BENEMÉRITA

UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DELA COMPUTACIÓN

SISTEMA DE CLASIFICACIÓN DE TEXTOS CORTOS QUE CONTIENEN DOBLE SENTIDO DENTRO DE REDES SOCIALES

TESIS PARA OBTENER EL TÍTULO DE:

INGENIERA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

ESTEFANIA GUZMAN FALCON

ASESORES:

MC. BEATRIZ BELTRÁN MARTÍNEZ

DRA. MIREYA TOVAR VIDAL

H. PUEBLA DE ZARAGOZA, JULIO 2015

Resumen

En los últimos años la tecnología ha crecido a pasos agigantados, cambiando de manera paulatina el estilo de vida de cada individuo, incluyendo su manera de interactuar y comunicarse con otros individuos. En la actualidad uno de los medios de comunicación más utilizados son las redes sociales, las cuales se utilizan para ponerse en contacto con familiares o amigos distantes, hacer nuevas amistades o incluso comunicarse con individuos con los que ya se tiene interacción diaria por otros medios.

Miles de mensajes circulan diariamente en las redes sociales y no existe garantía de que la gente realmente entienda el contenido de estos mensajes ya que puede darse el caso de que ocurran malinterpretaciones debido a un doble sentido que pudiera estar presente en el mensaje y no ser detectado por el lector por factores de lenguaje, contexto o geográficos .

En México, existe un humor característico que se utiliza para situaciones de la vida cotidiana, y es frecuente que dicho humor se utilice para darle doble sentido a las palabras o frases. Para algunas personas que no entienden un albur o doble sentido, puede llevarlas a una confusión y malinterpretación del mensaje y esto suele pasar de manera frecuente en redes sociales, donde hay que saber distinguir que es lo que realmente las personas quieren expresar.

Hoy en día uno de los principales modos de comunicación sobre todo entre los más jóvenes, es a través de twitter. Para los especialistas en procesamiento de Lenguaje Natural se hace necesario estudiar si en este tipo de mensajes se utiliza mucho el albur o no, y detectar cuales palabras generalmente obscenas y vulgares son las más empleadas.

En la presente trabajo se propone una metodología para detectar el albur en los tweets a partir del uso de un recurso léxico que contiene palabras que el mexicano usa con doble sentido. Se toma como datos de entrada un conjunto de tweets de la República Mexicana, y se realiza un análisis estadístico que permite identificar los estados de la república que más emplean el doble sentido, tanto con sentido vulgar, como obsceno.

El método utilizado fue realizar un modelo de clasificación el cual permitiera identificar que elementos del albur contiene un tweet, puede ser el caso de elementos vulgares u obscenos, obteniendo como resultado cuantas instancias estaban clasificadas correctamente y los resultados fueron alentadores por la cantidad de elementos clasificados.

Dedicatoria

Dedico esta tesis a mis padres Raúl Guzmán González y Eréndira Falcón Cazares por su apoyo, consejos y ánimos para que terminara mis estudios universitarios.

A mis abuelos Raúl Guzmán Cruz y María de Lourdes González Vera por su cariño, apoyo, motivación, gracias por estar a mi lado y verme como una hija.

A mi mejor amiga Rocio Galaviz Huerta, gracias por estar a mi lado estos 5 años de carrera y por compartir nuestras frustraciones y sueños juntas.

A Carlos Daniel Lima Romero por apoyarme a crecer como profesionista y persona, gracias por la motivación y cariño.

Agradecimientos

A mis asesoras: M.C. Beatriz Beltrán Martínez y a la Dra. Mireya Tovar Vidal por el apoyo y orientación para este trabajo de tesis.

Agradezco el apoyo incondicional de mis amigos en cada etapa de mi formación académica, gracias por las experiencias y conocimiento que compartimos.

Agradezco el apoyo al cuerpo académico de la Facultad de Ciencias de la Computación por la beca concedida, así como también a fundación TELMEX por la beca, para la culminación de esta tesis y carrera.

Índice general

Índice de figuras

Índice de Tablas

Introducción

El mexicano es característico por el humor que le pone a distintas situaciones sean buenas o malas, pero es mejor conocido por darle el doble sentido a palabras o frases. Para algunas personas que no entienden un albur, puede llevarla a una confusión y esto suele pasar mucho en redes sociales, donde hay que saber distinguir que es lo que realmente las personas quieren expresar, es por ello que se toma como motivación el análisis de tweets porque twitter es una de las redes sociales más usadas y puede ser de utilidad para hacer un análisis de lenguaje natural para detectar albures.

En el presente trabajo consiste en el uso de técnicas de procesamiento del lenguaje natural (PLN) para detección del albur en tweets. Inicialmente, se realiza una investigación sobre los trabajos relacionados con el albur o doble sentido. Además de una investigación sobre los conceptos básicos de procesamiento de lenguaje natural que servirán para el desarrollo de esta tesis. La finalidad es obtener un panorama para posteriormente como trabajo a futuro construir un sistema que analice tweets y permita definir si el mismo involucra el doble sentido o no.

Para esto se pretende reunir una cantidad aceptable de tweets de toda la república Mexicana para la fase de pruebas y evaluación. Para esto se hará uso de recursos léxicos, tales como diccionarios que contienen palabras que el mexicano usa con doble sentido, como es el caso del Diccionario de Mexicanismos editado por la academia mexicana de la lengua.

El objetivo principal es encontrar el sentido real de un tweet que permitirá a las personas comprenderlos. Para finalizar el estudio, se realiza un análisis estadístico que permite identificar los estados de la república que más emplean el doble sentido.

Para la realización de este proyecto se ha establecido el objetivo general:

“Diseñar un modelo de clasificación que sea capaz de procesar lenguaje natural para identificar el albur en tweets, mediante el desarrollo de un diccionario de palabras con albur y así generar reportes y estadísticas de los tweets analizados”.

Para llevar a cabo esta tarea se definen también los siguientes objetivos específicos:

1. Examinar las distintas técnicas de procesamiento del lenguaje natural.
2. Investigar los trabajos realizados con el manejo del albur mexicano.
3. Analizar el sistema, en base a las técnicas de procesamiento natural, para extraer tweets que contengan albur para formar los datos de entrenamiento y los de prueba del sistema.
4. Diseñar un modelo de clasificación.
5. Realizar pruebas en base a un conjunto de tweets.

La arquitectura de solución al problema, es utilizando técnicas de PLN y aprendizaje automático. Como parte del diseño del modelo de clasificación, detectar los puntos de mayor complejidad para análisis posteriores como el semántico.

La tesis se encuentra conformada de 4 capítulos distribuidos de la siguiente manera:

En el capítulo 1 se presenta el marco teórico el cual contiene descripción de algunos conceptos como el albur, locuciones verbales, etc. Así como también descripción de herramientas utilizadas en el presente trabajo, en el capítulo 2 se presenta el estado del arte, que son los antecedentes, los cuales introducen a la detección de humor en textos, en el capítulo 3 se presenta los enfoques propuestos que se utilizaron para este trabajo y bajo que metodologías previas, en el capítulo 4 se presentan las pruebas realizadas y los resultados obtenidos y por último se presentan las conclusiones y la bibliografía consultada para la realización del trabajo de tesis.

Capítulo 1

Marco teórico

En este capítulo se presenta la teoría en la cual se basa el presente trabajo, con el objetivo de dar a conocer algunos elementos que son parte fundamental.

1.1 ¿Qué es el lenguaje humano?

Para definir que es el lenguaje humano existen una variedad de definiciones, para expresar un concepto claro y concreto se tomará en cuenta dos:

La primera según la RAE[1], el Lenguaje se define como: el estilo o modo de hablar y escribir de cada persona en particular. Esta definición se describe elementos del lenguaje en la comunicación oral y escrita.

La segunda: El lenguaje es la capacidad humana adquirida por la que se comunican contenido a través de la palabra, oral o escrita, cualquier conjunto de signos que sirvan a un grupo humano para intercambiar mensajes. En esta última se describen elementos del lenguaje como son la interacción y su función [1].

En resumen el lenguaje es la capacidad del ser humano para comunicarse con otros a través de la escritura o lenguaje, que contiene cierto estilo para su comprensión.

Los seres humanos requieren de él para comunicarse, pero las preguntas son: ¿Es efectiva la manera de comunicación actualmente? ¿Cómo evolucionó hasta lo que ahora se conoce?

La lingüística que es la disciplina encargada del estudio del lenguaje, en épocas anteriores se limitaba a estudiar la estructura, gramática y puntuación. Pero ahora está consciente que el lenguaje se ve involucrado con elementos sociales y culturales.

A través del tiempo el lenguaje ha sufrido bastantes cambios en cada época, cambiando su estructura, agregando nuevos significados a las palabras u expresiones. Actualmente en la época digital se ha ido alterando la forma que se comunican las personas y a veces es complejo interpretar lo que se quiere decir si no se plantea adecuadamente el contexto. Cuando se aprende un nuevo idioma se cree que es complejo, pero es una idea errónea, debido a que se debe estar consciente siempre del contexto para tener una comunicación efectiva en un nuevo idioma. Se plantea la idea de contexto como la situación, circunstancia o escenario [3], donde se expresa el mensaje a transmitir. Este es construido a partir de las relaciones humanas, de estas relaciones el ser humano aprende a construir oraciones gramaticalmente correctas.

El idioma Español no es tan complejo como muchos lo ven, pero a veces el contexto de algunas palabras o expresiones hacen que parezca todo un reto aprenderlo. Ese caso aplica para el Español de México, cada oración o palabra por lo regular maneja un doble sentido.

[1] Diccionario de la Real Academia Española. http://www.rae.es/

Cuando una palabra u oración tiene doble sentido, este comportamiento es conocido como polisemia, donde estructuras gramaticales, sintácticas, semánticas y pragmáticas del lenguaje pueden ser alteradas, para eso es necesario conocer el contexto en que se está empleando, para entender el significado correcto y no exista un problema de comunicación.

En el lenguaje existe un componente fonético que accede que este tenga orden lógico para ser entendido y un componente semántico que permite que haya correspondencia entre significado y significante. Se atribuye significancia a un significado y entra en el juego de comunicaciones a partir de conductas simbólicas.

La relación significado-significante no es siempre la misma, las cosas no siempre quieren decir lo mismo. El lenguaje tiene un carácter polisémico y éste depende del lugar que ocupa el significante con respecto a otros significantes. Por lo tanto, el significante es el que le da sentido al lenguaje. Este se desarrolla a partir de la realidad por lo que esta debe ser representada mediante metáforas y metonimias.

1.2 Metáfora

La metáfora es un término polisémico referente a trasladar el sentido recto de las voces a otro figurado, “aplicación de una palabra o de una expresión a un objeto o a un concepto, al cual no denota literalmente, con el fin de sugerir una comparación (con otro objeto o concepto) y facilitar su comprensión” [6].

Una metáfora se genera por lo regular, al cambiar una palabra común a un contexto determinado por otra poco usual.

Para la comprensión de esta definición un ejemplo, en la frase: “*El inicio de la vida*”, se cambia ‘*inicio*’ por la palabra ‘*primavera*’, obteniendo “La primavera de la vida” [8].

En el ejemplo anterior se observa que primavera no se está empleando como la primera estación del año, sino una figuración de inicio, que en este contexto se refiere al comienzo de algo.

1.3 Metonimia

La metonimia en la literatura se refiere a la práctica de no utilizar la palabra formal de un objeto o sujeto y en lugar referirse a ello mediante el uso de otra palabra que está íntimamente ligada al nombre formal. Es la práctica de sustituir la palabra principal con una palabra que está estrechamente vinculada a ella [7].

Por ejemplo: La frase “Es el dueño absoluto de esos lares”, la palabra “lares” tiene el mismo significado que lugares, pero el empleo de esta es informal.

1.4 Locuciones verbales

Existen varias definiciones para estas unidades fraseológicas, para entender su función y de describir su carácter. Pero, antes de que se expongan algunas definiciones, se debe decir que para las locuciones también se han empleado otros términos como por ejemplo modismos, frases hechas o expresiones fijas [26].

La lingüista L. Ruiz Gurillo las define como:

"[...] sintagmas fijos que en ciertos casos presentan idiomaticidad"[27].

Esta autora considera como rasgos generales de las locuciones la fijación e idiomaticidad. Según ella la locución es, en primer lugar, un sintagma fijo y ocasionalmente la fijación viene acompañada de idiomaticidad, de modo que ambas se complementan.

A continuación se citaran las definiciones de los distintos diccionarios de la lengua española.

El Diccionario de la Real Academia Española las define de la siguiente manera:

Combinación estable de dos o más palabras, que funciona como oración o como elemento oracional, y cuyo sentido unitario no siempre se justifica, como suma del significado normal de los componentes.[2]

En el Gran diccionario de Uso del español actual aparece esta definición:

Combinación fija, formada por dos o más palabras, que componen un elemento oracional unitario y cuyo significado no se desprende necesariamente de los significados individuales de sus componentes.[3]

Por último se expone la definición del Diccionario de María Moliner:

Expresión pluriverbal de forma fija que se inserta en el habla como una pieza única, constituida por una oración simple o compuesta o una parte de oración.[4]

En resumen, se puede deducir de las definiciones precedentes que la fijación es el rasgo más importante de estas unidades. Cabe aclarar de la fijación de dos o más palabras. La fijación significa que un sintagma se reproduce siempre del mismo modo y que no admite grandes variaciones en su estructura.

1.4.1 Locución verbal

Las locuciones verbales están constituidas por un núcleo verbal, acompañado por sus complementos. Según G. Corpas Pastor presentan una gran diversidad morfosintáctica y a este tipo de las locuciones engloba los siguientes compuestos [25]:

- dos verbos + conjunción (pueden llevar también complementos) *ir y venir, dar y tomar*

[2] Real Academia Española (1992): Diccionario de la lengua española. 21. edición. Madrid, RAE.
[3] Sánchez, A. (2001): Gran diccionario de uso del español actual. Alcobendas-Madrid, Sociedad General Española de Librería.
[4] Moliner, M. (1986): Diccionario de uso del español. Madrid, Editorial Gredos

- verbo + pronombre *apañárselas*
- verbo, pronombre + partícula *tomarla con*
- verbo + partícula *asociada a éste dar de sí*
- verbo copulativo + atributo *ser el vivo retrato de alguien*
- verbo + complemento circunstancial *dormir como un tronco*
- verbo + suplemento *oler a cuerno quemado*
- verbo + objeto directo *costar un ojo de la cara*
- gran parte presenta fijación fraseológica en negativo *no tener dos dedos de frente*

Desde el punto de vista sintáctico expresan procesos y actúan como los predicados, con o sin complementos.

1.5 El Albur

En algún lugar, circunstancia o reunión cualquier persona que vive en México ha oído un albur, que puede causar gracia u ofensa dependiendo del contexto en que fue dicho y si fue entendido el mensaje "oculto" que este contiene. Pero que es realmente un albur, porque para algunos es un arte lingüístico que no cualquiera puede entender.

El albur es una manifestación de la cultura popular que contiene connotaciones sexuales que suelen ser vulgares y obscenas. Este se realiza a base de expresiones de doble sentido que aparenta manifestar una idea anodina e inocua [2].

A través de los años el albur ha ido trascendiendo las clases sociales y géneros, debido a que anteriormente era empleado únicamente por hombres que tenían trabajos medios como albañiles, mecánicos, choferes, etc.

Saber aplicar el albur, en pocas palabras "Alburear" requiere tener una habilidad verbal, ser experto en las palabras, darles la vuelta, modificarles el significado, torcerles la intención [5], para la manifestación de un doble sentido, este mensaje que por lo general alude al acto sexual y funciones del cuerpo. Pero no únicamente la persona que alburea debe tener esa capacidad sino también el receptor del este mensaje.

1.6 Vulgar y Obsceno

Las definiciones de vulgar y obsceno, según la RAE:

Vulgar[5]

Se define como vulgar a frases, palabras o expresiones; las cuales son groseras u ordinarias siendo catalogadas como impropias para personas cultas o educadas.

Obsceno[6]

[5] http://lema.rae.es/drae/srv/search?key=vulgar

Se define como obsceno a frases, palabras o expresiones; las cuales son impúdicas u ofensivas al pudor y en la mayoría de los casos con relación al sexo.

Ambas se denotan como groserías y que no son apropiadas para comunicarse en un ámbito formal, y su aplicación es en el contexto informal para expresar dichos, albures, chistes, etc.

Pero en resumen para diferenciar si una frase es vulgar u obscena se debe identificar primero si es un tipo de ofensa al pudor para descartar que sea obsceno y si no cuenta con ello se clasifica como vulgar.

1.7 Twitter[7]

Twitter es una red social que permite a los usuarios enviar y leer mensajes cortos de 140 caracteres llamados "tweets". Twitter fue creado en marzo de 2006 por Jack Dorsey, Evan Williams, Biz Stone y Noah Glass [24]; poniéndose en marcha en julio de 2006. El servicio rápidamente ganó popularidad en todo el mundo, con más de 100 millones de usuarios que en 2012 publicaron alrededor de 340 millones de tweets por día.

Para visualizar el contenido de twitter, los usuarios registrados pueden leer y publicar tweets, pero los usuarios no registrados sólo pueden leerlos. Los usuarios acceden a Twitter a través de la interfaz web, SMS, o aplicación para dispositivo móvil.

Actualmente twitter cuenta con 288 millones de usuarios activos mensualmente que envían alrededor de 500 millones de tweets por día y el 77% de las cuentas no pertenecen a usuarios de los EE.UU. [22].

La mecánica de twitter funciona en que cada usuario puede, tener una lista de "seguidos" (*following*) y de "seguidores" (*followers*). Los "seguidores" leerán los textos publicados por el "seguido" en sus páginas personales. Cuando se sigue a personas, sus Tweets se muestran instantáneamente en la cronología. Del mismo modo, los Tweets que uno publique se muestran en las cronologías de los seguidores. Se pueden seguir: amigos, celebridades, noticias o cualquier.

En cuestión de mensajes, un Tweet es una expresión de un momento o idea. Puede contener texto, fotos y videos. Los Tweets se comparten en tiempo real, todos los días.

[6] http://buscon.rae.es/drae/srv/search?id=UL47Ue02uDXX2j7Xanvv
[7] https://twitter.com/

Figura 1.1 Forma de un tweet

Cuando se escribe un tweet y se publica, quien lo lee puede realizar una serie de actividades con ese tweet, como responder (comentar un tweet), retwittear (Compartir un Tweet con seguidores e incluso añadir pensamientos antes de compartirlo), marcar como favorito un Tweet y hacerle saber al autor te gusta, como se muestra en la Figura 1.1 estas actividades en el tweet.

Lo anterior sólo es una parte de la funcionalidad de twitter, ya que por medio de esta red social se comparte mucha información que no simplemente es visualizada sino también categorizada para tener un acceso ordenado a esta, para ello se utilizan los famosos hashtags, que asignan un tema a un tweet.

La técnica de clasificación de twitter es el uso de etiquetas que relacionen el texto con un tema determinado. Las etiquetas (llamadas hashtags) van precedidas del símbolo "#", facilitando su identificación desde el buscador.

De esta forma se puede escribir y encontrar textos escritos de la siguiente forma:

“Infórmate de las principales #noticias sobre #educación en #México en…”, indicando que el texto hace referencia a noticias sobre educación en México. Si queremos encontrar todos los mensajes recientes relacionados con México, por ejemplo, podremos buscar usando dicha etiqueta en search.twitter.com filtrando también por el idioma en que ha sido escrito, obteniendo fácilmente el resultado. Ya que no todo el mundo clasifica sus mensajes, por lo regular se encontraran más resultados buscando por "México" que por "#México".

Las etiquetas más utilizadas en Twitter aparecen siempre en el menú lateral de la página, lo que ayuda a identificar rápidamente cuáles son los temas más comentados por los usuarios. Generalmente hacen referencia a grandes eventos internacionales, como elecciones, accidentes, atentados, estrenos de cine, etc.

1.8 TwitterSearch[8]

TwitterSearch fue desarrollado como parte de un proyecto sobre las redes sociales en el Carl von Linde-Akademie, una institución de la Technische Universität München[9]. Por lo tanto, TwitterSearch es un conjunto de herramientas de recolección de datos y no está implementando toda la API de Twitter, pero si la API de búsqueda y la API de línea de tiempo del usuario. La biblioteca es totalmente accesible a través del repositorio oficial en github y mantenido por Christian Koepp.

Esta biblioteca permite crear fácilmente una búsqueda sin tener que saber demasiado sobre los detalles de la API. Sobre la base de tal búsqueda puede incluso iterar a través de todos los tweets alcanzables de búsqueda de Twitter.

TwitterSearch está utilizando la API REST[10] sólo en la versión 1.1. En su versión reciente, la biblioteca está utilizando identificadores de tweets para navegar a lo largo de la lista disponible de tweets. Si lo hace, permite una iteración más flexible y eficiente que el método tradicional de usar páginas. También, TwitterSearch se construye para ser muy flexible en su uso por lo que es utilizable incluso en casos de uso exóticos. Todas las clases y sus métodos son probados contra las últimas versiones de Python2 y Python3 automáticamente. El estado actual de todas las ramas es visible a través de Travis CI[11].

1.8.1 Arquitectura

TwitterSearch consta de cuatro clases en Python para su funcionamiento: TwitterSearch, TwitterSearchOrder, TwitterUserOrder y TwitterSearchException [26].

TwitterSearch

Esta clase contiene la funcionalidad real de esta librería. Es responsable de transmitir correctamente sus datos a la API de Twitter y devolver los resultados a su programa después. Se configura mediante una implementación de TwitterOrder junto con las credenciales de Twitter válidos. Actualmente dos implementaciones diferentes son utilizables: TwitterUserOrder para recuperar la línea de tiempo de un determinado usuario y TwitterSearchOrder para acceder a la API de búsqueda de Twitter [26].

TwitterSearchOrder

Esta clase es para configurar todos los argumentos disponibles de la API de búsqueda de Twitter. También crea cadenas de consulta válidos que se pueden usar en otros entornos idénticos a la sintaxis de la Twitter API de búsqueda [26].

[8] https://github.com/ckoepp/TwitterSearch
[9] http://www.tum.de/
[10] Un tipo de arquitectura de desarrollo web que se apoya totalmente en el estándar HTTP
[11] https://travis-ci.org/recent

TwitterUserOrder

Esta clase configura todos los argumentos disponibles del punto final user_timeline de la API de Twitter. También crea una cadena de consulta válida de la configuración actual [26].

TwitterSearchException

Esta clase tiene que ver con excepciones. Toda excepción basada directamente en TwitterSearch consistirá en un código y un mensaje que describa el motivo de la excepción en breve.

1.9 WEKA

Weka es una colección de algoritmos para el aprendizaje automático y minería de datos. Los algoritmos bien se pueden aplicar directamente a un conjunto de datos o ser invocados desde su propio código Java. Weka contiene herramientas para los datos pre-procesados, para su clasificación, regresión, *clustering*, reglas de asociación, y la visualización. También es muy adecuado para el desarrollo de nuevos esquemas de aprendizaje automático.

Sus principales ventajas residen en la zona de clasificación, porque se han implementado dentro de una jerarquía de clases en Java orientado a objetos. Regresión, reglas de asociación y algoritmos de agrupamiento también se han implementado.

1.9.1 Clasificador

Cualquier algoritmo de aprendizaje en WEKA se deriva de la clase abstracta clasificador. Se necesita muy poco para un clasificador básico: una rutina que genera un modelo clasificador de una formación de datos (= *buildClassifier*) y otra rutina que evalúa el modelo generado en un conjunto de datos de prueba que no se ve (= *classifyInstance*), o genera una distribución de probabilidad para todas las clases (= *distributionForInstance*).

Un modelo clasificador es un mapeo complejo arbitrario de todo, pero un conjunto de datos que se atribuye al atributo de clase. La forma específica y la creación de esta asignación, o modelo, se diferencia de clasificador a clasificador. Por ejemplo, el modelo de Zeror sólo consta de un único valor: la clase más común, o la mediana de todos los valores numéricos en el caso de la predicción de un valor numérico (= aprendizaje de regresión). Zeror es un clasificador trivial, pero da un límite inferior en el rendimiento de un determinado conjunto de datos que deberían mejorarse significativamente por más clasificadores complejos. Como tal, es una prueba razonable de lo bien que la clase se puede predecir sin considerar los otros atributos.

Existen varios enfoques para determinar el rendimiento de los clasificadores. El rendimiento de la mayoría simplemente se mide contando la proporción de ejemplos predichos correctamente en un conjunto de datos de prueba que no se ve. Este valor es la precisión, que es también 1-ErrorRate. Ambos términos se utilizan en la literatura.

El caso más simple está utilizando un conjunto de entrenamiento y un conjunto de pruebas que son independientes entre sí. Esto se conoce como estimación de retención de salida. Para la estimación de la variación en las estimaciones de rendimiento, las estimaciones de cautividad a cabo se pueden calcular por re-muestreo repetidamente el mismo conjunto de datos, es decir, al azar reordenar y luego dividirlo en capacitación y de prueba con una proporción específica de los ejemplos, recogiendo todas las estimaciones sobre datos de prueba y calculando la media y desviación estándar de precisión.

Un método más elaborado es la validación cruzada. Aquí, no se especifica un número de pliegues. El conjunto de datos se reordena aleatoriamente y después se divide en n pliegues de igual tamaño. En cada iteración, un conjunto se utiliza para la prueba y los otros n-1 pliegues se utilizan para entrenar el clasificador. Los resultados del ensayo se recogen y se promedian sobre todos los pliegues. Esto le da a la estimación de validación cruzada de la exactitud.

Los pliegues pueden ser puramente aleatoria o ligeramente modificado para crear las mismas distribuciones de clase en cada pliegue como en el conjunto de datos completo. En el último caso, la validación cruzada se llama estratificada. Deja una validación cruzada (loo), significa que n es igual al número de ejemplos. Por necesidad, cv loo tiene que ser no estratificado, es decir, las distribuciones de clase en el equipo de prueba no están relacionados con los de los datos de entrenamiento. Por lo tanto loo cv tiende a dar resultados menos fiables. Sin embargo, es todavía muy útil en el tratamiento de pequeños conjuntos de datos, ya que utiliza la mayor cantidad de datos de entrenamiento del conjunto de datos.

1.9.2 Algoritmos de Clasificación

A continuación se enlistan algunos de los algoritmos de clasificación más utilizados en WEKA.

Lazy.IBk

Instancia basada en aprendizaje con un conjunto arreglado (Barrio). -K Establece el número de vecinos para su uso. IB1 es equivalente a IBK -K 1.

Functions.SMO

Máquina de soporte vectorial (lineal, kernel polinómico y RBF) con el algoritmo secuencial mínimo de optimización. Por defecto es SVM con kernel lineal, -E 5 -C 10 da una SVM con kernel polinomio de grado 5 y lambda = 10.

Trees.J48

Es un árbol de decisión, la perspectiva es en la raíz del árbol y determina la primera decisión. Va clasificado cada ejemplo en una hoja, cada hoja contiene la cantidad de elementos que fueron clasificados en ella.

1.9.3 Medidas de WEKA

Son las medidas devueltas por WEKA:

Accuracy

En los algoritmos de clasificación, indica el número o el porcentaje de ejemplos clasificados correctamente de todas las clases [35].

False Positive Rate (FP_Rate)

En los algoritmos de clasificación es la proporción de ejemplos que fueron clasificados en la clase x, pero pertenecen a una clase diferente, entre todos los elementos que no pertenecen a la clase x [35].

Precision

En los algoritmos de clasificación, es la proporción de ejemplos que realmente pertenecen a la clase x entre todos aquellos que fueron clasificados como pertenecientes a la clase x [35].

Recall

En los algoritmos de clasificación, es la proporción de ejemplos que fueron clasificados correctamente en la clase x, entre el total de ejemplos que realmente pertenecen a la clase x. Es similar a True Positive Rate [35].

True Positive Rate (TP_Rate)

En los algoritmos de clasificación, es la proporción de ejemplos que fueron clasificados correctamente en la clase x, entre el total de ejemplos que realmente pertenecen a la clase x. Es similar a Recall [35].

1.9.4 Validación Cruzada

Traducción de *Cross Validation*. Es un método utilizado por WEKA cuando no se le pasan dos archivos como parámetros. Para mandar llamar a las funciones de la aplicación WEKA, se requiere, como parte de los parámetros, que se especifiquen dos archivos de extensión .arff, uno de los cuales será utilizado como instancias de entrenamiento y el otro como instancias de prueba. Para el caso de que sólo se le especifique un solo archivo, la aplicación hará automáticamente el uso de la Validación Cruzada por 10, el cual consiste en reordenar aleatoriamente el conjunto de datos y dividirlos en 10 pliegues de igual tamaño. En cada iteración, un solo pliegue se utiliza para las pruebas y los restantes 9 pliegues se utilizan para el entrenamiento del clasificador. Los resultados de las pruebas se recogen y se promedian con los demás pliegues. Esto le da a la Validación Cruzada una estimación de la precisión en cuanto a la clasificación [35].

1.10 Procesamiento de Lenguaje Natural (PLN)

1.10.1 Definición de PLN

Procesamiento del Lenguaje Natural (PLN), también llamada Lingüística Computacional, tiene como objetivo desarrollar métodos computacionales y algoritmos para la comprensión y la

generación de las lenguas humanas [12]. Existe una clara delimitación entre el campo que se ocupa de la comprensión del lenguaje, que se llama comprensión del lenguaje natural, por sus siglas en inglés (*NLU: Natural Language Understanding*), y el campo que se ocupa de cuestiones con respecto a la generación o producción de lenguaje natural, que se llama Generación de Lenguaje Natural, por sus siglas en inglés (*NLG: Natural Language Generation*).

El campo de la PLN ha sido dominado por el campo NLU desde sus inicios en la década de 1950 hasta la actualidad, con pocas perspectivas de una representación más equilibrada en el futuro cercano entre campos de NLU y NLG. El predominio del campo NLU se explica por lo menos dos factores. En primer lugar, ha habido un deseo en el dominio de la informática para hacer que las computadoras entiendan el lenguaje humano. En segundo lugar, la generación de lenguaje se percibe con más fuerza, desde una perspectiva, que la comprensión del lenguaje [12].

Las aplicaciones del PLN son variadas, ya que su alcance tiene gran impacto son:

- Traducción automática
- Recuperación de la información
- Extracción de Información y Resúmenes
- Resolución cooperativa de problemas
- Tutores inteligentes
- Reconocimiento de Voz

Para la creación de un sistema PLN, se debe contar con una arquitectura la cual se sustenta en una definición del LN por niveles: fonológico, morfológico, sintáctico, semántico, y pragmático. Esta arquitectura se muestra en la Figura 1.2.

- **Nivel Fonológico:** Es la relación de las palabras con los sonidos que representan.
- **Nivel Morfológico:** Es la construcción de las palabras a partir de unas unidades de significado más pequeñas llamadas morfemas.
- **Nivel Sintáctico:** Es la unión de las palabras para formar oraciones, fijando el papel estructural que cada palabra juega en la oración y que sintagmas son parte de otros sintagmas.
- **Nivel Semántico:** Es el dignificado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir de la oración aislada.
- **Nivel Pragmático:** Este nivel trata de cómo las oraciones se usan en distintas situaciones y cómo el uso afecta al significado de las oraciones. Se reconoce un subnivel recursivo: discursivo, que trata de cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores [13].

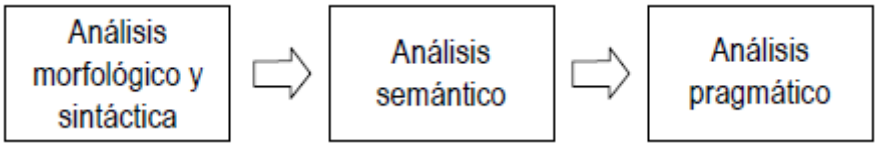

Figura 1.2 Arquitectura de sistema PLN

1.10.2 Niveles de análisis (palabras)

Análisis Léxico

Este análisis consiste en la creación de un vocabulario o diccionario que está conformado por palabras de un idioma o pertenecientes al uso en una región, a una actividad determinada, a un campo semántico dado, etc.

Una vez realizado ese diccionario se procede a hacer un análisis de frecuencia de éste en un texto determinado según el sentido para proceder con una clasificación del texto [9].

Análisis Sintáctico

El análisis sintáctico es, en el campo de la Lingüística, el análisis de las funciones sintácticas o relaciones de concordancia y jerarquía que guardan las palabras agrupándose entre sí en sintagmas u oraciones. Sin embargo a veces no está claro el límite entre la sintaxis y la morfología a estos, especialmente según el tipo de lengua de que se trate, también se suele denominar análisis morfosintáctico, aunque esta denominación se suele reservar para un análisis más profundo y detenido [9].

Análisis Morfológico

La morfología estudia la estructura de las palabras y su relación con las categorías gramaticales del lenguaje. El objetivo del análisis morfológico automático es llevar a cabo una clasificación morfológica de una forma de palabra.

Los lenguajes según sus características morfológicas dependen de la tendencia en la manera de combinar los morfemas que se clasifican en aglutinativos y flexivos [10].

Se dice que un lenguaje es aglutinativo si:

- Cada morfema expresa un sólo valor de una categoría gramatical.
- No existen alternaciones de raíces o las alternaciones que cumplen con las reglas morfonológicas que no dependen de la raíz específica, como, por ejemplo, armonía de vocales, etc.
- Los morfemas se concatenan sin alteraciones.
- La raíz existe como palabra sin concatenarse con morfemas adicionales algunos.

Algunos lenguajes son aglutinativos como el turco o el húngaro.

Y un lenguaje es flexivo si:

- Cada morfema puede expresar varios valores de las categorías gramaticales. Por ejemplo, el morfema *mos* en español expresa cumulativamente los valores de las categorías persona (tercera) y número (plural).
- Alternaciones de raíces no son previsibles, sin saber las propiedades de la raíz específica no se puede decir qué tipo de alternación se presentará.
- Los morfemas pueden concatenarse con ciertos procesos morfonológicos no estándares en la juntura de morfemas.
- La raíz no existe como palabra sin morfemas adicionales (por ejemplo, *escrib* no existe como palabra sin -ir, -iste, -ía, etc.).

Los lenguajes reflexivos son lenguas eslavas (ruso, checo, etc.) o románicas (español, portugués, etc.) [10].

Análisis Semántico

El Análisis Semántico es el proceso para comprobar el significado real de las palabras en un texto u oración.

La clave aquí consiste en el diseño de un sistema de macro/meso/micro-significados sobre la base de la estructura pragmática, bajo el supuesto teórico de que a cada x-acto corresponde un x-significado. Con el apareamiento del esquema semántico al esquema pragmático, se puede ya especificar y comprender mejor los actos textuales, ya que, además de la identificación de una acción, se dispone ahora de los distintos mapas representacionales asociados [11].

Pero en general el análisis semántico logra:

- Determinar los valores, normas y creencias del autor.
- Establecer la adecuación entre lo que 'hace' y lo que 'informa'.
- Precisar las relaciones entre los mapas representacionales del autor y los que corresponden al marco situacional.
- Determinar la coherencia, consistencia y relevancia de su información (trivialidades, redundancias, omisiones, etc.) [11].

Análisis Pragmático

Busca datos luego de definir la situación socio-espacio-temporal en la que se produjo el texto, contextuada dentro del respectivo marco situacional (el comercio, el entretenimiento, la recreación, la educación, la academia, la política, etc.) e identificada bajo un Macroacto. Estas definiciones previas permiten fijar un patrón interpretativo y evaluativo para la obtención de los datos.

Después de lo anterior, es conveniente el diseño de la estructura de macro/meso/micro-actos del texto, con el objeto de clarificar lo que 'va haciendo' el autor del texto a medida que desarrolla su discurso y de evaluar la eficiencia del texto. Pero ese diseño de la estructura

pragmática permite visualizar las intenciones del autor, especialmente si sus actos son bivalentes (i. e., si existen acciones encubiertas) [11].

Evidentemente, todo depende de las categorías de análisis de la investigación o de lo que se esté buscando. En general, el análisis pragmático provee información sobre lo siguiente:

- Las relaciones del autor con respecto a la situación socio-espacio-temporal y al marco situacional correspondiente: su grado de adaptación, su dominio de las convenciones del caso, su equilibrio sociocontextual, etc.
- Su eficiencia como actor-hablante: sus mecanismos de enlace, la cohesión y coherencia de sus meso/micro-actos, el esfuerzo invertido en la comunicación, etc.
- Su sistema de valores, motivaciones, creencias, etc; aunque buena parte de los valores se identifican en la instancia semántica, en esta instancia pragmática suele haber algunos 'actos' especiales que reflejan los valores y los sistemas normativos del autor. Tal es el caso de actos como prohibir algo, persuadir de algo, ensalzar algo, manifestar temor de algo, defender algo, etc., donde el nombre del acto (ubicado en el verbo) indica valores y normas, mientras que el 'algo' remite a un posterior análisis semántico que dirá, concretamente, en qué consisten tales valores y normas [11].

Capítulo 2

Estado del arte en PLN

En este capítulo se desglosan todos los trabajos previos de PLN que tienen como propósito la detección y clasificación de humor o doble sentido en textos.

2.1 Trabajos relacionados a PLN

A través de los años han surgido varios trabajos de investigación que han utilizado PLN, para clasificar y recuperar información, sin embargo no se ha realizado ningún sistema el cual sea capaz de detectar el significado real de una oración para el Español. Existen algunos trabajos que han utilizado herramientas como WEKA[12] para definir si textos cortos tienen un significado oculto, mejor conocido como el doble sentido.

2.1.1 Trabajos para el idioma Español

Rigoberto Ocampo en su trabajo de tesis [14], lleva a cabo un modelo de detección de humor en la rima, aliteración, albur, contenido adulto, chistes y dichos en el idioma español. La arquitectura de la solución que el utilizó se puede ver en la figura 2.1

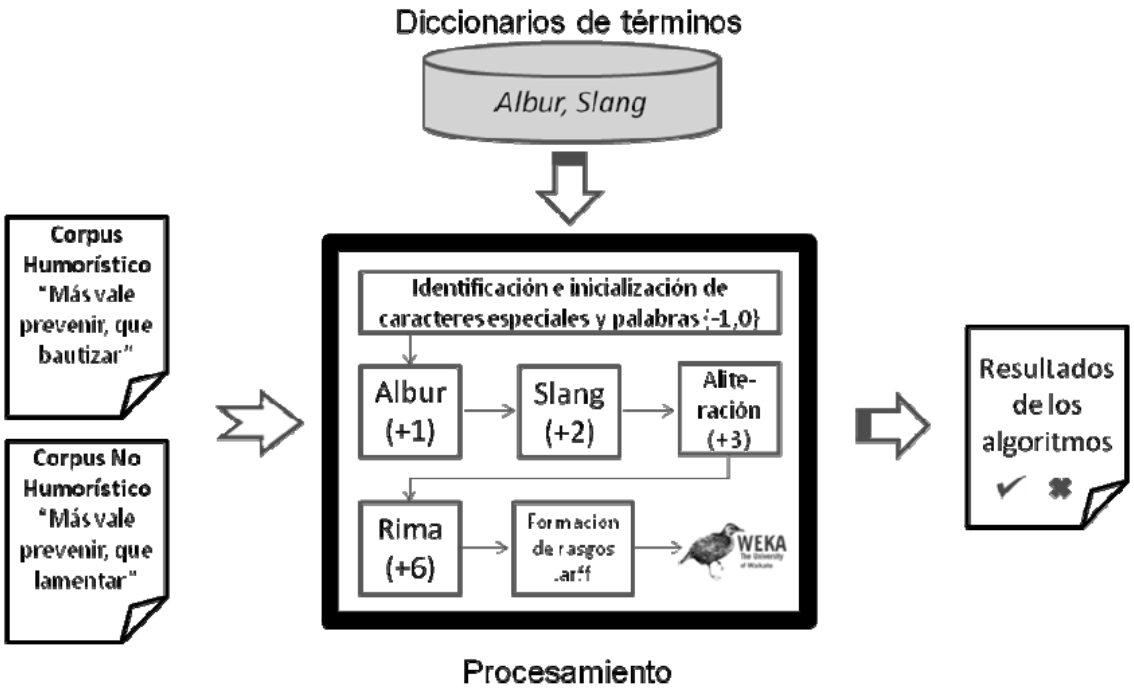

Figura 2.1 Proceso para la detección de humor [14].

[12] Colección de algoritmos de aprendizaje automático para tareas de minería de datos. http://www.cs.waikato.ac.nz/ml/weka/

En la parte que Ocampo denomino Albur, Slang, está compuesta por dos diccionarios de términos, uno con palabras utilizadas en el lenguaje del albur con 203 elementos, colectados del libro "Antología del Albur" [36] y otro con palabras de contenido sexual y otras palabras comunes en los chistes, colectados de diversos sitios de internet de 93 elementos.

Como corpus utilizó dos archivos de texto para su análisis, asegurándose de que uno de ellos estará compuesto por textos humorísticos de 3,676 elementos y el otro archivo de textos no humorísticos (dichos) de 10,000 elementos. Para la clasificación de los textos Ocampo no hizo ningún sistema de clasificación y prefirió obtener textos ya clasificados en internet, el proceso de colección de los textos cortos lo hizo sin ningún programa de por medio que hiciera la búsqueda y la depuración automática. Por el contrario, simplemente utilizó el buscador Google para ubicar algunas páginas que ofrecieran textos humorísticos (chistes) y posteriormente otras que ofrecieran textos no humorísticos en un formato similar al de los chistes, siendo estos los llamados dichos. Dando como resultado un conjunto de datos de entrada de 3,676 chistes y 10,000 dichos. La longitud para los chistes era un máximo de 50 palabras y de los dichos de 35 palabras.

El procesamiento lo hace a través de la herramienta de Detección Automática de Humor en Textos Cortos en Español (DAHTCE), que Ocampo elaboró para su trabajo pero no describe mucho su estructura pero si su funcionamiento:

Recibe dos archivos, a suponerse, uno con textos humorísticos y otro con textos no humorísticos, los cuales procesa por separado identificando cada una de las palabras y caracteres especiales, almacenándolos en una estructura de datos conformada por una cadena de caracteres y una variable entera donde se inicializan con un valor, (-1) para los caracteres especiales y (0) para las palabras. Una vez identificadas todas la palabras y los caracteres, DAHTCE utilizará los módulos de albur, contenido adulto, rima y aliteración, para identificar, por cada renglón del archivo (cada renglón es un chiste o un dicho), los atributos que se encuentran en sus palabras.

A continuación se describen cada módulo que ocupa DAHTCE:

- **Módulo de Albur:** El módulo de albur identifica, a través del diccionario de términos de albur, si alguna de las palabras del texto (mayores a tres caracteres, para evitar la búsqueda con artículos indefinidos, conjunciones, etc.) presenta esta característica.

- **Módulo de Contenido Adulto:** El módulo de Contenido Adulto identifica, a través del diccionario de términos de contenido adulto, si alguna de las palabras del texto (igualmente mayores a tres caracteres, por las razones arriba expuestas) presenta esta característica. En dado caso que así sea, se modifica el valor inherente a la palabra incrementándolo en dos unidades, para denotar este rasgo en ella.

- **Módulo de Aliteración:** El módulo de Aliteración identifica en los textos esta característica tomando en cuenta únicamente las tres primeras letras de cada palabra (de igual forma, para palabras mayores a tres letras), porque son suficientes para

identificar este rasgo y se comparan con las tres primeras letras de las demás palabras. En caso de encontrar igualdad entre ellas, se suman 3 unidades al valor que trae originalmente la palabra (recordando que la palabra puede estar relacionada a un valor de 0, 1 o 2; dependiendo si resultó ser una palabra con albur o contenido adulto), para indicar que también forma parte de una aliteración, siendo características no exclusivas entre las palabras.

- **Módulo de Rima:** El módulo de rima identifica en los textos esta característica analizando únicamente las 3 últimas letras de cada palabra. En caso de encontrar igualdad entre ellas, se suman 6 unidades al valor que trae originalmente la palabra (debido a que en esta ocasión las palabras pueden contener ya valores entre 0 y 5 producto de los análisis anteriores) para indicar que también forma parte de una rima. Cabe aclarar, que se identifica la rima consonante.

- **Formación de los Rasgos:** El módulo formación de rasgos tiene por objetivo la formación del vector de los rasgos, el cual consiste en el análisis final del número asociado a las palabras. Éste indicará los atributos que tiene cada una de ellas. De cada frase se realiza un conteo de cuántas palabras resultaron con cada atributo para determinar los atributos generales de dicho texto; de modo que al finalizar el conteo se tiene para cada frase los valores: SinAlbur, ConAlbur, SinAdulto, ConAdulto, SinAliteracion, ConAliteracion, SinRima, ConRima, dependiendo de la presencia o no de cada atributo en la frase. Por último, se concatenan todos estos valores junto con una etiqueta de Humorístico o NoHumorístico, dependiendo si las frases provienen del archivo con ejemplos positivos o del archivo con ejemplos negativos. Los resultados son unos vectores de datos como el siguiente: {SinAlbur, ConAdulto, SinAliteracion, ConRima, Humoristico}. El resultado final de este módulo es la creación de un archivo .arff (Attribute Relation File Format), que es el tipo de archivos que procesa WEKA.

- **Módulo WEKA:** El módulo WEKA invoca, con el archivo .arff, los siguientes algoritmos:

1. bayes.AODE
2. bayes.AODEsr
3. bayes.BayesNet
4. bayes.HNB
5. bayes.NaiveBayes
6. bayes.NaiveBayesSimple
7. bayes.NaiveBayesUpdateable
8. bayes.WAODE
9. rules.ConjunctiveRule
10. rules.DecisionTable
11. rules.OneR
12. rules.Prism
13. rules.ZeroR
14. lazy.IB1
15. lazy.IBk
16. lazy.KStar
17. lazy.LBR
18. lazy.LWL
19. trees.DecisionStump
20. trees.Id3
21. trees.J48
22. trees.RandomForest

23. trees.REPTree
24. functions.Logistic
25. functions.MultilayerPerceptron
26. functions.RBFNetwork
27. functions.SimpleLogistic
28. functions.SMO
29. functions.VotedPerceptron
30. functions.Winnow
31. misc.HyperPipes
32. misc.VFI

Los valores que tienen las palabras, que fueron mencionados en los módulos se muestran en la tabla 2.1:

Tabla 2.1 Valores de las palabras de los textos al finalizar el procesamiento.

Valor de la palabra	Atributos de la palabra
-1	Carácter especial
0	Sin atributos
1	Albur
2	AdultSlang
3	Aliteración
4	Albur, Aliteración
5	AdultSlang, Aliteración
6	Rima
7	Albur, Rima
8	Adulto, Rima
9	Aliteración, Rima
10	Albur, Aliteración, Rima
11	AdultSlang, Aliteración Rima

Por último, DAHTCE extrae los resultados más relevantes arrojados por todos los algoritmos, los cuales son: *Accuracy*, *True Positive Rate*, *False Positive Rate, Precision* y *Recall*; necesarios para la comparación de la precisión de los algoritmos.

La metodología experimental que utilizó Ocampo en [14] consta de lo siguiente: en cada experimento se proponen dos archivos con diferentes cantidades de ejemplos positivos (o textos humorísticos) y ejemplos negativos (o textos no humorísticos). Ambos archivos se ingresan a la aplicación DAHTCE, la cual extrae los atributos de albur, contenido adulto, aliteración y rima. Una vez extraídos los atributos de los dos archivos, se forman vectores de rasgos por cada texto y con ellos se crea otro documento de extensión .arff, que es el tipo de archivos con los que trabaja WEKA.

Posteriormente, con esta información y desde la misma aplicación DAHTCE, se mandan ejecutar 32 algoritmos de clasificación de WEKA elegidos al azar en tiempo de diseño. Por último se concatenan todos estos resultados en un solo registro y se extraen los parámetros Accuracy, True Positives Rate, False Positives Rate, Precision y Recall, para su análisis y para la obtención de las conclusiones finales.

Se analizan los algoritmos para tres casos particulares que el considero de suma importancia para encontrar el mejor de ellos en la detección del humor en textos cortos en español: El caso para cuando se analizan las 3,675 instancias positivas y las 10,000 instancias negativas como

un solo conjunto de datos, el cual llamó CasoTodasInstancias. El caso para cuando se analizan 3,675 instancias positivas y 3675 instancias negativas (estas últimas tomadas al azar), como un solo conjunto de datos Caso-3675-Heterogéneo. Y por último el caso para 2833 instancias positivas y 2833 instancias negativas que presentan cuando menos un rasgo distintivo del humor. Caso-2833-Homogéneo.

Las medidas que se analizan en todos los algoritmos son Accuracy; así como True Positive Rate, False Positive Rate, Precision y Recall (todas ellas con posibilidad de ser calculadas a partir de la Confusion Matrix), dado que son las medidas más relevantes extraídas por los algoritmos, al indicar claramente las cantidades y proporciones de los ejemplos que fueron clasificados correcta e incorrectamente.

El concluye su trabajo afirmando que los algoritmos bayes.AODE, lazy.LBR y misc.VFI como los mejores algoritmos para la detección de humor en textos cortos en español. Sin embargo también concluye que en el caso del albur falta mucho para poderse detectar gran parte del mismo, dado que muchas veces el albur se presenta como juegos de palabras muy complejos que todavía no está dentro de los alcances de la Lingüística para su detección eficientemente.

2.1.2 Trabajos para el idioma Inglés

Otros trabajos relacionados han sido desarrollados por Rada Mihalcea, pero para el idioma inglés [15]. Por ejemplo, Rada y Attardo en [15] hacen uso de características de estilo de textos humorísticos tales como la aliteración[13], la antonimia y el contenido adulto; así como de características basadas en contenido y de una combinación de ambas características, para hacer una clasificación.

Pero uno de los trabajos destacados por Rada Mihalcea, es que realizo con Strapparava [16] y [17] el cual consistió en el reconocimiento y clasificación automático de humor, su investigación estaba basada para el humor que se encuentra en una sola línea, ya que es una frase corta con efectos cómicos en muy pocas palabras (generalmente 15 o menos) y contienen una estructura lingüística variada: sintaxis simple, uso deliberado de dispositivos retóricos (por ejemplo, la aliteración, rima), y el uso frecuente de construcciones lingüísticas.

Se formularon el problema de reconocimiento de humor como una tarea de clasificación tradicional, la alimentaron con ejemplos positivos (buen humor) y negativos (no humorísticos) a un clasificador automático. El conjunto de datos de humor que emplearon consistió en una sola línea que se recolectó de la web utilizando un proceso de *bootstrapping*[14] automático. En concreto, utilizaron tres conjuntos de datos negativos diferentes: títulos de noticias Reuters, los proverbios y las sentencias de la *British National Corpus (BNC)*. Los resultados de la clasificación son pasables, con figuras de precisión que van desde 79,15% (Citas / BNC) al 96,95% (de una sola línea / Reuters).

La organización del trabajo se llevó de la siguiente manera: primero, se describieron los conjuntos de datos humorísticos y no humorísticos, proporcionando detalles sobre el proceso

[13] Ret. Figura que, mediante la repetición de fonemas, sobre todo consonánticos, contribuye a la estructura o expresividad del verso.

[14] El bootstrap es un tipo de técnica de remuestreo de datos que permite resolver problemas relacionados con la estimación de intervalos de confianza o la prueba de significación estadística.[31]

de *bootstrapping* basado en la Web, que se utiliza para construir una colección muy grande de una sola línea.

Para poner a prueba su hipótesis de que las técnicas de clasificación automática representan un enfoque viable para el reconocimiento de humor, lo que necesitaron en primera instancia era un conjunto de datos que consta de dos ejemplos humorísticos (positivos) y no humorísticos (negativos). Tales conjuntos de datos se pueden utilizar para aprender automáticamente modelos computacionales para el reconocimiento de humor y al mismo tiempo evaluar el desempeño de tales modelos.

Ellos destacan que grandes cantidades de datos de entrenamiento tienen el potencial de mejorar la exactitud del proceso de aprendizaje, y al mismo tiempo proporciona una visión de cómo cada vez más grandes conjuntos de datos pueden afectar a la precisión de clasificación.

La construcción manual de un gran conjunto de datos de una sola línea puede ser problemático, a pesar de que la mayoría de los sitios web o listas de correo que hacen disponibles tales bromas no suelen listar más de 50 a 100 de una sola línea. Para enfrentar este problema, implementaron un algoritmo de programa previo basado en la Web capaz de recoger automáticamente una gran cantidad de bromas en una sola línea a partir de una lista corta como semilla, que consiste en un par de chistes identificados manualmente. El proceso de bootstrapping que ellos implementaron se ilustra en la figura 2.2.

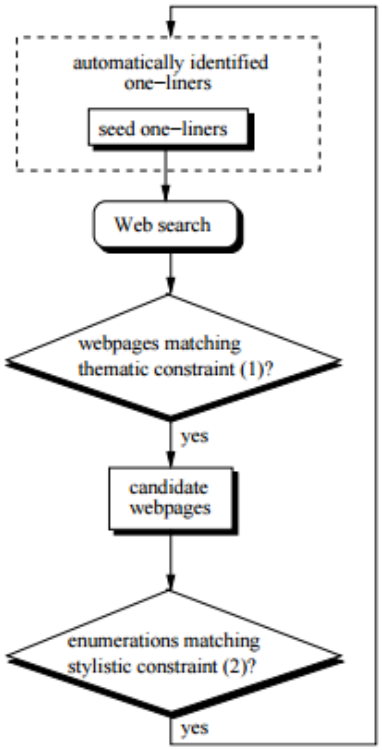

Figura 2.2 Proceso de bootstrapping [16].

Este proceso funciona que, a partir de la muestra inicial, el algoritmo identifica automáticamente una lista de páginas web que incluyan al menos un conjunto de datos de una sola línea, a través de una búsqueda simple que se lleva acabo con un motor de búsqueda en

la web. A continuación, las páginas que se encuentran para analizar son HTML que se identifican automáticamente que son de una sola línea y se añade al conjunto de semillas.

El proceso se repite varias veces, hasta que se recogen suficientes de una sola línea. Un aspecto importante de cualquier algoritmo de *bootstrapping* es el conjunto de restricciones utilizado para dirigir el proceso y evitar tanto como sea posible la adición de entradas con ruido. El algoritmo utiliza: una restricción temática aplicada al tema de cada página y una restricción estructural, explotando anotaciones HTML que indican texto de género similar. La primera restricción se implementa mediante un conjunto de palabras clave de los cuales al menos uno tiene que aparecer en la URL de una página web recuperada, por tanto, potencialmente limitar el contenido de la página web a un tema relacionado con esa palabra clave. El conjunto de palabras clave utilizadas en la implementación actual se compone de seis palabras que indican explícitamente contenidos relacionados humor: de una sola línea, humor, broma, divertido.

Rada y Strapparava, exponen que dos iteraciones del proceso de *bootstrapping*, comenzó con una pequeña muestra de semillas de 1001 chistes, como resultado en un gran conjunto de cerca de 24.000 de una sola línea. Después de retirar los duplicados utilizando una medida de similitud de cadenas basada en la más larga subsecuencia común métrica, se quedan con un conjunto final de aproximadamente 16.000 de una sola línea, que se utilizan en los experimentos de reconocimiento de humor. Toman en cuenta que ya el proceso de recolección es automático y las entradas con ruido también son posibles. La verificación manual de una selección al azar muestra de 200 de una sola línea indica un promedio de 9% el ruido potencial en el conjunto de datos, que está dentro límites razonables, ya que no parece significar algún impacto en la calidad del aprendizaje.

Teniendo su conjunto de datos de entrenamiento, utilizaron dos algoritmos de clasificación de textos de uso frecuente: Naïve Bayes y máquinas de soporte vectorial (SVM), estos fueron seleccionados por su diversidad de metodologías de aprendizaje.

Rada explica que el algoritmo Naïve Bayes es como estimar la probabilidad de una categoría determinada de un documento con probabilidades conjuntas de palabras y documentos. Los clasificadores que utilizan Naïve Bayes asumen independencia a cada palabra, pero a pesar de esta simplificación, se desempeñan bien en la clasificación de texto. Si bien existen versiones de clasificadores Naïve Bayes (variaciones de multinomial y multivariado Bernoulli), ellos emplearon el modelo multinomial, previamente demostrado ser más eficaz [29].

Para el caso de Máquinas de Vectores Soporte (SVM) son clasificadores binarios que buscan encontrar el hiperplano que mejor separa un conjunto de ejemplos positivos de un conjunto de ejemplos negativos, con el máximo margen. Aplicaciones de los clasificadores SVM a categorización de textos llevaron a algunos de los mejores resultados reportados en la literatura [30].

Para la obtención de resultados, ellos llevaron a cabo varios experimentos para obtener información sobre diversos aspectos relacionados con una tarea de reconocimiento automático de humor: precisión clasificación utilizando rasgos estilísticos y basados en

contenido, los tipos de aprendizaje, el impacto del tipo de datos negativos, el impacto de la metodología de clasificación.

Todas las evaluaciones se realizaron utilizando validación cruzada. La línea de base para todos los experimentos es del 50%, que representa la precisión de la clasificación obtenida si una etiqueta de "buen humor" (o "no-humorística") sería asignada por defecto a todos los ejemplos en el conjunto de datos.

En una primera serie de experimentos, evaluaron la precisión de la clasificación utilizando rasgos estilísticos de humor específico: la aliteración, antonimia, y el argot de adultos. Estas son las características numéricas que actúan como heurística, y el único parámetro necesario para su aplicación es un umbral que indica el valor mínimo admitido para una declaración para ser clasificado como humorística (o no humorísticos).

Estos umbrales aprenden de forma automática mediante un árbol de decisión que se aplica sobre un pequeño subconjunto de ejemplos humorísticos / no humorísticos (1.000 ejemplos). La evaluación se realiza en los ejemplos restantes 15.000, los resultados se presentan en la tabla 2.2.

Tabla 2.2 Resultados de experimentos con aliteración, antonimia y Slang adulto [16].

Heurística	One-liners Reuters	One-liners BNC	One-liners Proverbs
Aliteración	74.31%	59.34%	53.30%
Antonimia	55.65%	51.40%	50.51%
Slang Adulto	52.74%	52.39%	50.74%
Todo	76.73%	60.63%	53.71%

El segundo grupo de experimentos se ocupaba de la evaluación de las características basadas en el contenido para el reconocimiento humor. La tabla 2.3 muestra los resultados que obtuvieron con los tres conjuntos diferentes de ejemplos negativos, con los clasificadores de texto Naïve Bayes y SVM.

Tabla 2.3 Resultados de los clasificadores Naïve Bayes y SVM [16].

Clasificador	One-liners Reuters	One-liners BNC	One-liners Proverbs
Naïve Bayer	96.67%	73.22%	84.81%
SVN	96.09%	77.51%	84.48%

Los resultados que obtuvieron en los experimentos de clasificación automática revelan el hecho de que los enfoques computacionales representan una solución viable para la tarea de reconocimiento de humor, y el buen desempeño se puede lograr utilizando técnicas de clasificación basados en las características estilísticas y de contenido. Los resultados también muestran que es más difícil distinguir humor con respecto al texto normal (por ejemplo, frases BNC).

Además de los tres conjuntos de datos negativos, también hicieron un experimento utilizando un corpus de sentencias arbitrarias extraídas al azar de los tres conjuntos negativos. El

reconocimiento humor con respecto a este conjunto de datos mixtos negativo resultó en 63.76% de precisión para las características estilísticas, 77,82% para las características basadas en el contenido utilizando Naïve Bayes y 79,23% usando SVM. Estas cifras son comparables a los reportados anteriormente para una sola línea / BNC, lo que sugiere que los resultados experimentales reportados anteriormente no reflejan un sesgo introducido por los conjuntos de datos negativos, ya que se obtienen resultados similares cuando el humor el reconocimiento se realiza con respecto a los ejemplos negativos arbitrarias.

Por ultimo realizaron un último experimento en el cual utilizaron distribuciones desiguales de clase. Para cada uno de los tres tipos de ejemplos negativos, se construyó un conjunto de datos utilizando el 75% de ejemplos no humorísticos y 25% ejemplos humorísticos. Aunque la línea de base en este caso es mayor (75%), las técnicas automáticas de clasificación de reconocimiento de humor mejora aún más esta línea de base.

Los rasgos estilísticos conducen a una precisión de clasificación de 87,49% (de una sola línea / Reuters), 77,62% (Citas / BNC) y 76,20% (de una sola línea / Proverbios), y las características basadas en el contenido utilizan un clasificador Bayes resultado en cifras de precisión del 96,19% (de una sola línea / Reuters), 81,56% (de una sola línea / BNC) y 87,86% (de una sola línea / Proverbios).

En trabajos posteriores Mihalcea trabaja con Pulman en [19], en donde analizan de forma detallada otras dos características que, según teorías psicológicas, están presentes de manera muy frecuente en el humor: centrado en el humano y la polaridad negativa.

En [19] se investigó el problema del reconocimiento automático de humor, proporcionando un profundo análisis de dos de las características más frecuentemente observados de texto humorístico: humano-centrismo y la polaridad negativa. A través de experimentos llevados a cabo en dos colecciones de textos humorísticos, ellos demuestran que estas propiedades de humor verbal son consistentes a través de diferentes conjuntos de datos.

Para su trabajo, utilizaron un corpus de una sola línea, así como un nuevo conjunto de datos que consiste en artículos de noticias humorísticas. Al tener en cuenta dos bases de datos diferentes.

El primer conjunto de textos humorísticos utilizado para estas pruebas se obtuvo por medio del método de *bootstrapping* automático que utilizo Mihalcea en [16] que solo recolectaba textos de una sola línea.

El segundo conjunto se compone de historias diarias del periódico "*The Onion*", del cual tomaron una publicación semanal satírica con los artículos irónicos sobre noticias de actualidad, en particular en las historias de los Estados Unidos. La recolección fue de todos los artículos publicados en agosto 2005 hasta marzo 2006, que dio lugar a un conjunto de datos de aproximadamente 2.500 artículos de noticias. Realizaron una limpieza de todas las etiquetas HTML, suprimieron la cabecera que contiene información específica del periódico y también eliminaron todos los artículos de noticias que estaban fuera del rango de longitud 1000-10.000 caracteres. Este proceso les dejó conjunto de datos de 1125 noticias con contenido humorístico.

Para la parte de clasificación, aparte de utilizar ejemplos positivos (buen humor), también necesitaron un conjunto de textos negativos (graves). Para cada conjunto de datos buen humor, se construyó una colección de ejemplos negativos, identificado como textos que no son buen humor, pero similar en estructura y composición de los ejemplos humorísticos.

Ellos no querían que los clasificadores automáticos para aprender a distinguir entre los ejemplos humorísticos y no humorísticos basados simplemente en la longitud del texto o las diferencias de vocabulario obvias. Sino que buscaban era que los clasificadores identificaran características humor específico, mediante el suministro de ejemplos negativos similares en la mayoría de sus aspectos a los ejemplos positivos, pero diferentes en su efecto cómico.

Para cada conjunto de datos humorísticos, recopilaron un número igual de ejemplos no humorísticos, mezclando textos a partir de tres o cuatro fuentes diferentes. El propósito de la búsqueda de diferentes fuentes para la construcción del conjunto de datos no-humorísticos negativo es para evitar el sesgo que podría introducirse por una fuente o un género específico.

Para los de una sola línea, crearon un conjunto de datos negativos que consistía en una mezcla de frases siguiendo las mismas restricciones de longitud (10-15 palabras). Combinaron: (1) los títulos de Reuters, extraído de noticias publicadas en la agencia de noticias Reuters durante un período de un año (20/08/1996 a 19/08/1997), (2) Proverbios extraídos de una colección proverbio en línea, (3) Corpus (BNC) frases Nacionales británica y (4) las sentencias de la colección *Open Mind* sentido común de los estados de sentido común.

Para los artículos de noticias, los ejemplos negativos se obtuvieron de tres fuentes diferentes: (1) los artículos extraídos de Los Angeles Times, (2) *newstories* del Servicio de Información de difusión de Relaciones Exteriores y (3) textos extraídos de la *British National Corpus*. Todos los ejemplos no humorísticos fueron obligados a tener una estructura similar a los artículos de "*The Onion*" historias con una longitud de 1,000 - 10,000 caracteres.

Los clasificadores empleados para este trabajo fueron Naïve Bayes y máquinas de soporte vectorial (SVM), ya utilizados anteriormente por Mihalcea en [16].

Para todas las evaluaciones se realizaron con estratificados de diez veces las validaciones cruzadas, para estimaciones precisas. La línea de base para todos los experimentos es del 50%, que representa la precisión de la clasificación obtenida si una etiqueta de "humorística" (o "no-humorística") sería asignada por defecto a todos los ejemplos en el conjunto de datos. La tabla 2.4 muestra las precisiones de clasificación obtenidos con cada uno de los clasificadores.

Tabla 2.4 Exactitud en la clasificación para los dos conjuntos de datos de humor [19].

Clasificador	One-liners	News articles
Naïve Bayer	76.69%	88.00%
SVN	79.23%	96.80%

Los resultados indican que los datos humorísticos y no humorísticos son claramente separables, utilizando exclusivamente las características lingüísticas. No es sorprendente que la precisión de la clasificación para los artículos de noticias es más alta que para los de una sola línea, muy probablemente debido al mayor tamaño de los documentos en la colección los

newstories. La diferente brecha entre la SVM y Naive es la precisión de clasificación de Bayes puede ser probablemente atribuido a la misma razón, con el clasificador SVM que conduce a resultados cercanos al 100% en el caso de los *newstories*, pero a los resultados ligeramente peores que los obtenidos con el Naive Bayes clasificador en el caso de los de una sola línea.

En un análisis previo de las características de humor verbal [17], han tratado de identificar y clasificar las características humor específicos basados en contenidos característicos de la serie de datos de una sola línea.

Examinaron manualmente las funciones basadas en el contenido más discriminativos aprendidas durante el proceso de clasificación de texto, han tratado de clasificarlos en clases semánticas. Las siguientes clases de palabras son:

Vocabulario centrado humano

Bromas parecen hacer constante referencia a escenarios relacionados con humanos, a través del uso frecuente de palabras como usted, yo, hombre, mujer, hombre, etc. Por ejemplo, la palabra que solo ocurre en más del 25% de los de una sola línea ("Siempre se puede encontrar lo que no busca"), mientras que la palabra que se produce en aproximadamente el 15% de los de una sola línea ("de todas las cosas que he perdido, yo falte mi mente más"). Esto apoya las sugerencias anteriores hechas por Freud [32], más tarde por Minsky [33], que la risa es a menudo provocada por sentimientos de frustración causadas por nuestra cuenta, en algún torpe, comportamiento.

Negación

Textos humorísticos parecen incluir a menudo formas de las palabras negativas, como: no lo hace, no es, no lo hacen. Un gran número de las bromas en la colección contiene alguna forma de negación, por ejemplo, "El dinero no puede comprar amigos, pero te dan una mejor clase de enemigo", o "Si al principio no tienes éxito, el paracaidismo no es para ti."

Orientación negativa

Además de las formas verbales negativos, chistes parecen contener también un gran número de palabras con una polaridad negativa, tales como adjetivos con connotaciones negativas como: malo, ilegal, incorrecto ("Cuando todo viene a tu manera, estás en el carril equivocado") o los nombres con una carga negativa, por ejemplo, error, error, el fracaso ("error Usuario: sustituir usuario y pulse cualquier tecla para continuar"). Tanto las formas verbales negativos y las palabras con orientaciones negativas son posibles reflejos de las teorías basadas incongruencia de humor.

Comunidades profesionales

Muchos chistes parecen apuntar a las comunidades profesionales que a menudo se asocian con situaciones divertidas, tales como: abogados, programadores, policías. Por ejemplo, alrededor de 100 de una sola línea en nuestra colección otoño en esta categoría, por ejemplo, "Hacía tanto frío que el invierno pasado vi un abogado con las manos en los bolsillos."

Debilidad humana

Finalmente, el último significativamente grande categoría semántica que han identificado se refiere a eventos o entidades que a menudo se asocian con "débiles" momentos humanos, incluidos los nombres tales como: la ignorancia, la estupidez, problemas ("Sólo los adultos tienen problemas con las botellas a prueba de niños"), cerveza, alcohol ("Todo el mundo debe creer en algo, creo que voy a tener otra cerveza"), o verbos como dejar de fumar, robar, mentir, beber ("Si no se puede beber y conducir, entonces ¿por qué las barras tienen aparcamiento un montón? "). Como se mencionó antes, este tipo de vocabulario parece relacionarse con las teorías del humor que explican la risa como un efecto de frustración o torpes sentimientos, cuando terminamos riendo "a nosotros mismos" [33].

En un nivel superior, estas características se pueden clasificar en dos clases principales. En primer lugar, el hombre centrado en el vocabulario, las comunidades profesionales y la "debilidad" humana se pueden agrupar en la categoría más amplia de egocentrismo humano. En segundo lugar, la negación, la orientación negativa, y la "debilidad" humano todos tienen que ver con la categoría más amplia de la orientación de polaridad. Ellos analizaron cada una de estas categorías, a su vez, y obtuvieron evidencia de una alta correlación entre el texto humorístico y cada una de estas dos características.

Centralidad humana

Para una evaluación más sólida de la propiedad humana centralidad de los textos humorísticos, implementaron un sistema que mide el peso de las características más discriminatorias aprendidas del proceso de clasificación de texto con respecto a las clases semánticas dadas consideradas relevantes para la centralidad humana.

Específicamente, comenzaron creando una lista de características más destacadas para el conjunto de datos humorístico. A partir de las características identificadas como importantes por el clasificador de Bayes ingenuo (un umbral de 0,3 se utilizó en el proceso de selección de características), seleccionaron todas esas características que tienen un peso total superior a un umbral dado T, donde se calcula un peso característica para cada categoría (humorístico / no humorística) y se determina como la probabilidad de ver la función en una categoría dada. A continuación, calcularon la puntuación humorística de una función como la relación entre el peso en el corpus de humor y el peso total en todo el corpus mixto. Esto da como resultado una puntuación dentro del [0-1] intervalo, con un valor cercano a 1 indica un representante de función para los textos humorísticos, y un valor cercano a 0 correspondiente a alta prominencia cuenta para el conjunto de datos no humorístico. En las evaluaciones se informa a continuación, se utiliza un umbral T de 100, lo que les permitió extraer las 1.500 principales características más discriminatorias para cada conjunto de datos.

A continuación, dada una cierta clase semántica, midieron el peso de esa clase semántica con respecto a las características más discriminatorias sumando los pesos correspondientes, y normalizando con respecto al tamaño de la clase semántica. Por ejemplo, suponiendo una clase semántica que incluye las palabras que yo, yo, yo mismo, con las puntuaciones humorísticos de 0,88, 0,65 y 0,55, respectivamente, medidos en el conjunto de datos buen

humor, el peso de la clase semántica dada se mide entonces como (0.88 + 0.65 + 0,55) / 3 = 0,69[15].

Mediante el uso de clases semánticas, pudieron generalizar sobre la palabra individuo características aprendido de salida de los clasificadores y derivar categorías de representante de palabras para los datos de buen humor. Teniendo en cuenta que una clase semántica que no tiene correlación con las características de humor de un texto dará lugar a un peso aproximadamente igual (0,50), medido en los textos humorísticos y no humorísticos.

Para medir la característica humana centralidad de los textos humorísticos, para cada conjunto de datos ellos extrajeron las 1.500 principales características más discriminatorias, y posteriormente se midió el peso de cuatro clases semánticas que consideramos relevante para la propiedad de centralidad humana: las personas, los grupos sociales, las relaciones sociales, y los pronombres personales. Las tres primeras categorías se derivan automáticamente de una clase llamada *WordNet* y la cuarta categoría se construye una lista exhaustiva de todos los pronombres personales en el idioma Inglés. Aunque al principio pensaron que esta clase *WordNet* los ayudaría a descubrir la categoría de comunidades profesionales, en una inspección más cercana, resulta que los nombres pertinentes para este tipo de comunidades (por ejemplo, programador, abogado) están representados en la clase semántica de la persona. En cambio, la categoría de grupo social incluye más organización relacionada sustantivos, como la iglesia, universidad, o consejo, que no son necesariamente representativos para el texto humorístico.

Polaridad Orientación

La segunda característica humor que investigaron, tiene que ver con la orientación de polaridad del humor. En el análisis manual previo [19] habían observado un uso frecuente de las formas verbales en los textos humorísticos negativos, así como otras palabras con orientación negativa (por ejemplo, los adjetivos negativos), o que denota humana "debilidad". Con el fin de tome este análisis con el siguiente paso, e investigar en mayor escala la orientación de polaridad del humor, se ha implementado una herramienta para el análisis automático de sentimiento, y se utiliza esta herramienta para anotar los dos conjuntos de datos humorísticos utilizados.

A partir de un conjunto de datos con anotaciones para la orientación "positiva" y "negativa", implementaron un sistema de clasificación que tiene la capacidad para indicar automáticamente la orientación semántica de un texto. En concreto, están utilizando el conjunto de datos de 10.662 fragmentos cortos de texto introducidas en [34], y alimentaron las y los 5.331 fragmentos de "negativos" 5331 "positivos" en un Clasificador bayesiano. En una de diez veces la validación cruzada experimento, la precisión del sistema se determinó como 78,15%, que se compara favorablemente con los resultados previos reportados en el mismo conjunto de datos [34].

[15] Correspondientemente, el peso de la clase semántica en los textos no humorísticos se mide como 1-0,69 = 0,31.

Pero uno de los trabajos más innovadores es el de Rada Mihalcea et al. [18], [20], que trabaja en conjunto otra vez con Strapparava, en donde exploran diversos modelos computacionales para la resolución de incongruencias, que es una de las teorías del humor más ampliamente aceptadas; la cual sugiere que el humor se debe a la mezcla de dos cuadros opuestos de interpretación posible para un enunciado.

Para evaluar los modelos de incongruencia en el humor, construyeron un conjunto de datos que consta de 150 montajes, cada uno de ellos seguido de cuatro posibles continuaciones de los cuales sólo uno tenía un efecto cómico. Por consiguiente, la tarea se presenta como una tarea de resolución incongruencia, y la exactitud de los modelos se define como su capacidad para identificar la continuación de humor entre los cuatro proporcionado.

El conjunto de datos fue creado en cuatro pasos. Primero tomaron 150 datos de una sola línea que fueron seleccionados al azar del conjunto de datos de humor utilizado en [16]. Después, cada dato de una sola línea se dividió manualmente en una puesta a punto y una línea de perforación, la razón de esta fue el hecho de que lo requerían para minimizar las diferencias entre los cuatro finales alternativos manteniéndolos cortos.

Para que las frases tuvieran sentido y estuvieran completas le proporcionaron la puesta a punto a 10 anotadores humanos para completar la frase. También proporcionaron un rango para el número de palabras a ser utilizadas, el cual fue determinado como una función del número de palabras en la línea. Y como último paso hicieron el filtrado, que se hizo para asegurarse de que las alternativas no tenían errores gramaticales o de ortografía, era coherente, y no tienen un efecto cómico.

Realizando experimentos con los conjuntos de datos, demostraron que las técnicas de clasificación automática para el idioma inglés pueden ser con éxito aplicadas a la tarea de reconocimiento de humor. Los resultados experimentales obtenidos en los conjuntos de datos a gran volumen mostraron que los enfoques computacionales se pueden utilizar de manera eficiente para distinguir entre textos humorísticos y no humorísticos. Ellos observaron mejoras significativas con respecto a priori en líneas de base conocidos, con cifras de precisión que van desde 79,15% (de una sola línea / BNC) a 96,95% (*One-liners* / *Reuters*), que comparan favorablemente con la línea de base de 50%. Este fue el primer resultado de este tipo reportado en la literatura, ya que no contaban con conocimiento de ningún trabajo previo de la interacción entre el humor y las técnicas para clasificación automática [18].

Hasta el momento se han obtenido resultados significativos en el desarrollo de herramientas computacionales para el idioma inglés, sin embargo para el idioma español, el avance ha sido muy lento.

Capítulo 3

Enfoques propuestos

Para la detección del albur es importante identificar los elementos que lo componen para tener un punto de comparación, así mismo generar una propuesta y para ello se requieren los tweets para corroborar.

Con la revisión de trabajos para detección de humor, se seleccionaron algunos métodos para obtener una solución concreta, en la figura 3.1 se muestra la arquitectura de solución que se desarrolla en el este trabajo.

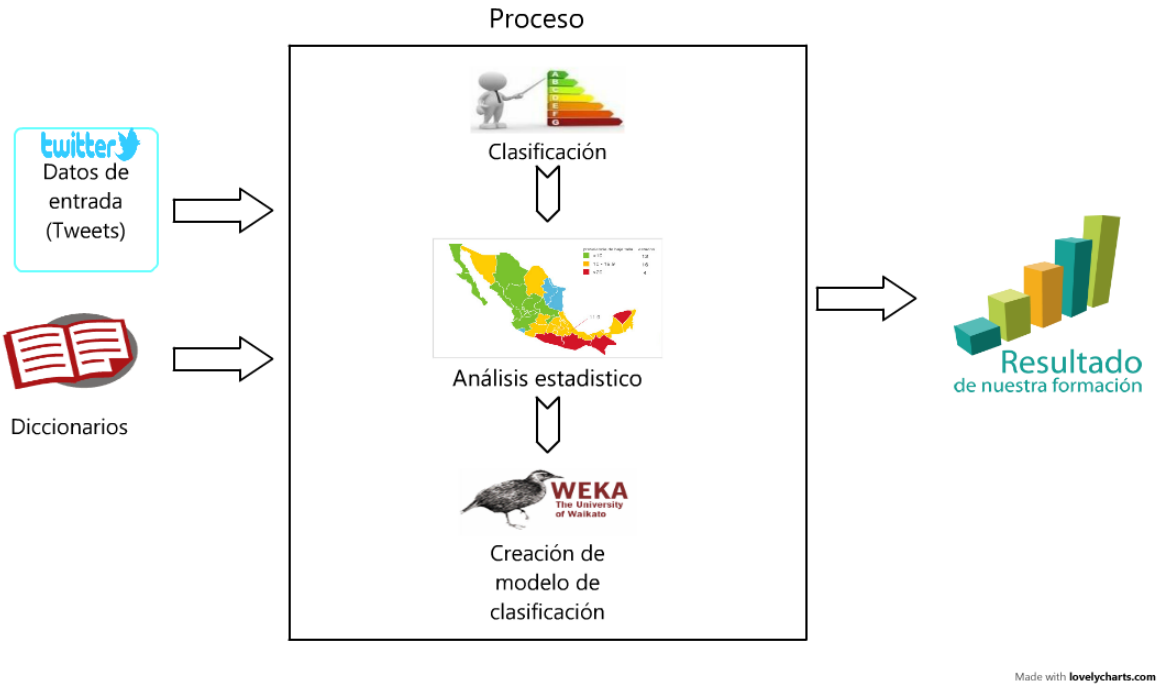

Figura 3.1 Arquitectura de solución

Se considera la metodología que uso Rada Mihalcea en [16] para armar el conjunto de datos, ella emplea textos humorísticos y no humorísticos de una sola línea, en este caso se emplea algo similar solo que en textos cortos, como los tweets, que tienen 140 caracteres como máximo de longitud.

Para la construcción del conjunto de datos no se recurrió a implementar algún sistema para la extracción como lo había realizado Mihalcea en [16], [17], [18] ya que el sistema que ella realizó busca en varias páginas web y lo que se requiere para este trabajo es extraer solo de un sitio web que es Twitter. Los requerimientos que debe tener el sistema es la extracción de grandes cantidades de tweets en varias cuentas y de varias regiones de la república mexicana, después de una larga búsqueda el que cumplía con los requisitos fue el sistema TwitterSearch[16].

[16] Link de descarga: https://github.com/ckoepp/TwitterSearch/archive/master.zip

3.1 Conjunto de datos

3.1.1 Extracción de información o tweets

Para la extracción de datos, se realizaron una serie de pasos para poder hacer uso de la API de Twitter: Primero se procedió en crear una aplicación en https://apps.twitter.com/ , para llevar a cabo este paso, se debe contar con una cuenta de Twitter, ya que se haya creado una, al ingresar a la página pide una serie de datos, como se muestra en la figura 3.2.

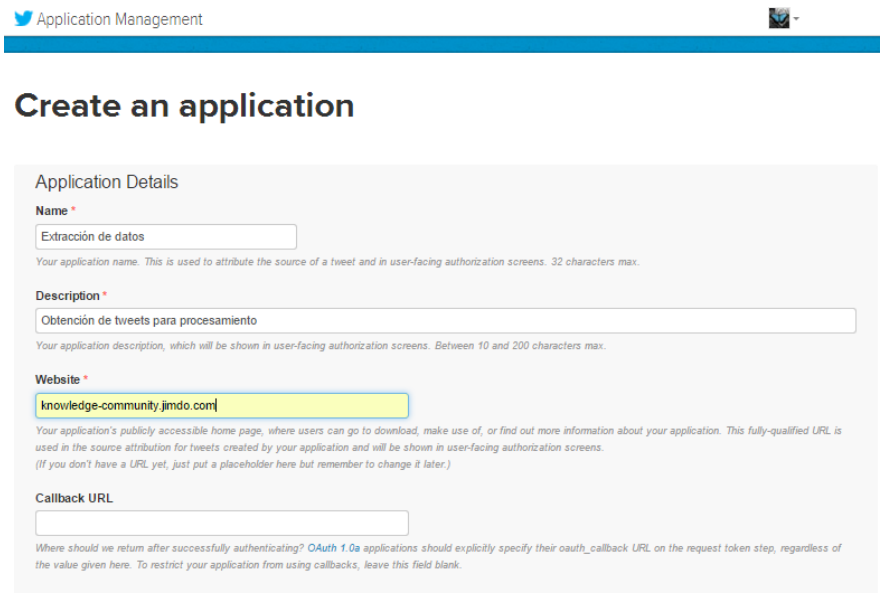

Figura 3.2 Registro de aplicación en Twitter[17]

Se completaron los campos como *Nombre* que es el nombre que llevara la aplicación, *Descripción* donde se coloca una breve descripción de la aplicación y por último el campo *Website* que si se cuenta con un sitio web se coloca, en caso de no contar poner alguna dirección que sea fácil recordar para cambiarla a futuro.

Cuando se haya creado la aplicación se obtendrá la siguiente información como *consumer_key*, *consumer_secret*, *access_token*, *access_token_secret*. Esta información se observa en la figura 3.3.

Ya teniendo los datos necesarios para hacer uso de la aplicación, se procede con la instalación de TwitterSearch, descargándola de la página https://github.com/ckoepp/TwitterSearch .

[17] https://apps.twitter.com/app/new

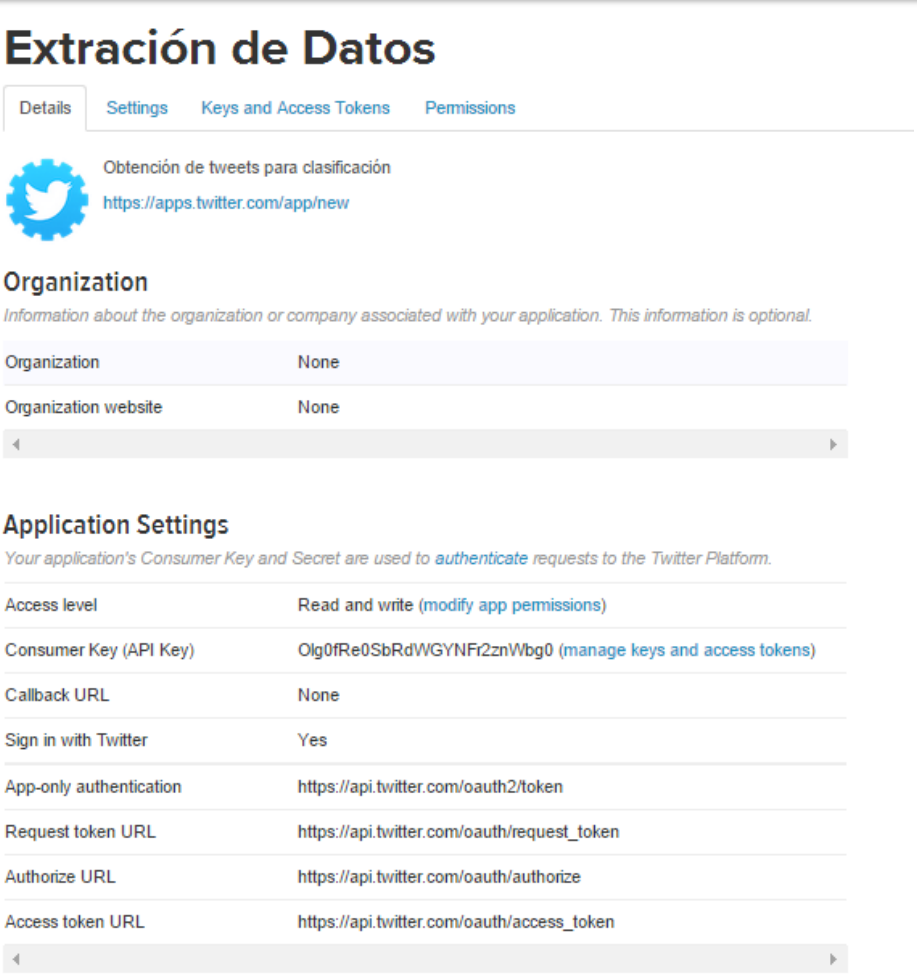

Figura 3.3 Datos de la aplicación

Después de descargar la aplicación, debe ser instalada en una computadora que contenga una distribución de Linux, en este caso puede ser Ubuntu. Para instalarlo solo basta escribir en consola el siguiente comando **sudo pip install TwitterSearch** y se instalara sin ningún problema. Pero en caso de que la versión de Ubuntu no cuente con el comando **pip**, solo basta ejecutar antes el siguiente comando **sudo apt-get install python-pip**.

Concluida la instalación se procede con la extracción de tweets, para ejecutar la aplicación se requiere la ejecución de un programa en lenguaje Python, basándose en el ejemplo que ofrece la página oficial de la aplicación, se le aumentaron algunas cosas para filtrar los tweets que se van a extraer.

Como se muestra en la figura 3.4, se observa la parte más importante del programa donde se puede ver los datos extraídos anteriormente de Twitter cuando se creó la aplicación, así como los delimitadores para filtrar la información en este caso que sean tweets en español. Para el caso de la extracción por estado, se recolectaron aquellos tweets que estuviesen localizados en un radio no mayor a 10 kilómetros de la capital de cada estado de la república mexicana (geolocalización).

```
try:
    tso = TwitterSearchOrder() # create a TwitterSearchOrder object
    tso.setKeywords(['Puebla']) # Permite definir todas que palabras clave debe tener
    tso.setLanguage('es') # Para visualizar tweets unicamente en español
    tso.setGeocode(19.044419, -98.198599, 10, km=True) # centro zocalo de Puebla con un radio de 10 kilometros
    tso.setCount(100) # Solo da los primero 100 resultados por pagina
    tso.setIncludeEntities(False) # No da la información de la entidad

    # Datos extraidos de la aplicación
    ts = TwitterSearch(
        consumer_key = 'jTx4y2RMHfSoWyyy6UPC2A',
        consumer_secret = 'ChZ5SXqPDfCLoAAp6bDSpoGFnlOSyIPV1eQTmMs',
        access_token = '163995934-65saoSGWiEljjMmHaVz3IIcB7SXfjxntrvS9HxU4',
        access_token_secret = 'RsGC0ZFwnU4CBLPORli4s7DFJU2IXEdJUjXnyKT9G62cD'
     )
    f = open(archivo, "w") # crear archivo donde se guardaran los tweets, si ya existe reemplaza el contenido
    for tweet in ts.searchTweetsIterable(tso): # this is where the fun actually starts :)
        f.write( '@%s tweeted: %s\n' % ( tweet['user']['screen_name'], tweet['text'] )) # escribir los tweets en el archivo
    f.close() # cerrar el archivo
except TwitterSearchException as e: # take care of all those ugly errors if there are some
    print(e)
```

Figura 3.4 Programa de ejecución en Python

Ejecutando el programa varias veces, para la obtención tweets de cada estado, se encontró un detalle, ya que por cada consulta Twitter limita las consultas por lapsos de 15 minutos, por lo que se debe esperar ese tiempo entre consultas. Pero terminada la extracción se obtuvo un conjunto de datos preliminar por estado.

Se recolecto un total de 548,243 tweets para ser utilizados, se realizó un cálculo de los datos, el cual indica que se tiene un promedio de 14.340 palabras por tweet, con un vocabulario total de 581,109 palabras y en el cual interactuaron 173,339 usuarios con un promedio de 3.167 tweets por usuario, esto se puede ver en la Tabla 3.1.

Tabla 3.1 Información del corpus de tweets por estado y total

Estado	Tweets	Mínimo de palabras	Máximo de palabras	Promedio de palabras	Vocabulario	Tokens
AGU	14,426	1	32	11.452	23,351	165,200
BCN	13,644	1	33	11.064	23,563	150,961
BCS	11,145	1	30	10.897	18,813	121,443
CAM	15,496	1	36	11.201	22,671	173,569
CHH	12,652	1	39	11.576	21,293	146,458
CHP	12,440	1	31	10.980	19,656	136,590
COA	13,581	1	34	11.667	21,898	158,445
COL	14,498	1	32	12.281	24,728	178,054
DIF	13,452	1	32	11.857	26,305	159,500
DUR	15,837	1	33	11.643	24,401	184,397
GRO	11,012	1	32	10.748	18,502	118,353

Estado	Tweets	Mínimo de palabras	Máximo de palabras	Promedio de palabras	Vocabulario	Tokens
GUA	15,021	1	32	11.136	24,385	167,271
HID	14,304	1	33	10.874	21,349	155,546
JAL	14,201	1	31	11.693	25,394	166,056
MEX	13,208	1	36	11.767	25,311	155,413
MIC	14,643	1	31	11.186	23,104	163,797
MOR	12,288	1	32	11.439	21,242	140,560
NAY	12,892	1	33	11.932	22,854	153,828
NLE	14,354	1	33	10.455	22,999	150,072
OAX	11,159	1	36	12.333	19,221	137,623
PUE	13,415	1	33	11.092	21,143	148,797
QRO	16,274	1	32	10.372	23,390	168,798
ROO	14,369	1	33	11.974	21,975	172,055
SIN	16,169	1	32	10.814	23,518	174,858
SLP	16,473	1	32	11.676	26,194	192,341
SON	12,856	1	31	9.210	17,653	118,405
TAB	15,459	1	31	9.969	21,906	154,117
TAM	16,183	1	36	10.156	21,567	164,360
TLA	12,225	1	33	11.168	20,478	136,525
YUC	12,976	1	31	10.924	22,326	141,755
ZAC	15,394	1	32	11.798	24,138	181,614
TOTAL	548,243	1	39	12.340	581,109	6,765,487

3.1.2 Preprocesamiento de los datos

Sin embargo, los tweets extraídos contenían elementos que no iban a ser indispensables para la parte de clasificación y análisis. Como hashtags, usuarios, links, emoticones, etc. Estos elementos no influyen para determinar si un texto contiene doble sentido o elementos con tendencia vulgar u obscena.

Por lo tanto se llevaría acabó una limpieza de estos elementos para agilizar el procesamiento de los datos en los pasos posteriores, primero, se verificó como estaban almacenados los tweets en el documento de texto obtenido de TwitterSearch (se generó un documento de texto por cada estado de la república), en la figura 3.5 se muestra cómo están los tweets extraídos:

```
@Ramiro_Pedroza tweeted: I'm at Villa India (Aguascalientes, MÃ©xico) http://t.co/RsjqLzz6IG
@ReschMoris tweeted: Este fin estuvo  @gretelresch @AndreaReyesh @casandraresch @hreyess8
@JeNnYJeNnY18 tweeted: Jajaja Asi Es!!! ;) http://t.co/XmHChWHIkT
@JaAC9510 tweeted: @pauvarelam jajaja ya veeeeen
```

Figura 3.5 Tweets extraídos

Como se observa en la figura 3.5, el primer elemento que se encuentra es el usuario, después la palabra “tweeted” acompañada con “:” y finalmente el tweet, que puede contener texto, referencia a otros usuarios, links, emoticones, etc. Así que se procedió remover esos elementos, quedando los tweets de la siguiente manera:

```
niegan proteccion por violencia intrafamiliar a una mujer en guanajuato se la otorga la
no permitas que nadie te diga para que sirves en esta vida el unico que lo puede saber eres tu
me enferma que me bardeen boludo me molesta me saca me todo
```

Figura 3.6 Tweets limpios

Como se observa en la figura 3.6, solo se muestra la oración que es únicamente el texto del tweet, revisando que la información era correcta en un documento prueba, este paso se aplicó a cada documento de cada estado.

3.2 Diccionarios

Para realizar el siguiente paso que es clasificar los datos, es necesario tener un punto de comparación que permita realizar esa clasificación para ello se construyeron dos diccionarios compuestos de palabras y oraciones, que ayudaran para la clasificación de los tweets, para la realización de los diccionarios se tomaron en cuenta los aspectos característicos de albur, lo vulgar y obsceno. El procedimiento de realización de un diccionario fue basado en el trabajo de Rigoberto Ocampo en [14].

Para la creación de los diccionarios, se ocupa como base el diccionario de mexicanismos[18], el cual cuenta con una clasificación y ejemplos de los usos que tiene cada palabra en México, se tomaron en cuenta aquellos que estaba categorizados con VUL y OBSC, haciendo referencia a palabras vulgares y obscenas respectivamente, obteniendo como resultado 2 diccionarios, uno de vulgaridades y el otro de obscenidades, con 409 y 361 frases respectivamente.

A pesar de que se conocía que palabras iban a ser usadas en los diccionarios por los prefijos OBSC y VULG, no era del todo eficiente tomar todas las palabras clasificadas o exclusivamente la palabra, debido a que varias tienen otros significados que no reflejan ser obscenos o vulgares. Tal es el caso del siguiente ejemplo:

andar. INTR. supran. Visitar un lugar, estar en él no permanentemente: “Marilú ahora anda por Europa”.

Figura 3.7 Primer significado de andar

En el primer significado de “andar”, en la Figura 3.7 es sólo la palabra, este no tiene ningún sentido vulgar, por lo tanto se requiere analizar otro de sus significados, para observar que componentes afectan para que cambie el sentido, por ejemplo:

[18] Diccionario de Mexicanismo, Academia Mexicana de la Lengua.

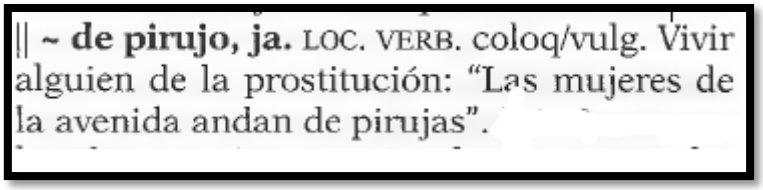

Figura 3.8 Significado vulgar de andar

En la figura 3.8, "andar" es una locución verbal o frase compuesta que incluye el verbo andar y como se observa el sentido de la palabra cambió. Por lo tanto para agilizar y obtener mayor precisión en la detección de frases obscenas o vulgares, se deben tomar en cuenta las palabras que proceden o suceden a una palabra para obtener completa la expresión y facilite el procesamiento en un texto corto.

Se localizaron algunas palabras que con todo y complementos, era muy difícil diferenciar si se aplicaba un doble sentido o no, y en estos casos es indispensable analizar el contexto, para ello se requiere un nivel de análisis semántico que por ahora no será aplicado. Lo anterior se concluyó debido a que se hizo un análisis léxico de prueba con el conjunto de tweets, como resultados se obtuvieran todos aquellos tweets que tuvieran estas palabras, la mayoría no tenían ningún albur, por lo tanto para este trabajo, estas palabras no son incluidas en los diccionarios.

3.3 Clasificación

Retomando el trabajo de Mihalcea [16], que ella solo empleaba dos tipos de textos que eran humorísticos y no humorísticos, se siguió el concepto de manejar una clase que no contuviera ningún contenido vulgar u obsceno llamándola ninguna, entonces en un principio, para realizar la clasificación de lo los datos se tomaron en cuenta que existirían tres clases: tweets vulgares, obscenos y ninguna. Para realizar esta clasificación se emplearían los diccionarios elaborados y se realizó un análisis léxico para encontrar frecuencias de palabras o frases en tweets.

El mecanismo fue el siguiente: Se tomó primero el diccionario de obscenidades y se comparan tweet por tweet con cada elemento del diccionario y las fue clasificando, en caso de no encontrar una frecuencia el tweet era clasificado como ninguna. Después se tomó el segundo diccionario para la comparación, aunque aquí el proceso cambio debido a que ya los tweets ya estaban clasificados en primera instancia, así que si encontraba algún tweet que tenía alguna vulgaridad, solo quitaba la etiqueta ninguna y le asignaba la nueva que era vulgaridad.

La clasificación se realizó de manera correcta y todo parecía estar bien, pero había un caso que no se había contemplado desde un inicio y era que en los diccionarios existían frases o palabras que estaba catalogadas como vulgares y obscenas, entonces que pasaba con estos casos, ya que estaban siendo ignorados, no podía pasar por alto este comportamiento y se concluyó en adicionar una clase adicional llamada mezclado debido a que en esos casos no se podía decidir que solo cumpliera ser vulgar u obscena, ya que no sería lo suficientemente preciso para los modelos de clasificación. Se modificó el proceso de clasificación en la segunda parte donde utiliza el diccionario de vulgaridades, validando que si la frase ya ha sido etiquetada como obscena y encuentra elementos vulgares, que la clasifique como mezclado.

Se realizó un programa en AWK[19] para el proceso de clasificación de los tweets, en la figura 3. 9 se muestra un extracto de la segunda parte de la clasificación donde se da el caso de que existan tweets con la clasificación mezclado.

```
linea="";
for (i=1; i<NF; i++)
   linea = linea " " $i; # Obtener tweet
for (x in obsc)
  if (match (x, linea)!= 0)
    {
     if (match (obsc[x], "NINGUNA")!= 0 && match ($NF,"VULGARIDAD")!= 0)
       {
        obsc[x] = "VULGARIDAD";
       }
     if (match (obsc[x], "OBSCENIDAD")!= 0 && match ($NF,"VULGARIDAD")!= 0)
       {
        obsc[x] = "MEZCLADO";
       }
    }
```

Figura 3.9 Extracto de código en AWK

3.3.1 Balanceo de clases

Una vez finalizado la clasificación, al revisar la cantidad de tweets clasificados, dominaban más los tweets clasificados como vulgares, que las clases estaban completamente desbalanceadas y por ende al realizar los modelos de clasificación iba a predominar esa clase. Para tener consistencia en los datos de entrada de los modelos de clasificación, se realizó un proceso de balanceo quedando la misma cantidad de tweets correspondiente a la clase menos representativa en dicho conjunto, en este caso había sido la de obscenidades. El proceso de balanceo fue realizado por programa en AWK, en cual aplicaba la selección de los tweets manera aleatoria, quedando al final 130 tweet por cada categoría.

[19] AWK es un lenguaje de programación diseñado para procesar datos basados en texto, ya sean ficheros o flujos de datos

Capítulo 4

Experimentos y pruebas realizadas

En este último capítulo se expondrán los procedimientos y resultados de los experimentos que se llevan a cabo con el conjunto de datos que se cuenta.

4.1 Análisis estadístico

Para fines de estudio, se realiza una estadística para verificar que estados de la republica tienden a utilizar con mayor frecuencia elementos vulgares u obscenos en tweets, en comparación de otros estados, así como cuales son las expresiones más utilizadas a nivel república mexicana.

Se realiza el análisis sobre los tweets ya clasificados sin balancear, utilizando los diccionarios de vulgaridades y obscenidades, el proceso es el siguiente:

Se realiza un conteo de cuantos tweets estaban clasificados como vulgares u obscenos o ambos, para la obtención de cuantas incidencias había por estado, dando como resultando en primer lugar para las vulgaridades, el estado de Querétaro (QRO), seguido de Tabasco (TAB) y Nuevo León (NLE), y en último lugar se posiciona Sinaloa (SIN), como se puede observar en la Tabla 4.1.

Tabla 4.1 Frecuencia de ocurrencia de vulgaridades por estado

Estado	Ocurrencias	Estado	Ocurrencias
QRO	408	MOR	247
TAB	350	MIC	240
NLE	347	ZAC	230
PUE	345	MEX	225
HID	325	YUC	224
TAM	319	GRO	220
CHP	311	DF	217
VER	306	BCN	211
GUA	305	ROO	203
AGU	305	COA	182
CHH	303	COL	181
TLA	284	NAY	179
CAM	275	OAX	171
SON	264	BCS	167
SLP	260	DUR	166
JAL	254	SIN	102

Después se procede con un análisis léxico para obtener la frecuencia de palabras vulgares más usadas y los estados que más las ocupan, se genera un top cinco de palabras o frases vulgares que dentro de todos los tweets obtuvieron mayor ocurrencia, las cuales se pueden mostrar en

la Figura 4.1, y que fueron tomadas a partir de la Tabla 4.2, donde se muestra los estados que más utilizan dicha palabra o frase.

Tabla 4.2 Palabras vulgares con mayor ocurrencia en total.

Frase	Ocurrencias	Estados
pedo	1256	**QRO**, TAB, PUE
pendejo	935	**QRO**, CHH, NLE
no mames	601	HID, TAB, TAM
pendeja	401	CAM, ZAC, HID
que pedo	382	PUE, **QRO**, CHH

Como se observa en la tabla 4.2 para las palabras vulgares la que tuvo mayor ocurrencia fue la palabra pedo con 1,256 ocurrencias, los estados que la presentaron con mayor frecuencia fueron: Querétaro, Tabasco y Puebla. Otro aspecto que resalta en esta palabra es que los estados que encabezan el mayor uso de vulgaridades, utilizan con frecuencia esta palabra. Para interpretación gráfica de resultado, se utilizó la interfaz de programación de aplicaciones de google[20], se realizó una gráfica de barras para representar el top 5 de palabras como se muestra en la figura 4.1 y un mapa para representar los resultados de la tabla 4.2 con la información que se obtuvo, como se muestran en la figura 4.2.

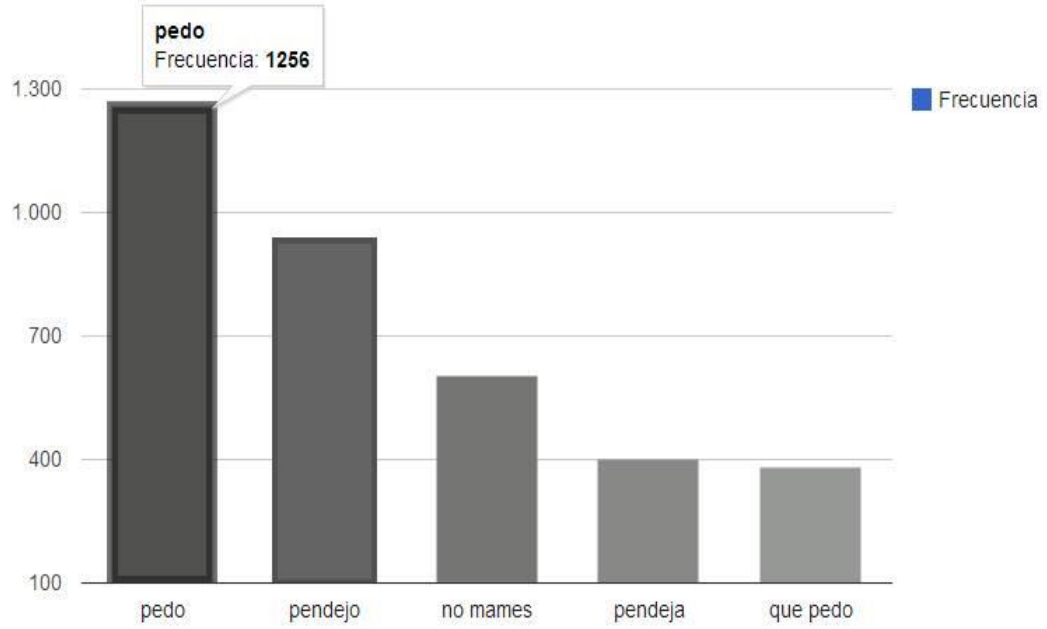

Figura 4.1 Palabras o frases vulgares con mayor frecuencia.

[20] https://developers.google.com/maps/?hl=es

Figura 4.2 Mapa de mayor incidencia de vulgaridades, en el estado de Querétaro.

Posteriormente, se realiza otro análisis para cada estado, con el objetivo de obtener dentro de los tweets que estados manejan dentro de su vocabulario mayor cantidad de obscenidades, en las tres primeras posiciones se encuentra Chihuahua (CHH), Zacatecas (ZAC) y Guanajuato (GUA), respectivamente, y en último lugar se posiciona Sonora (SON). Tabla 4.3

Tabla 4.3 Palabras obscenas con mayor ocurrencia en total.

Estado	Ocurrencias	Estado	Ocurrencias
CHH	29	CAM	19
ZAC	**28**	**NAY**	**18**
GUA	28	SIN	17
AGU	**28**	**SLP**	**16**
MIC	27	HID	16
MEX	**27**	**TLA**	**15**
VER	26	COA	15
QRO	**25**	**BCS**	**15**
JAL	24	CHP	14
TAB	**23**	**PUE**	**13**
DF	23	YUC	12
NLE	**22**	**COL**	**12**
BCN	21	ROO	11
MOR	**20**	**GRO**	**8**
TAM	19	DUR	8
OAX	**19**	**SON**	**1**

Se obtuvo el top cinco palabras o frases obscenas que dentro de todos los tweets obtuvieron mayor ocurrencia las cuales se muestran en la Figura 4.3 y Figura 4.4, éstas fueron tomadas a partir de la Tabla 4.4, la palabra que mayor ocurrencias tuvo fue coger con 152, a diferencia del análisis con vulgaridades que los estados que encabezaban la lista eran los que mayor usaban la palabra, en obscenidades solo se encuentra uno de los primeros.

Tabla 4.4 Palabras obscenas con mayor ocurrencia en total.

Frase	Ocurrencias	Estados
coger	152	MEX, **CHH**, JAL
mamar	95	DF, QRO, VER
pito	92	MIC, **CHH**, NAY
panocha	32	**CHH**, GUA, COL
mamaste	32	**CHH**, TAM,COA

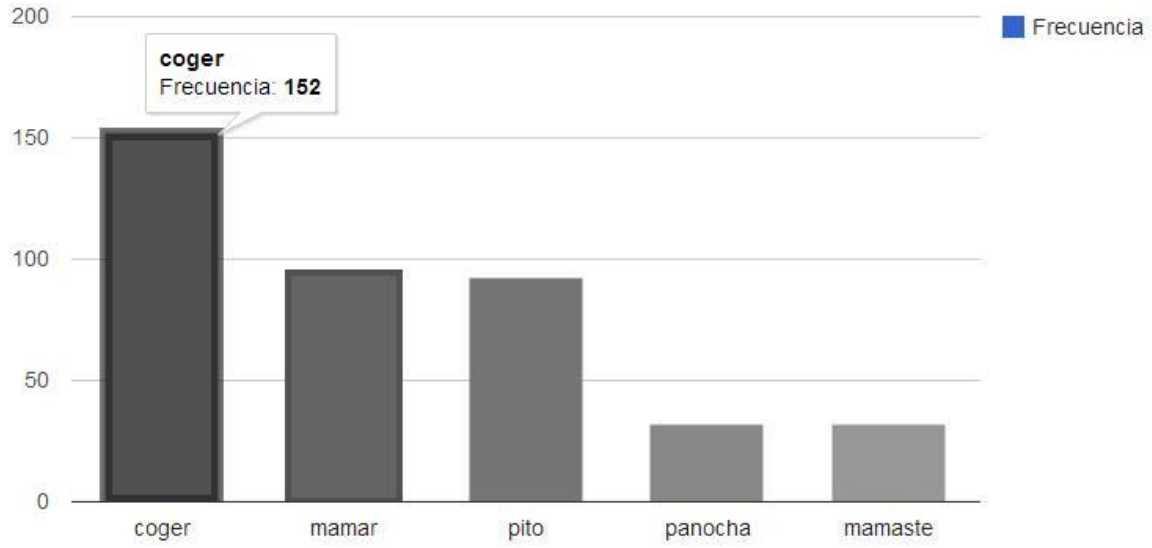

Figura 4.3 Palabras o frases obscenidades con mayor frecuencia.

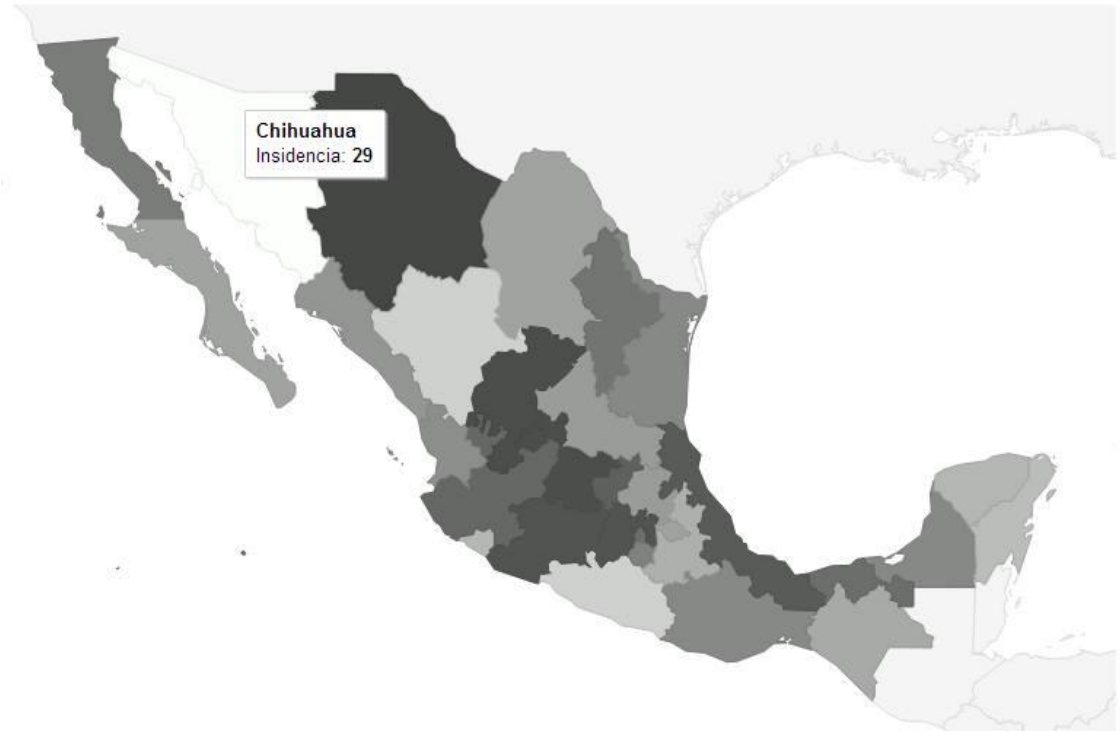

Figura 4.4 Mapa de mayor incidencia de obscenidades, en el estado de Chihuahua.

4.2 Modelo de clasificación

El análisis datos de entrada y la detección de las palabras obscenas y vulgares más utilizadas en las diferentes regiones, ayudó a reforzar y validar la clasificación de los tweets que forman parte del corpus de entrenamiento.

Una vez desarrollado el corpus de entrenamiento, se procede a generar un modelo de clasificación de todas las palabras que aparecen en cada tweet. No se aplica ningún procesamiento adicional al corpus construido ya que con el procesamiento que se había realizado como limpieza y balaceo, se consideró que no se requería otro adicional.

Para el desarrollo de los modelos de clasificación se utilizó la herramienta Weka. En particular los algoritmos de clasificación empleados fueron: Vecino más cercano (IBK), Máquina de soporte vectorial (SMO), Bayes Multinomial y Árbol de decisión (J48).

En la Tabla 4.5 se pueden ver los resultados obtenidos, aplicando validación cruzada con 10 pliegues. Como se puede observar el algoritmo J48 es el que nos brinda mejores resultados con un 76.35%.

Tabla 4.5 Clasificación de tweets con cuatro clases

Algoritmo	Correctamente clasificadas	Incorrectamente clasificadas
IBK	50.77%	49.23%
SMO	75%	25%
Bayes multinomial	69.81%	30.19%
J48	**76.35%**	23.65%

Como el número de muestras positivas de cada clase es pequeño (solo 130 tweets por categoría), se puede apreciar que los resultados en el modelo de clasificación no fueron muy altos, ya que los algoritmos de clasificación trabajan mejor con 2 categorías (elementos positivos y negativos) como en los trabajos [16] y [17] de Mihalcea, por este motivo se decidió desarrollar entonces dos corpus de entrenamiento, considerando solamente dos categorías: Obscenidad y Ninguna o Vulgaridad y Ninguna. Ambos corpus están balanceados, las categorías Obscenidad y Ninguna cuentan con 582 tweets por categoría, mientras que las categorías Vulgaridad y Ninguna se componen por 6945 cada una.

Las Tablas 4.6 y 4.7, muestran los resultados obtenidos aplicando validación cruzada con 10 pliegues.

Tabla 4.6 Clasificación con las clases Obscenidad y Ninguna

Algoritmo	Correctamente clasificadas	Incorrectamente clasificadas
IBK	70.53%	29.47%
SMO	**91.07%**	8.93%
Bayes multinomial	85.14%	14.86%
J48	84.28%	15.72%

Tabla 4.7 Clasificación con las clases Vulgaridad y Ninguna

Algoritmo	Correctamente clasificadas	Incorrectamente clasificadas
IBK	75.04%	24.96%
SMO	**98.90%**	1.10%
Bayes multinomial	88.54%	11.46%
J48	97.46%	2.54%

Como puede apreciarse los mejores resultados fueron ofrecidos por el algoritmo de clasificación máquina de soporte vectorial, con los datos por defecto que ofrece la herramienta Weka.

El hecho de que haya mejorado la precisión del modelo de clasificación se debe en gran medida al número de muestras positivas en cada clase, y por el hecho de que es más simple para el algoritmo de clasificación cuando se trabaja con 2 clases, que con 4 clases.

Conclusiones

En este trabajo de tesis de cumplió con cada uno de los objetivos planteados, cabe resaltar que la exhaustiva investigación realizada de las distintas técnicas de PLN nos fue de utilidad, para el desarrollo de un modelo de clasificación. Conociendo las técnicas no podía faltar la revisión de trabajos previos, lo cual revelo que existen pocos trabajos referentes al albur debido a la complejidad que este tiene.

En este trabajo se presenta una primera aproximación para la identificación de frases obscenas y vulgares en mensajes de twitter. En los resultados obtenidos en el análisis estadístico por cada estado de la república mexicana, se puede observar que los mensajes de este tipo se emplean más en los estados de: Guanajuato, Estado de México y Jalisco para obscenidades y en el caso de vulgaridades los estados de la republica detectados que usan este tipo de frases son: Querétaro, Puebla e Hidalgo.

En los experimentos realizados con los modelos realizados, los mejores resultados obtenidos fueron de los modelos de 2 clases, debido a que se tiene un mejor control, ya que para 4 clases existe mayor probabilidad de error y los resultados obtenidos lo demuestran.

El corpus y el modelo de clasificación desarrollado es nuestro primer acercamiento para intentar educar a las nuevas generaciones a comunicarse de manera correcta por redes sociales.

Entre las mejoras de este trabajo es realizar un análisis semántico, el cual consiste en obtener el significado de cada elemento de la oración en base al contexto, aplicar este análisis ayudará a detectar con más precisión el doble sentido en tweets, debido a lo que demuestra el análisis léxico que es limitado en cuestiones de contexto, porque una simple comparación palabra a palabra, no es lo más acertado para el desarrollo de un sistema que sirva como intérprete de textos.

Otra propuesta seria aumentar el número de tweets para entrenamiento del modelo de 2 clases Obscenidad y Ninguna, porque la cantidad de tweets empleada está muy limitada.

Como trabajo a futuro este tipo de análisis puede ayudar a que sistemas como traductores o robots puedan procesar frases y obtener el significado correcto, debido a la estructura del albur es compleja pero a la vez muy completa, por la polisemia de las palabras.

Bibliografía

[1] E. Camacho Hernández y I. Chorres Chacón, "DESARROLLO NEUROLINGÜÍSTICO DEL LENGUAJE", San José, Costa Rica, 2001.

[2] J. Mejia Prieto, "Albures y refranes de Mexico", Mexico: Panorama, 2005.

[3] T. A. van Dijk, "Discurso y contexto", GEDISA, 2013.

[4] H. Beristáin, «"El Albur", Mexico, 2000.

[5] M. CÓRDOBA URBANO, "Lenguaje y desarrollo", *Innovación y experiencias educativas,* nº 36, 2010.

[6] P. J. Chamizo Domínguez, Ensayistas, 2005. [En línea]. Available: http://www.ensayistas.org/critica/retorica/chamizo/. [Último acceso: 8 2 2015].

[7] Figuras Literarias, [En línea]. Available: http://figurasliterarias.org/content/metonimia. [Último acceso: 8 2 2015].

[8] D. Vázquez, Metafora y Analogia en Aristoteles, *Tópicos,* nº 28, pp. 85-116, 2010.

[9] Wikilengua del Español, [En línea]. Available: http://www.wikilengua.org/index.php/An%C3%A1lisis_sint%C3%A1ctico. [Último acceso: 22 02 2015].

[10] A. Gelbukh, G. Sidorov y F. Velásquez , Análisis morfológico automático del español a través de generación.

[11] J. Padrón, Análisis del discurso e Investigacion Social, Caracas, 1996.

[12] P. Vasile Rus, Encyclopedia of Sciences and Religions, Memphis: Springer Netherlands, 2013.

[13] A. Cortez Vásquez, H. Vega Huerta y J. Pariona Quispe, Procesamiento de lenguaje natural, *RISI,* vol. 6, nº 2, pp. 35-44, 2009.

[14] R. Ocampo (2010). Detección automática de humor en textos cortos en español (tesis de Maestria). Instituto Politecnico Nacional. México.

[15] R. Mihalcea y S. Attardo, Making Computers Laugh, de *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, 2005.

[16] R. Mihalcea y C. Strapparava, Making Computers Laugh: Investigations in Automatic Humor Recognition, de *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, 2005.

[17] R. Mihalcea y C. Strapparava, Learning to Laugh (Automatically): Computational Models for Humor Recognition, *Computational Intelligence,* 2006.

[18] R. Mihalcea y C. Strapparava, Laughter Abounds in the Mouths of Computers: Investigations in Automatic Humor Recognition, *INTETAIN,* pp. 84-93, 2005.

[19] R. Mihalcea y S. Pulman, Characterizing Humor: An exploration of Features in Humorous Texts, *CICLing,* pp. 337-347, 2007.

[20] R. Mihalcea, C. Strapparava y S. Pulman, Computational Models for Incongruity Detection in Humour., *CICLing,* pp. 364-374, 2010.

[21] WEKA, [En línea]. Available: http://weka.wikispaces.com/Primer. [Último acceso: 01 03 15].

[22] Twitter. [En línea]. Available: https://about.twitter.com/es/company. [Último acceso: 08 Marzo 2015].

[23] J. D. Polo, Twitter... para quien no usa Twitter, España: Bubok, 2010.

[24] Wikipedia, [En línea]. Available: http://en.wikipedia.org/wiki/Twitter. [Último acceso: 22 3 2015].

[25] G. Corpas Pastor, Manual de fraseología española, Madrid: Gredos, 1997.

[26] S. Dovrtělová, *LOCUCIONES VERBALES,* Brno, 2008.

[27] L. Ruiz Gurillo, Las locuciones en español actual, Madrid: Arco/Libros, 2001.

[28] C. Koepp, twittersearch, 2014. [En línea]. Available: https://twittersearch.readthedocs.org/en/latest/basic_usage.html#architecture. [Último acceso: 20 05 2015].

[29] M. a. K. Nigam, A comparison of event models for Naive Bayes text classification, de *In Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[30] Joachims, Text categorization with Support Vector Machines: learning with many relevant features, de *In Proceedings of the European Conference on Machine Learning*, 1998.

[31] R. Ledesma, Introduccción al Bootstrap. Desarrollo de un ejemplo acompañado de software de aplicación, *Tutorials in Quantitative Methods for Psychology,* vol. 4, nº 2, pp. 51-60, 2008.

[32] S. FREUD, Der Witz und Seine Beziehung zum Unbewussten. Deutike, Vienna, 1905.

[33] M. MINSKY, Jokes and the logic of the cognitive unconscious. Tech. rep, MIT Artificial Intelligence Laboratory, 1980.

[34] B. A. L. L. PANG, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, de *In Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 2004.

[35] Boukaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., y Scuse, D. (2010). "WEKA manual for version 3-6-2". The University of Waikato.

[36] Hernández, V. (2006). "Antologia del Albur", ISBN: 1-4196-2447-4