA V. USO DE LA FUNCIÓN "STRINGSIM"

```
prob <- stringsim ('ALTO','ALTAR' ,method = 'jw',p=0.08);
```

palabra character vary	peso numeric(4,2)	nivel charact	mensaje_analisis text
corrupto	0.50	M	vivimos tiempo donde
ignorante	0.50	M	periodismo pais e
miserable	0.70	M	creer estas almas
estupido	0.70	M	jajajajaja causa risa
parido	0.90	A	presidente difer
borrego	0.30	B	presidente difer
perro	0.50	M	jorge guaman has borra
ladron	0.50	M	igualitos lado lado
ladron	0.50	M	igualitos lado lado
corrupto	0.50	M	hablando feriados ban
meco	0.40	M	presidente rafael
culiar	0.80	A	presidente rafael
conche	0.80	A	conoce historia rec
ano	0.50	M	conoce historia rec
corrupto	0.50	M	cerrara esa corru
huevada	0.50	M	pana tantas huevadas
pedo	0.20	B	cuente asambleita
ano	0.50	M	sale gran interce
cojo	0.30	B	sale gran interce
corrupto	0.50	M	gente mayoria
mediocre	0.60	M	xq aca
pene	0.30	B	ecorae traduce nekas
pedo	0.20	B	suena perdio
tonto	0.30	B	justamente gente us
payaso	0.30	B	cierra ojete payaso

Fig. 2. Resultados del Proceso de Análisis de Agresividad al de las palabras más utilizadas por los usuarios

En la Fig.2, se observa las palabras del diccionario que están inmersas dentro del comentario. Además, se observar el peso de cada palabra y el nivel de agresividad.

En la Fig.3, se muestra el resultado del procesado de texto, que duró aproximadamente 15 minutos para procesar 3762 comentarios, teniendo en cuenta que cada comentario tiene aproximadamente entre 20 a 1200 palabras. El resultado con una probabilidad del 90% de similitud entre las palabras, se encontraron 1318 coincidencias con el diccionario.

IV. Evaluación De Resultados

La muestra utilizada para esta investigación fue de 2791 usuarios en trece publicaciones distintas. Antes de empezar a encontrar la relación de los niveles de agresividad de los usuarios, se realizará una exploración de la información que se ha obtenido, producto del resultado del algoritmo de clasificación de palabras. La Tabla 6 muestra las diez palabras más utilizadas en los comentarios.

TABLA VI. LAS DIEZ PALABRAS MÁS UTILIZADAS

Palabras	Peso	Nivel de agresividad	Frecuencia	Probabilidad
Ano	0.50	M	91	0,069
Asco	0.50	M	85	0,064
Basura	0.50	M	77	0,058
Borrego	0.30	B	68	0,052
Burro	0.50	M	52	0,039
corrupto	0.50	M	42	0,032
Ladrón	0.50	M	39	0,030
Mierda	0.80	A	37	0,028
Puta	1.00	A	35	0,027
Rata	0.70	M	33	0,025

Al realizar la primera consulta a la base de datos, sobre las 1318 incidencias que hubieron, se encontró que 91 de las 1318 se refieren a la palabra “ano”, lo que significa que el 0,069% de las personas que han participado en esas publicaciones tienen esa palabra en su vocabulario.

En segundo lugar, le sigue la palabra “asco” sinónimo de repugnancia hacia algo, la probabilidad de estar presentes en sus vocabularios es del 0,064%. Además, se observa que el nivel de agresividad de las top 10, en su mayoría es media y le sigue la alta, lo que indica que los usuarios tienen un comportamiento explosivo inicial y durante el intercambio de opiniones.

Si se realiza una sumatoria de la probabilidad se tiene que el 42,41% de los usuarios que han publicado utilizan cualquiera de estas palabras durante sus conversaciones. La Fig.3, muestra la división porcentual de las palabras más utilizadas por los usuarios.

Fig. 3. División porcentual de las palabras más utilizadas por los usuarios

- Frecuencia del nivel de agresividad

TABLA VII. Frecuencia del nivel de agresividad en los individuos de investigación

Nivel de agresividad	Frecuencia	Porcentaje
Mediano	860	0,65
Alta	243	0,18
Baja	215	0,16

Al observar la Tabla 7, se verifica que el nivel de agresividad más frecuente en los individuos investigados es el mediano. Es decir, tienen un comportamiento explosivo, al inicio y durante el intercambio de ideas y en sus vocabularios están inmersas las palabras de la Tabla 6.

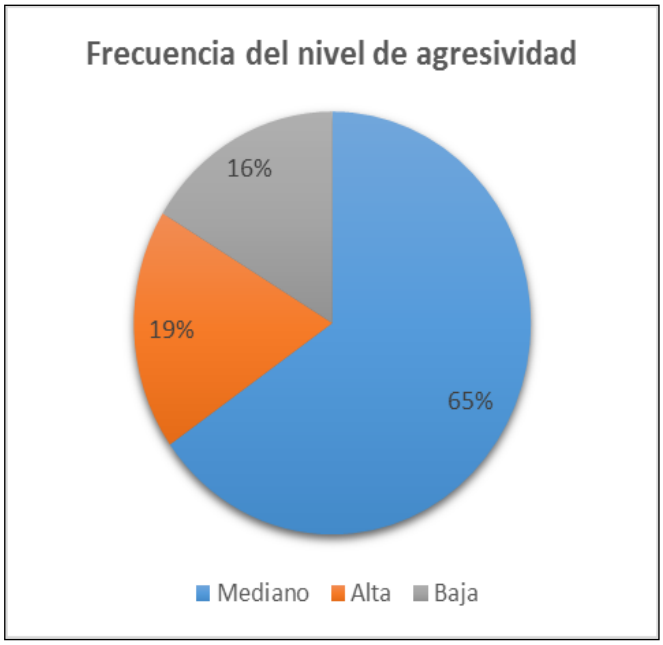

Fig. 4. Frecuencia de niveles de agresividad

Sin embargo, lo más preocupante, es el segundo lugar de los niveles de agresividad, porque en su vocabulario hay palabras agresivas consideradas las más ofensivas de todas. Por lo tanto, se puede inferir que el 18% de los individuos investigados son altamente ofensivos.

Además, en la Fig.4, se observa que un 16% de los individuos investigados poco o casi nada agresivos con otras personas.

- Top 10 de los individuos con nivel de agresividad alta.

La Tabla 8, lista los individuos con nivel de agresividad alta y también, lista los números de insultos realizados hacia otra persona. Para poder identificar los individuos, se realizó una consulta a la base de datos de los 10 primeros individuos con insultos graves ordenados descendentemente y como resultado tenemos ese listado. Pero si se desea saber más, sobre las palabras más utilizadas por estas personas, ver la Fig.5.

TABLA VIII. TOP TEN DE INDIVIDUOS CON NIVEL DE AGRESIVIDAD ALTA

Nombre	N° Insultos	Nivel
Alfredo Martínez	10	A
David Villacrés	4	A
Derek Córdova	3	A
Diego Galarraga Villalba	6	A
Edgar Ayovi	5	A
Fernando Juan Guamán	4	A
Jorge Luis Charvet Castro	6	A
Kira Ryūzaki	6	A
Macías Vélez Andrius	4	A
Tardo Estebas	4	A

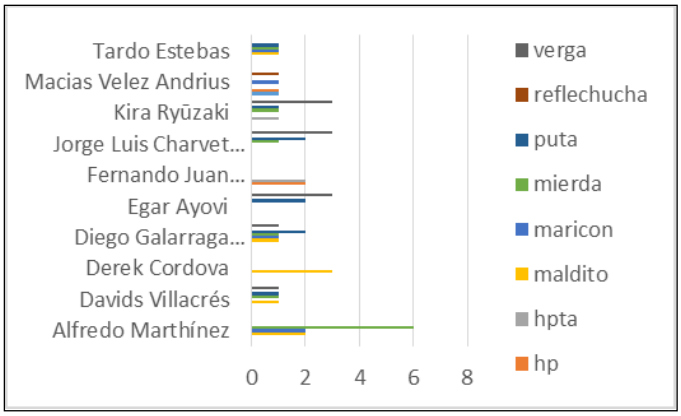

Fig. 5. Individuos agresivos vs palabras comunes

V. TRABAJOS RELACIONADOS

En la investigación de WebCrowling [28] se evidencias técnicas y modelos matemáticos para extracción de datos en sitios web con el objetivo de analizar la información y tener una percepción del contenido. A esto se le conoce como patrones de opinión que se utilizan en el marketing para ofrecer productos a consumidores. También, existen otras investigaciones como análisis sentimental [29], que consiste en analizar los comentarios de las personas para determinar la parte subjetiva de los individuos, cuyos fines son la obtención de perfiles de consumidores.

VI. CONCLUSIONES Y TRABAJOS FUTUROS

Como conclusiones generales cabe establecer que al explorar la información, después del proceso de clasificación, se encontró con niveles altísimos de insultos en gran mayoría de los comentarios, tal es el caso de la Tabla 6, donde se demuestra que existen un 42,41% de personas que utilizan palabras ofensivas para expresarse. Además, la gran mayoría de las palabras son agresivas, para ser específicos un 65% usa palabras de nivel de agresividad mediano mientras que otra parte conformando por el 18% utiliza frase con nivel de agresividad alto y tan solo el 16% de los individuos no son tan violentos.

Este trabajo de investigación demuestra que de 2791 usuarios, el 65% de ellos tiene un vocabulario medio agresivo, el 18% un vocabulario agresivo y el 16% poco agresivos. Por lo tanto, se puede inferir que el 65% de las personas que publica un mensaje en fan pages de política, son medianamente agresivos y tiene un comportamiento medio violento.

Como trabajo futuro se planea implementar una herramienta en tiempo real, que interactúe directamente con la red social Facebook y mediante una "app" embebida en Facebook, determine si es adecuado o no aceptar la solicitud de un desconocido. Además, las aplicaciones de este tipo de minería de texto tienen una gran utilidad a nivel de marketing porque se podría obtener perfiles de consumidores.

REFERENCIAS

[12] C. Domínguez,. "Las Redes Sociales. Tipología, uso y consumo de las redes 2.0 en la sociedad digital actual." Documentación de las Ciencias de la Información 33 (2010): 45-68 [online] http://revistas.ucm.es/index.php/DCIN/article/view/DCIN1010110045A/18656.

[13] Phillips, Whitney. "LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online." First Monday 16.12 (2011).

[14] Levenshtein VI (1966). "Binary codes capable of correcting deletions, insertions, and reversals".

[15] Fernando Juárez, Alba Dueñas & Yamilé Méndez, 17-05-2005, ISSN 1697-2600, [online] http://www.redalyc.org/pdf/337/33760108.pdf

[16] Mauricio Batallas Bustamante, "Agresividad, Hostilidad e Ira en adolescentes que juegan video juegos." ,2014:19-24 [online] http://dspace.udla.edu.ec/bitstream/33000/3441/1/UDLA-EC-TPC-2014-04(S).pdf

[17] K. Cela, W. Fuertes, C. Alonso, F. Sánchez, , Revista Estilos de Aprendizaje, nº5, Vol. 3, 04-2010, [online] http://learningstyles.uvu.edu/index.php/jls/article/view/123/86.

[18] M. Gonzalez, "Patrones de Comportamiento", 28-09-2010, ISBN 978-84-9717-323-0, [online] http://www.elcampamentodedios.com/28sep10b.pdf .

[19] M. Montes, M. Gómez, A. Gelbukh, and A. López López. "Minería de texto empleando la semejanza entre estructuras semánticas." Computación y Sistemas 9.1 (2005): 63-81.

[20] Aggarwal, Charu C., and ChengXiang Zhai. Mining text data. Springer Science & Business Media, 2012.

[21] Elder IV, John, and Thomas Hill. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, 2012.

[22] Buneman, Peter, and Robert E. Frankel. "FQL: a functional query language." Proceedings of the 1979 ACM SIGMOD international conference on Management of data. ACM, 1979.

[23] "Diccionario de palabras ofensivas",2016. [online] https://1drv.ms/f/s!AvddUrAdljaYgmMi5eCx34HydZh0

[24] Wilfredo Martel, "Algoritmo de Ordenación y Clasificación de palabras ofensivas", 2016. [online] https://1drv.ms/f/s!AvddUrAdljaYgmbDEokfXJZmr4hq

[25] "Resultados de la investigación", 2016. [online] https://1drv.ms/x/s!AvddUrAdljaYgmKnrf_ZNcGS1e76

[26] "Algoritmo de limpieza de datos", 2016. [online] https://1drv.ms/f/s!AvddUrAdljaYgmh9ie20r4RPjL34

[27] Wilfredo Martel, "Video del proceso de extracción y análisis de datos", 2015.[online] https://1drv.ms/f/s!AvddUrAdljaYgmowCe7RzxynL2vp

[28] http://www.palermo.edu/ingenieria/pdf2013/12/12CyT_01webcrawling.pdf.

[29] J,Akaichi, Z. Dhouioui, M. J. Lopez-Huertas Perez. "Text mining Facebook status updates for sentiment classification,", System Theory, Control and Computing (ICSTCC), 2013 17th International Conference, 10- 2013 ,ISBN 978-1-4799-2227-7.