

INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

LABORATORIO DE LENGUAJE NATURAL Y PROCESAMIENTO DE TEXTO

**“Análisis automático de opiniones
de productos en redes sociales”**

T E S I S

**PARA OBTENER EL GRADO DE
MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

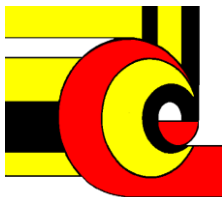
PRESENTA

LIC. HUGO LIBRADO JACOBO

DIRECTORES DE TESIS:

DR. GRIGORI SIDOROV

DR. ALEXANDER GELBUKH



MÉXICO, D.F., ENERO DEL 2016



INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 18:30 horas del día 14 del mes de diciembre de 2015 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Análisis automático de opiniones de productos en redes sociales"

Presentada por el alumno:

LIBRADO

Apellido paterno

JACOBO

Apellido materno

HUGO

Nombre(s)

Con registro:

B	1	3	0	0	9	0
---	---	---	---	---	---	---

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron **APROBAR LA TESIS**, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de Tesis

Dr. Grigori Sidorov

Dr. Alexander Gelbukh

Dr. Ildar Batyrshin

Dr. Francisco Hiram Calvo Castro

Dra. Sofía Natalia Galicia Haro

Dr. Marco Antonio Moreno Ibarra

PRESIDENTE DEL COLEGIO DE PROFESORES



Dr. José Alfonso Villa Vargas

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN
EN COMPUTACIÓN
DIRECCIÓN

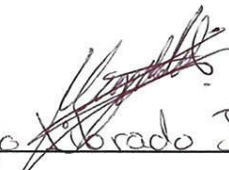


INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México el día 15 del mes diciembre del año 2015, el (la) que suscribe Hugo Librado Jacobo alumno (a) del Programa de Maestría en Ciencia de la Computación con número de registro B130090, adscrito al Centro de Investigación en Computación, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección del Dr. Grigori Sidorov y Dr. Alexander Gelbukh y cede los derechos del trabajo intitulado Análisis automático de opiniones de productos en redes sociales, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección hugo74@gmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.


Hugo Librado Jacobo
Nombre y firma

RESUMEN

El rápido avance de las tecnologías de la información y la comunicación en las últimas décadas, ha permitido al ser humano crear una nueva sociedad, una sociedad de la información. Dentro de toda la información que se genera hoy en día, podemos encontrar una nueva forma de comunicación, pero sobre todo, una nueva forma de opinión, la opinión en redes sociales.

Según la Real Academia Española (RAE), la opinión es un dictamen o juicio que se forma de algo cuestionable, y podríamos agregar, qué es algo que los seres humanos hacemos natural y cotidianamente, por lo cual, con la llegada de la Web 2.0 y el rápido crecimiento de las redes sociales, opinar en redes sociales se popularizó rápidamente. Podemos encontrar opiniones sobre productos, política, noticias, personas famosas, etcétera. Sin embargo, la capacidad humana para analizar la información de estas opiniones tiene un límite, el cual en cierta forma, mediante el procesamiento del lenguaje natural se puede superar.

En la presente tesis se propone una metodología y se desarrolla una aplicación que permite el análisis de textos cortos de opinión, clasificándolos en muy positivos, positivos, neutros, negativos, muy negativos y sin opinión o sentimiento. Enfocándonos en textos cortos del idioma español y considerando una de las redes sociales más populares de la actualidad, Twitter.

ABSTRACT

Fast progress on information and communication technologies in past decades, has allowed to human beings to develop a new society, an information society. Within all generated information nowadays, we can find new communication ways, but especially, a new form of opinion, social networks opinion.

According to Royal Spanish Academy (RAE), opinion is a sentence or judgment that is developed from something questionable, and we could add, that is something that human beings daily and naturally do, whereby, with arriving of Web 2.0 and fast growing of social networks, opinions in social networks became popular. We can find opinions about products, politics, news, famous people, so on. However, human ability to analyze information of opinions has a limit, which somehow, with natural language processing can be overcome.

In this thesis it is proposed a methodology and it is developed an application that enables the analysis of short texts, ranking them as very positive, positive, neutral, negative, very negative and unfeeling. And we focus on short texts of Spanish language and consider one of the most popular social networks today, Twitter.

AGRADECIMIENTOS

Agradezco a mis asesores por su paciencia, guía y apoyo brindado para la realización de este trabajo.

Agradezco a mis padres por apoyarme siempre incondicionalmente.

Agradezco a mis amigos por todo el apoyo y palabras de aliento a continuar mi camino hasta alcanzar mis propósitos.

Agradezco al Laboratorio de Procesamiento del Lenguaje Natural, al Centro de Investigación en Computación, al Instituto Politécnico Nacional y a CONACyT por brindarme la oportunidad de continuar preparándome y alcanzar uno más de mis objetivos.

ÍNDICE

AGRADECIMIENTOS	VI
RESUMEN	IV
ABSTRACT.....	V
CAPÍTULO 1. INTRODUCCIÓN	13
1.1 ANTECEDENTES	13
1.2 PLANTEAMIENTO DEL PROBLEMA.....	16
1.3 OBJETIVOS GENERALES Y ESPECÍFICOS	17
1.4 JUSTIFICACIÓN	18
1.5 ALCANCES Y LIMITES.....	19
1.6 ESTRUCTURA DE LA TESIS	20
CAPÍTULO 2. MARCO TEÓRICO	21
2.1 LA OPINIÓN EN REDES SOCIALES	21
2.2 EL ANÁLISIS DE OPINIÓN EN LAS REDES SOCIALES.....	22
2.3 EL LENGUAJE EN LAS REDES SOCIALES	24
2.4 NIVELES DEL LENGUAJE NATURAL	25
2.4.1 Nivel Fonético / fonológico.....	26
2.4.2 Nivel morfológico	26

2.4.3	<i>Nivel sintáctico</i>	26
2.4.4	<i>Nivel semántico</i>	27
2.4.5	<i>Nivel pragmático</i>	27
2.4.6	<i>Nivel discursivo</i>	28
2.5	SISTEMAS CORRECTORES Y DE NORMALIZACIÓN DE PALABRAS	28
CAPÍTULO 3. ESTADO DEL ARTE		29
CAPÍTULO 4. ANÁLISIS Y DISEÑO DE LA APLICACIÓN		40
4.1	PROPÓSITO DE LA APLICACIÓN.....	40
4.2	PREPROCESAMIENTO.....	41
4.2.1	<i>Eliminación de texto no útil</i>	42
4.2.2	<i>Normalización y acentuación de términos</i>	43
4.2.3	<i>Algoritmo de Levenshtein</i>	45
4.2.4	<i>Diagrama de secuencia del preprocesamiento</i>	46
4.3	SELECCIÓN DE CARACTERÍSTICAS	47
4.3.1	<i>Reglas para la selección de características</i>	50
4.4	PROCESAMIENTO Y DETERMINACIÓN DE CLASE.....	53
4.4.1	<i>Primera fase de evaluación: Presencia o ausencia de polaridad</i>	54
4.4.2	<i>Segunda fase de evaluación: orientación global positiva, negativa o neutral</i>	55

4.4.3	<i>Tercera fase de evaluación: polaridad final</i>	56
4.4.4	<i>Diagrama de secuencia</i>	56
CAPÍTULO 5. IMPLEMENTACIÓN DEL SISTEMA		58
5.1	SECUENCIA LÓGICA GENERAL DEL SISTEMA.....	58
5.2	RECURSOS LINGÜÍSTICOS.....	60
5.2.1	<i>Diccionario de frecuencia de palabras de la REA</i>	60
5.2.2	<i>Diccionario de Freeling</i>	61
5.2.3	<i>Corpus de entrenamiento y de pruebas</i>	62
5.2.4	<i>Estándar de oro</i>	62
5.3	ESTRUCTURA DEL SISTEMA.....	65
5.3.1	<i>Preprocesamiento</i>	65
5.3.2	<i>Selección de características</i>	66
5.3.3	<i>Clasificación de los documentos</i>	66
5.4	SALIDAS	67
5.5	CONFIGURACIONES ALTERNAS	68
CAPÍTULO 6. PRUEBAS Y RESULTADOS		71
6.1	CORPUS USADO EN LA EVALUACIÓN	71
6.2	MECANISMO DE EVALUACIÓN	75
6.2.1	<i>Exactitud (Accuracy)</i>	76

6.2.2	<i>Precisión</i>	76
6.2.3	<i>Exhaustividad (Recall)</i>	77
6.2.4	<i>Medida-F (F1-measure)</i>	77
6.3	PRUEBAS REALIZADAS Y RESULTADOS OBTENIDOS	78
6.3.1	<i>Pruebas y resultados de exactitud del sistema</i>	78
6.3.2	<i>Pruebas y resultados de la medida F1 sobre las polaridades</i>	81
6.4	COMPARACIÓN DE RESULTADOS CON EQUIPOS DE TASS 2014	83
CAPÍTULO 7. CONCLUSIONES		85
7.1	APORTACIONES.....	85
7.2	PRODUCTOS DESARROLLADOS.....	85
7.3	TRABAJOS FUTUROS	86
BIBLIOGRAFÍA		87
GLOSARIO		91

Índice de figuras

Figura 3.1. Vista conjunta del proyecto y flujo (Dave et al., 2003).	29
Figura 3.2 Arquitectura del sistema de clasificación de tweets (Jiménez Zafra et al., 2014).	35
Figura 4.1 Proceso que sigue una palabra en el preprocesamiento.	45
Figura 4.2 Secuencia general del algoritmo de Levenshtein.	46
Figura 4.3. Secuencia del preprocesamiento.	47
Figura 4.4 Secuencia para clasificación de los textos.	57
Figura 5.1 Secuencia lógica general del sistema.	59
Figura 5.2 Ejemplo de codificación del corpus en la sección de entrenamiento.	63
Figura 5.3 Ejemplo de codificación del corpus en la sección de pruebas. ...	64
Figura 5.4 Ejemplo del formato del archivo QREL.	64
Figura 5.5 Ilustración del preprocesamiento de un documento.	65
Figura 5.6 Ejemplo del módulo de selección de características.	66
Figura 5.7 Ejemplo del módulo de procesamiento.	67
Figura 6.1 Grafica de distribución del corpus TASS en la sección de entrenamiento.	74
Figura 6.2 Grafica de distribución del corpus TASS en la sección de pruebas.	75

Índice de tablas

Tabla 4.1 Ejemplos de eliminación de texto no útil.	42
Tabla 6.1 Distribución de tweets del corpus TASS por polaridad en la sección de entrenamiento.	73
Tabla 6.2. Distribución de tweets del corpus por polaridad en la sección de pruebas.	74
Tabla 6.3 Pruebas aleatorias de clasificación.	79
Tabla 6.4 Resultados obtenidos usando palabras como características.	80
Tabla 6.5 Resultados obtenidos usando lemas como características.	80
Tabla 6.6 Precisión, exhaustividad y medida F del sistema aleatorio, por palabras y por lemas.	81
Tabla 6.7 Precisión, exhaustividad y medida F por clase para cada sistema.	82
Tabla 6.8 Resultados obtenidos en el taller TASS 2014 y resultados obtenidos con las propuestas realizadas.	83

Capítulo 1. INTRODUCCIÓN

1.1 Antecedentes

Nos encontramos en una revolución tecnológica [23], la tercera para ser precisos, y como tal, las tecnologías en general avanzan, pero en especial las tecnologías de la información y comunicación que en consecuencia modifican el desarrollo de la sociedad, actividades tan cotidianas como la comunicación interpersonal adoptan nuevos instrumentos que la facilitan y al mismo tiempo la modifican para dar paso a nuevas formas de interacción. Sumado a lo anterior y según el sociólogo canadiense Marshall McLuhan, podríamos considerar a los medios de comunicación como una extensión de las personas, algo así como un tercer ojo o una segunda boca¹, por lo que en ese contexto podemos considerar a los teléfonos inteligentes, computadoras, así como los sistemas de cómputo, como una extensión del cuerpo del ser humano que ayudan a comunicar y a procesar la información de forma más fácil y efectiva.

Por otra parte, McLuhan también acuñó el término Aldea Global², lo cual describe las consecuencias de la transformación de la cultura material, que principalmente se ven reflejadas en los diferentes medios de comunicación que usamos hoy en día. Actualmente, es fácil comunicarse con alguien que se encuentre al otro lado del planeta, leer opiniones y comentarios de

¹ McLuhan, M. (1994). Understanding media: The extensions of man. MIT press.

² McLuhan, M., & Powers, B. R. (1996). La aldea global. Barcelona.

personas que usaron o compraron productos o servicios, incluso sin conocer a la gran mayoría de estas personas; pero finalmente, existe esa facilidad de comunicación que convertiría al mundo actual en una gran aldea global.

Dentro de la tercera revolución tecnológica, encontramos a internet y a lo que se conoce como Web 2.0 [2], que desde sus inicios en 2003 el usuario adquiere un nuevo rol, pasando de ser consumidor de contenidos a ser también productor de los mismos, de receptor pasivo a emisor de juicios y opiniones. Posteriormente, con la creciente moda por el uso de redes sociales, que ha ido en aumento de 2005 a la fecha, se reconfiguró nuevamente la participación del usuario promedio, permitiendo que sus opiniones, pensamientos y sentimientos, lleguen a más personas en la red, así como la creación de círculos sociales virtuales con usuarios con quienes podía tener una mayor afinidad.

Sin duda, las redes sociales contribuyeron a fomentar la evolución de los medios de comunicación electrónicos en su forma escrita, al mismo tiempo ha generado interés de otras áreas del conocimiento, siendo la lingüística computacional una de esas áreas.

En la lingüística computacional o procesamiento del lenguaje natural, que se interesa por estudio y la modelación del lenguaje humano mediante métodos computacionales, podemos ubicar la tarea de minería de opinión y análisis de sentimiento, que cobró un especial interés en la web 2.0 y las redes sociales por la posibilidad de explorar la inmensa cantidad de información que se genera en ellas todos días.

El análisis de opiniones de usuarios a cerca de productos o servicios que proporciona una empresa, es una actividad que se ha realizado tradicionalmente de diferentes formas, en diferentes etapas y contextos, ejemplo de ello son los estudios de mercado, estudios de impacto de un

producto en el mercado, análisis de resultados, etcétera. Las herramientas utilizadas van desde los formularios en papel, entrevistas y sondeos, hasta formularios electrónicos; aunque gracias a los avances en el procesamiento del lenguaje natural, se abren nuevas posibilidades para el análisis de opinión mediante las redes sociales.

Si usualmente, el recabar opiniones y analizarlas involucra una interacción directa con el usuario (costosa en tiempo y dinero), ahora y mediante las redes sociales surge una nueva opción, analizar la opinión textual de los usuarios, quienes muchas veces dan su opinión y expresan su sentir sobre el producto o servicio consumido. Esta nueva opción es particularmente importante para empresas que desean conocer la opinión sincera de sus consumidores, políticos que cuidan su imagen ante los electores, o incluso para cualquier persona interesada en adquirir información sobre algo que desconoce.

El análisis de opiniones en internet y redes sociales sobre productos (considerando que productos pueden ser objetos de consumo, servicios, música, películas, eventos sociales o incluso la imagen de una persona) es algo que la mayoría de las personas con acceso a internet realiza casi de forma instintiva en situaciones con poca o nula información del producto.

Como seres humanos, y diferenciándonos de otras especies, tenemos la capacidad de reflexión y análisis de información, capacidad que se ha desarrollado con el propósito de transformar la información que recibimos en conocimiento, y posteriormente usar ese conocimiento en beneficio propio o colectivo [3], sin embargo, tenemos la desventaja de pagar un alto costo en tiempo y esfuerzo al realizar esa tarea, limitando la cantidad de información que podemos analizar; por otra parte, las computadoras a pesar de tener la capacidad de procesar una gran cantidad de información, estas no cuentan

con la capacidad de reflexión, por lo que crear sistemas que en cierta forma les dé esa capacidad representa un gran avance y utilidad.

Páginas web como city-data.com, yelp.com y tripadvisor.com dan información a los usuarios y les permite opinar sobre algún tema de interés en específico. Redes sociales como Facebook y Twitter, permiten una interacción más amplia y diversa entre los usuarios, cuentan con millones de suscriptores en todo el mundo, por lo que se pueden encontrar opiniones más diversas, sobre más temas, aunque más subjetivas y complejas.

1.2 Planteamiento del problema

Los seres humanos somos seres sociales, por tal motivo, al realizar tareas como comprar un celular, algún electrodoméstico, votar por algún político o cualquier otra situación de la cual no tenemos conocimiento sólido, realizamos una consulta en nuestros círculos sociales (familia y amigos) que nos ayuden a tomar la decisión. Así mismo, atendiendo a la necesidad de socializar con las personas en nuestros círculos sociales, damos nuestra opinión sobre las buenas o malas adquisiciones que realizamos.

Con el aumento de los beneficios y posibilidades de interacción de los usuarios en internet y el crecimiento de las redes sociales, las opiniones individuales y la comunicación interpersonal, que comúnmente se realizaba en círculos sociales pequeños, se modifican de tal forma que en la actualidad una opinión realizada en alguna red social virtual como Twitter o Facebook puede llegar a millones de personas dependiendo de la popularidad del emisor y/o el impacto del mensaje.

El análisis de información y el proceso de transformación de la información en conocimiento, son capacidades exclusivas de los seres humanos; involucran y relacionan numerosos procesos como la memoria, análisis de contexto, ventajas, desventajas, apropiación del conocimiento, entre otros aspectos. El análisis de opiniones en texto es una tarea personal que en la actualidad realizamos de forma constante, pero realizarla usando un programa de cómputo, que analice automáticamente las opiniones, resulta ser una tarea sumamente complicada. Además, se podría decir que dicha tarea es un campo de reciente exploración, por lo cual, constantemente surgen nuevas propuestas y herramientas.

Por lo anterior, contribuir mediante una propuesta que facilite el análisis de las opiniones clasificándolas en una escala como: muy positivas, positivas, neutras, negativas o muy negativas, y que además realice la tarea con buena precisión sería una buena aportación en el campo.

1.3 Objetivos generales y específicos

Objetivo general:

Desarrollar una aplicación de software capaz de analizar textos cortos del idioma español extraídos de la red social Twitter y determinar la polaridad de cada documento.

Objetivos particulares:

- Investigar métodos de estado de arte que resuelven el mismo problema,
- Encontrar o desarrollar un corpus que se adapte a nuestros propósitos,

- Determinar cuál es línea base a superar para la solución del problema,
- Proponer una metodología alterna en comparación las existentes,
- Desarrollar un sistema que implemente la metodología propuesta,
- Realizar experimentos con el sistema desarrollado y analizar los resultados,
- Comparar los resultados obtenidos con los resultados de otros sistemas en un contexto similar.

1.4 Justificación

El análisis de información y la toma de decisiones son actividades que forman parte de la vida cotidiana de toda persona, aunque siempre existirá un límite en la capacidad humana. Las capacidades humanas pueden ser ampliadas mediante el uso de la tecnología como equipos y sistemas de cómputo. Por otra parte, explorar y explotar los datos que se generan en las redes sociales todos los días se ha convertido en un tema de mucho interés por parte de particulares, empresas y académicos.

Twitter es una de las redes sociales más populares de la actualidad, incluso en muchas ocasiones los medios masivos de comunicación tradicionales (radio y televisión) usan como referencia los *trending topics* (tendencias) y noticias difundidas en dicha red. Twitter es usado para expresar opiniones, juicios, pensamientos e información de interés particular y social, siendo los “tweets” (mensajes de texto de hasta 140 caracteres) su principal cualidad.

La estructura de los comentarios en Twitter, si la tienen, es diversa debido a la restricción de caracteres permitidos; por lo que la realización de minería de opinión puede tener diferentes aproximaciones y metodologías, de esta forma, en la presente tesis se propone una metodología donde los términos

de un mensaje son considerados como multipolares (muy positivo, positivo, neutro, negativo, muy negativo o sin polaridad) y se crea un programa que clasifica textos cortos, siendo la multipolaridad la base del sistema.

Es un hecho que el análisis de opiniones y sentimientos sobre corpus en el idioma español se ha dado en mucha menor medida que en el idioma inglés, por lo que es grato trabajar en el idioma español, no solo por ser el idioma oficial de México, sino que también por ser la segunda lengua más hablada en el mundo después del mandarín³.

1.5 Alcances y límites

Se ha construido un clasificador automático de textos cortos del idioma español. Este clasificador se ha entrenado y probado sobre el corpus de *tweets* que proporciona la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) en su taller anual. El corpus cuenta con más de 68,000 *tweets* recolectados entre noviembre de 2011 y marzo de 2012, los cuales fueron realizados por personalidades y celebridades de habla hispana sobre diferentes temas.

Por otra parte, Twitter impone restricciones⁴, algunas de ellas son: el número de búsquedas (permitiendo únicamente 180 para usuarios y 450 para aplicaciones), otra restricción es que prohíbe la distribución⁵ del texto de los

³ Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>.

⁴ <https://dev.twitter.com/rest/public/rate-limiting>

⁵ <https://dev.twitter.com/es/overview/terms/agreement-and-policy>

mismos, pero si permite la distribución del Id de usuario y del tweet, haciendo posible recuperar el texto de los mismos.

1.6 Estructura de la tesis

Capítulo 1. Muestra la información general acerca del proyecto: planteamiento, objetivos generales y específicos, justificación, alcances y límites.

Capítulo 2. Presenta información acerca de la opinión en redes sociales, retos al trabajar con redes sociales virtuales y niveles del lenguaje natural.

Capítulo 3. Presenta el estado del arte con respecto al análisis automático de opiniones en redes sociales.

Capítulo 4. Se presenta nuestra aproximación para el análisis automático de opiniones en redes sociales.

Capítulo 5. Describe la implementación del sistema y los recursos adicionales utilizados.

Capítulo 6. Presenta las pruebas realizadas al sistema y los resultados obtenidos.

Capítulo 7. Enlista las conclusiones y plantea el trabajo a realizarse a futuro.

Capítulo 2. MARCO TEÓRICO

En este capítulo se describen conceptos fundamentales en torno al análisis de opinión en la red social Twitter. Además, se describen los principales métodos y recursos que serán de apoyo para el desarrollo de la herramienta de software.

2.1 La opinión en redes sociales

Una red social de personas puede ser algo complejo de definir, ya que puede tener diferentes tamaños y alcances según las características que se consideren en la red social. Algunas de características pueden ser el parentesco, la relación que se lleva con la persona para considerar parte de la red social, el estatus social, la frecuencia con la que se tiene contacto con la persona o la cercanía emocional. Los antropólogos Hill & Dunbar [24] estiman que el promedio de una red social de un individuo es alrededor 153 personas e influyen factores como la edad, estado civil, género, personalidad, nivel educativo, ocupación e ingresos.

La opinión, es una capacidad que todo ser humano practica cotidianamente de diferentes formas, es un enunciado que se afirma como verdadero sin tener garantía de su validez, surge de la necesidad de comunicarse y colaborar con los demás. La opinión se puede estudiar desde un enfoque de social, humanístico, psicológico, lingüístico, filosófico, político, administrativo, de marketing, etcétera. La opinión puede representar algo tan simple y común como la cotidianeidad de una persona o algo sumamente complejo

como tratar de cuantificar algo que es subjetivo, dependiendo del campo desde donde se aborde.

La persona que emite una opinión busca aportar información con el propósito de provocar una reacción, por lo que aunque se pueda tener una opinión individual, esta adquiere mayor importancia cuando se comparte a nivel interpersonal, de aquí que, la opinión siempre se dé en círculos sociales de diferentes magnitudes que rodean al individuo. Una opinión puede ser emitida por un líder social o por una persona común, puede ser expresada de forma oral, de forma escrita o incluso de forma gesticulada.

Con el rápido crecimiento de las redes sociales virtuales, se puede decir que los modos de opinión, han sumado a sus filas la opinión escrita (electrónica) en redes sociales virtuales, abriendo nuevas posibilidades tanto para quien emite el mensaje, como para quienes fungen como receptores del mensaje.

2.2 El análisis de opinión en las redes sociales

Las redes sociales (comunidades virtuales de comunicación en internet) son una herramienta que permite a las personas comunicarse e interactuar, en ellas se puede escribir o leer opiniones de otros usuarios sobre algún tema de interés. Si bien, la opinión en redes sociales se colocó rápidamente en un lugar privilegiado en la sociedad en general, también despertó el interés por parte de la comunidad académica y empresarial, al llegar con ellas la posibilidad de medir y analizar las opiniones de libre acceso.

El análisis de opinión en las redes sociales, o minería de opinión y análisis de sentimiento aplicado a redes sociales, es un campo joven y que a la fecha se encuentra desarrollo, por lo que, día con día se proponen métodos que

permiten un mejor análisis de texto y mejores resultados. No obstante, sigue siendo una tarea difícil debido a la subjetividad que esta tarea implica, e incluso, los seres humanos muchas veces no coincidimos cuando intentamos clasificar una opinión o comentario.

Se puede decir que la minería de opinión y el análisis de sentimiento inicio formalmente a partir de 2001 con trabajos como el de C. Cardie “*Combining low-level and summary representations of opinions for multi-perspective question answering*”, de S. Das y M. Chen “*Yahoo! for Amazon: Extracting market sentiment from stock message boards*”, de K. Dave, S. Lawrence y D. M. Pennock “*Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*”, entre muchas otras publicaciones posteriores que prosiguieron sobre el tema [5].

El análisis automático de opiniones de productos en redes sociales es una actividad que a la fecha ha ido en aumento, incluso páginas web como socialmention.com, semantria.com, bottlenose.com, voxco.com o sentimentalytics.com, ofrecen herramientas que permiten el monitoreo de opiniones en redes sociales.

Existen diferentes aproximaciones relacionadas con el tema, de las cuales, la mayoría se centra en el análisis de opinión de textos del idioma inglés, aunque muchas veces también es posible adaptar estas metodologías al idioma español. Por otra parte, abordar al problema de análisis de opinión en redes sociales en el idioma español resulta interesante y provechoso, esto debido al escaso trabajo que hay en este idioma, además de la posibilidad de crear metodologías alternas, nuevas o complementarias al abordar el tema desde una perspectiva diferente. Según Lera Boroditsky, el lenguaje que hablamos afecta nuestra percepción del mundo y nos obliga a desarrollar diferentes habilidades cognitivas [25], lo que se traduce en perspectivas diferentes.

Los textos de Twitter tienen la peculiaridad de ser cortos (máximo 140 caracteres), por lo que se tiene menos texto, y por tanto, se tiene menos indicios que aporten información relevante, haciendo más difícil su análisis y su clasificación. Realizar la tarea de análisis y clasificación de opiniones sobre textos cortos se puede volver un tema muy complejo, subjetivo e incluso difuso.

2.3 El lenguaje en las redes sociales

El lenguaje humano es un sistema de comunicación estructurado que sigue reglas combinatorias de signos y contexto de uso, ha sido una pieza fundamental para transmitir conocimientos de generación en generación y gracias a ello las sociedades contemporáneas son posibles.

El lenguaje siempre ha estado expuesto a diversos factores que lo afectan y lo modifican. En la actualidad, podemos ver como constantemente el lenguaje escrito es modificado en las redes sociales, se agregan o se adoptan nuevas palabras, se dan nuevos significados a otras, entre otros fenómenos.

El lenguaje que se usa en redes sociales no es cuidado y muchas veces es creado para ahorrar caracteres, agilizar y facilitar la comunicación. Los usuarios en redes sociales no están condicionados a seguir reglas o convenciones de la escritura, lo que aumenta la complejidad de las tareas que implican trabajar con este texto.

Si hacer que las computadoras, de cierta forma, entiendan el lenguaje ya es una tarea complicada, hacer que entiendan el lenguaje escrito de las redes sociales es aún más complejo, ya que muchas veces no se cuenta con una

estructura en los textos o características necesarias para ser entendido en algún nivel como se ve aborda en el siguiente punto.

2.4 Niveles del lenguaje natural

La tarea principal de la lingüística computacional consiste en la creación de modelos que sean entendibles para las computadoras. Como lo argumentan el Dr. Gelbukh y el Dr. Sidorov [26], se puede tratar de desarrollar un modelo de lenguaje completo, sin embargo, es preferible dividir el objeto en partes y construir modelos más pequeños y simples del lenguaje que lo describan.

El lenguaje natural se puede dividir en 6 niveles:

1. Nivel fonético / fonológico
2. Nivel morfológico
3. Nivel sintáctico
4. Nivel semántico
5. Nivel pragmático
6. Nivel discurso

Cada uno de estos niveles tiene características que lo definen, pero también cuentan con similitudes que pueden compartir en diferentes niveles, esto como consecuencia de pertenecer a algo que las engloba, el lenguaje natural. A continuación una breve descripción de cada nivel con base en el libro: Procesamiento automático del español con enfoque en recursos léxicos grandes (A. Gelbukh, G. Sidorov, 2010).

2.4.1 Nivel Fonético / fonológico

En este nivel se exploran las características del sonido como parte esencial del lenguaje hablado, por lo que se realizan implementaciones relacionadas con sistemas de reconocimiento de voz y síntesis del habla. Existiendo mayor éxito en la tarea de síntesis de voz que la de reconocimiento.

En este nivel también se estudia la posición del sonido, en relación con otros sonidos, en comparación con otros idiomas.

2.4.2 Nivel morfológico

En este nivel se explora la estructura interna de las palabras como sufijos, prefijos, raíces y flexiones; y las categorías gramaticales como género y número. Aquí se puede estudiar a una gran diversidad de lenguas, algunas con similitudes y otras con muchas diferencias en relación con las reglas que las rigen.

Los principales problemas que resuelven se relacionan con el desarrollo de sistemas de análisis y síntesis morfológica automática. Aunque existe la metodología y hay sistemas funcionando para muchos idiomas, hace falta una estandarización de módulos.

2.4.3 Nivel sintáctico

En el nivel de la sintaxis se analizan las relaciones entre las palabras dentro de la frase y generalmente se usa alguno de los dos modelos principales para la representación de las relaciones, dependencias y constituyentes, en el primero se marcan las relaciones entre palabras con flechas, en el segundo, las relaciones se marcan en forma de árbol binario.

Las principales tareas que se exploran aquí se relacionan con métodos para análisis y síntesis automática de texto, siendo una tarea más fácil el desarrollo de generadores, que el desarrollo de los analizadores sintácticos o *parsers*, que aún es un problema abierto.

2.4.4 Nivel semántico

En el nivel de la semántica el objetivo es “entender” la frase, por lo que se busca identificar el sentido de todas las palabras e interpretar las relaciones sintácticas, buscando obtener como resultado *redes semánticas*, que representan los conceptos y las relaciones entre ellos del texto analizado. También se pueden obtener *grafos conceptuales*, que son muy parecidos a las redes semánticas.

Otras tareas en el nivel de la semántica son la definición del sentido de las palabras, tarea complicada aún para los seres humanos; y la desambiguación automática de sentidos de palabras, tarea que ni siquiera los seres humanos podemos realizar si no existe un contexto.

2.4.5 Nivel pragmático

En este nivel se trata de establecer las relaciones entre la oración y el mundo externo, o dicho de otra forma, lo que interesa a la pragmática son las intenciones del autor del texto o del hablante. Las oraciones que tienen como característica particular ser acciones por sí mismas o performativas, son otro ejemplo de exploración en este nivel.

Debido a que existen números tropiezos a nivel semántico, es muy difícil continuar la cadena de análisis hasta este nivel.

2.4.6 Nivel discursivo

En este nivel se amplía el ámbito de exploración, es decir, se consideran varias oraciones y ya no solo una. Dichas oraciones, mantienen relaciones entre sí y se forma algo conocido como discurso.

Un problema que se aborda en este nivel es la resolución de correferencia o también llamadas relaciones anafóricas. Aunque existen algoritmos que alcanzan hasta 90% de exactitud en la solución de correferencia, resolver el 10% restante aun es una tarea difícil.

2.5 Sistemas correctores y de normalización de palabras

Dentro del procesamiento del lenguaje natural existen métodos para corregir los errores de ortografía y también para normalizar palabras.

Algoritmos como el de Levenshtein, que miden la distancia de una palabra a otra, realizan dicha tarea. Para ello es necesario contar con enorme corpus del texto en el idioma en que se está trabajando, o un diccionario de palabras. El algoritmo, básicamente toma una palabra A, que se va a corregir, y busca en sus recursos una palabra B que sea igual o se asemeje con menos cambios, tomando la palabra con la que se realizan menos modificaciones para ser iguales.

Por otra parte, la normalización de términos que contienen caracteres repetidos, se puede solucionar de forma sencilla mediante un sistema con reglas y excepciones predefinidas.

Capítulo 3. ESTADO DEL ARTE

Los primeros trabajos que relacionan con la minería de opinión en internet surgen junto con la web 2.0 y con las nuevas capacidades que adquiere el usuario, la posibilidad de ser receptor y productor de opiniones a la vez. En ese contexto, Kushal Dave et al., publican “*Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*”, artículo en cual abordan la riqueza que existe en internet sobre opiniones de productos y proponen una herramienta que selecciona y sintetiza las opiniones.

El método que proponen Dave et al. (2003) [15], se basa en aprendizaje automático, para lo cual, el proceso inicia con el uso de opiniones estructuradas para pruebas y entrenamiento, continua con la identificación de características y medición de métodos adecuados para finalmente determinar si las opiniones son positivas o negativas. Su método queda ilustrado en la figura 3.1.

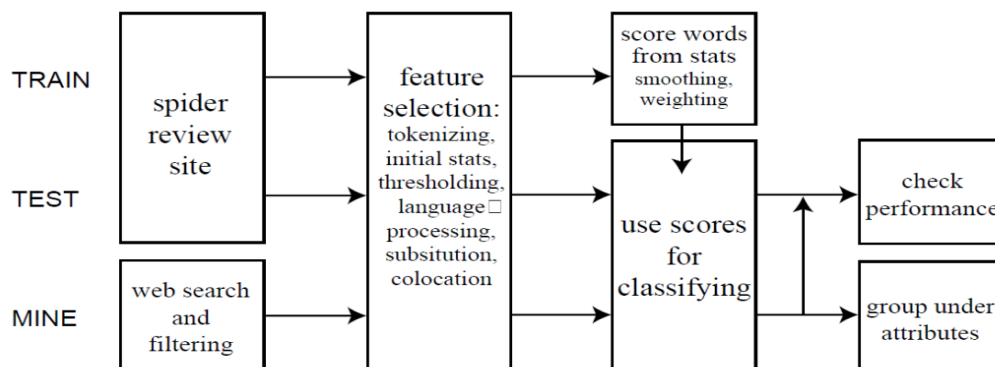


Figura 3.1. Vista conjunta del proyecto y flujo (Dave et al., 2003)

En el año 2005, Jonathan Read en su publicación “*Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification*” expone que los métodos propuestos hasta entonces para la clasificación de sentimientos habían demostrado resolver con éxito el problema, pero bajo la condición de tener que existir una buena relación entre los datos de entrenamiento y los de prueba, por lo que un sistema entrenado con datos de reseñas de películas no obtendría tan buenos resultados si se prueba con reseñas de automóviles.

Además de la dependencia temática, también propone que existe una dependencia en el dominio y en el tiempo, es decir, un clasificador entrenado con opiniones de productos no es efectivo para evaluar el sentimiento de artículos de noticias, además, este sería efectivo solo por un lapso de tiempo. En el artículo se propone una metodología para determinar la polaridad por medio de aprendizaje automático con datos de entrenamiento etiquetados, donde se usa al lenguaje en conjunción con emoticonos tomando como corpus de entrenamiento el texto usado en *Usenet newsgroups* [7].

En 2009 Alec Go et al., en su artículo “*Twitter Sentiment Classification using Distant Supervision*” abordan la clasificación de sentimientos, positivos y negativos sobre textos cortos de la red social Twitter, con el propósito de lograr que su aproximación sea útil para consumidores que buscan opiniones antes de adquirir algo, o para empresas que buscan monitorear las opiniones acerca de sus marcas. El equipo propone una metodología usando *machine learning* para clasificar el sentimiento en los mensajes de Twitter, esto mediante el uso de supervisión distante sobre los términos de consulta; el corpus de entrenamiento que usan consiste en mensajes de Twitter que contienen emoticonos, se basan en mensajes de este tipo para determinar la polaridad de las palabras del mensaje y alimentar al corpus de entrenamiento.

En su aproximación, establecen como línea base una categorización simple basada en el número existencias de palabras, positivas y negativas clasificadas por Twittratr (sistema clasificador de *tweets*), para determinar la polaridad. Hacen un comparativo de resultados obtenidos con clasificadores Naive Bayes, entropía máxima y Maquinas de Soporte Vectorial; consideran unigramas, bigramas y etiquetado de partes de oración; los mejores resultados fueron obtenidos con los clasificadores Naive Bayes y Entropía Máxima usando unigramas y bigramas en conjunto como características para el entrenamiento [8].

En 2010, Alexander Pak y Patrick Paroubek, en su publicación “*Twitter as a Corpus for Sentiment Analysis and Opinion Mining*”, retoman el tema de la clasificación de *tweets*, recolectando mensajes de Twitter para usarlos como corpus. Construyen una herramienta que clasifica los mensajes en positivos, negativos y neutros, usando los algoritmos Naive Bayes y SVM [9].

En 2013, Grigori Sidorov et al., en su publicación “*Empirical Study of Machine Learning for Opinion Mining in Tweets*”, describen los resultados obtenidos de una serie de experimentos en los que demuestran como diferentes factores afectan la precisión de los algoritmos de aprendizaje máquina. Para experimentación usaron un corpus de 32,000 *tweets* en español de los cuales 8,000 fueron clasificados manualmente para el entrenamiento en las categorías: positivo, negativo, neutral o noticioso. Los algoritmos de aprendizaje que implementaron fueron: *Naïve Bayes*, *Decision Tree* y *Support Vector Machines*. Los factores que modificaron fueron: el tamaño de los n-gramas, el tamaño del corpus, el número de clases de sentimiento, el balance de los corpus y probaron varios dominios, con la finalidad de encontrar el mejor balance de configuraciones para la clasificación de *tweets* en español [4].

Otras investigaciones que se enfocan en el idioma español son promovidas por la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), quienes cada año, desde 2012, organizan un taller de análisis de sentimiento [10], en donde los participantes proponen diferentes metodologías para resolver alguna de las tareas propuestas con respecto a textos cortos de Twitter, siendo la tarea de análisis de sentimiento y clasificación de temas las más populares.

El Taller de Análisis de Sentimientos de la SEPLN o TASS, cuenta con su propio corpus creado para las diferentes tareas, con el que los participantes realizan experimentos y evalúan resultados. En la edición 2014, propusieron 4 tareas diferentes: Análisis de sentimiento a nivel global con 5 y 3 etiquetas, clasificación de temas, detección de aspectos y análisis de sentimiento basado en aspectos, de las cuales, el análisis de sentimiento a nivel global es la tarea que tuvo mayor participación.

En la última edición del TASS, el grupo ELiRF-UPV integrado por Lluís-F. Hurtado y Ferran Pla, de la Universidad Politécnica de Valencia, obtuvo el mayor puntaje en la mencionada tarea sobre otros 6 equipos participantes. A continuación se da una breve descripción de las metodologías propuestas por los equipos que participaron en el Taller.

El equipo ELiRF-UPV de la Universidad Politécnica de Valencia integrado por Lluís-F. Hurtado y Ferran Pla, en su publicación “ELiRF-UPV en TASS 2014: Análisis de Sentimientos, Detección de Tópicos y Análisis de Sentimientos de Aspectos en Twitter” [16] describen las metodologías propuestas para todas las tareas del taller, siendo la tarea de determinación de polaridad la que más nos interesa. El equipo ELiRF-UPV realiza un preprocesamiento de los textos antes de abordar cualquier tarea, para ello adaptaron un tokenizador *tweets* llamado *Tweetmotif*, usaron *Freeling* como lematizador, detector de entidades nombradas y etiquetador morfosintáctico. Abordan la tarea de

determinación de polaridad como un problema de clasificación, usando máquinas de soporte vectorial (SVM) junto con diccionarios de polaridad de lemas y de palabras. Realizan tres aproximaciones diferentes en la primera y segunda usan uni-gramas, mientras que en la tercera usan n-gramas: En la primera aproximación toman como características los coeficientes td-idf de los lemas de las palabras que aparecen en el *tweet*, más el número de lemas positivos y negativos con base al diccionario de lemas; En la segunda aproximación usan como características los coeficientes td-idf de las palabras más el número de palabras positivas y negativas según el diccionario de palabras; En la tercera aproximación realizan una valoración de 6 sistemas diferentes con 1, 2 y 3 gramas de lemas y palabras, de donde eligen la polaridad con mayor valor.

El equipo de *Elhuyar Fundazioa*, integrado por Iñaki San Vicente Ronzal y Xabier Saralegi Urizar, participó exclusivamente en la primera tarea de TASS 2014, en su artículo “Looking for Features for Supervised Tweet Polarity Classification” (Buscando Características para Clasificación Supervisada de Polaridad de Tuits) [17] describen su aproximación, la cual básicamente consistió en la creación de un sistema basado en máquinas de soporte vectorial (SVM) que combina la información extraída a partir de léxicos de polaridad con características lingüísticas.

El léxico de polaridad usado consiste en una compilación propia (*ElhPolar*), junto con otros recursos léxicos de polaridad existentes como *Mihalcea's Lexicon*, *SO-CAL lexicón*, y *Spanish Emotion Lexicon*, con el propósito de ampliar la cobertura de léxico. Usan como base el “ElhPolar”, que es una construcción de léxico de polaridad del equipo Elhuyar usado en la edición 2013, fue creado traduciendo un léxico existente en inglés y extrayendo palabras positivas y negativas del corpus de entrenamiento de TASS, además, las polaridades fueron corregidas manualmente, también fue enriquecido con frases comunes compuestas del español que fueron

compiladas manualmente, sumado a eso, los léxicos de polaridad existentes que usan para ampliar la cobertura son: *Mihalcea's Lexicon "full strength"* (Perez-Rosas, Banea, and Mihalcea, 2012), *SO-CAL lexicon* (Taboada et al., 2011), *Spanich Emotion Lexicon (SEL)* (Sidorov et al., 2013). Usan la implementación SMO de los algoritmos de Maquinas de Soporte Vectorial de Weka en su sistema supervisado de clasificadores. Las características lingüísticas que evalúan en el sistema son n-gramas del español con significado especial, por ejemplo "Valer la pena" o "perro faldero", otra características que consideraron son los signos de puntuación, en específico el uso de signos de exclamación e interrogación, y finalmente consideran la negación como un modificador de polaridad, para ello crean el equivalente negado de cada característica y léxico en su modelo de aprendizaje.

El equipo LyS, del Departamento de Computación, de la Universidad da Coruña e integrado por David Vilares, Yeraí Doval, Miguel A. Alonso y Carlos Gómez-Rodríguez, en su publicación "LyS at TASS 2014: A Prototype for Extracting and Analysing Aspects from Spanish tweets" (LyS en TASS 2014: Un prototipo para la extracción y análisis de aspectos en tuits) [18], propone una aproximación basada en aprendizaje automático para la solución de la tarea de análisis de sentimiento a nivel global. Para ello, primero realizan un preprocesamiento del texto de los *tweets*, seguido de un etiquetado de partes de la oración y análisis de dependencias, cabe destacar que únicamente usaron el corpus oficial de entrenamiento etiquetado, además de adaptar un etiquetador de partes de oraciones *the Brill (1992) tagger* incluido en el *NLTK framework* con datos de entrenamiento de *The Ancora corpus* (Taulé, Martí, and Recasens, 2008), y finalmente, para el análisis de dependencias usaron *MaltParser* (Nivre et al., 2007) y *Ancora corpus*.

El equipo SINAI-ESMA, de la Escuela Politécnica Superior de Jaén, España, integrado por Salud María Jiménez, Eugenio Martínez, M. Teresa Martín y L. Alfonso Ureña, en su publicación "SINAI-ESMA: An unsupervised approach

for Sentiment Analysis in Twitter” (SINAI-ESMA: Una aproximación no supervisada para análisis de sentimiento en Twitter) [22], explican su aproximación no supervisada para la tarea de clasificación global de opiniones expresadas en textos cortos, que se basa en el uso de léxico de opinión y aplicación de una heurística sintáctica. Lo interesante de esta publicación es que son pocos los sistemas que abordan el problema desde una perspectiva no supervisada.

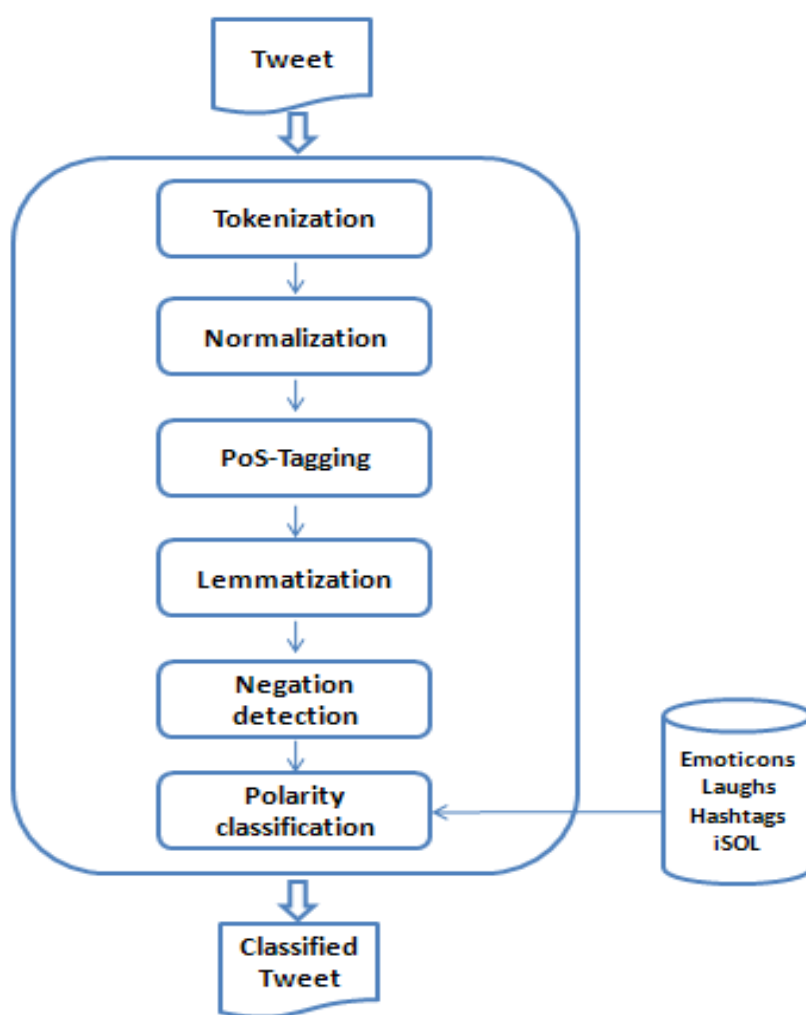


Figura 3.2 Arquitectura del sistema de clasificación de *tweets* (Jiménez Zafra et al., 2014).

El proceso que siguen es el siguiente: tokenización del *tweet*, normalización del texto, etiquetado de partes de oración, lematización, detección de negación, clasificación de polaridad con apoyo de emoticonos, hashtags y texto positivo como “jajaja”. Para la tokenización usan una adaptación al idioma español de *the Christopher Potts’ tokenizer*, que puede consultarse en la web sentiment.christopherpostts.net/; para la normalización del texto mal escrito programaron un corrector basado en distancia, desarrollado por Peter Norvig (norvig.com/spell-correct.html), basando en parte su programa en el recurso lingüístico iSOL, que contiene una lista de palabras representativas de opinión en el español (Dolores-Molina et., al, 2014). Posteriormente etiquetan las partes de la oración y obtienen el lema de cada token, seguido de una detección de negación que influirá en la determinación de la polaridad final, para finalmente decidir mediante su sistema la polaridad de *tweet*. Su proceso se puede apreciar mejor en el siguiente diagrama de la figura 3.2.

El equipo JRC-IPSC de la Comisión Europea del Centro de Investigación Conjunta y el Instituto para la Protección y Seguridad del Ciudadano, integrado por José M. Perea-Ortega y Alexandra Balahur, en su publicación “*Experiments on feature replacements for polarity classification of Spanish tweets*” (Experimentos sobre sustituciones de características para la clasificación de *tweets* en español) [19], proponen una sustitución de características (signos de puntuación repetidos, emoticonos y palabras de opinión) para los corpus proporcionados usando aprendizaje automático para realizar la clasificación de los *tweets*, en específico *Support Vector Machine Sequential Minimal Optimization* (SVM SMO) de WEKA (www.cs.waikato.ac.nz/ml/weka).

El proceso que siguieron consistió en realizar un preprocesamiento tanto para el corpus de entrenamiento como para el corpus de pruebas, en cual removieron direcciones web, números y normalizaron los símbolos y letras repetidos. No realizaron *stemming* y algunas palabras no útiles fueron

removidas para algunos experimentos usando una lista pequeña de *Snowball* (snowball.tartarus.org/algorithms/spanish/stop.txt) para español, lista que fue modificada manualmente removiendo 228 de las 325 palabras. En cuanto a recursos semánticos utilizaron USELESP (*Unified SEntiment LExicon for Spanish*), herramienta que desarrollaron con el propósito de integrar recursos semánticos existentes en diferentes lenguajes en un léxico único del español para análisis de sentimientos, en este caso *Spanish JRC* (Steinberger et al., 2011) y *eSOL lexicon (enriched Spanish Opinion Lexicon)* (Molina-González et al., 2013), también utilizaron *SentiStrength* (Thelwall et al., 2010) que consiste en 106 emoticonos relacionados con un peso de sentimiento (1 y -1). El propósito del reemplazo es poner bajo la misma etiqueta a diferentes características con el mismo sentimiento, para reducir el número de características durante el proceso de aprendizaje. Los reemplazos realizados fueron signos de puntuación repetidos, emoticonos por positivos y negativos por su equivalente en palabra, y finalmente palabras afectivas por su equivalente polar (*hpositive*, *positive*, *hnegative* o *negative*), según la categoría asignada por su recurso USELESP. Adicionalmente usaron una técnica llamada *skip-grams*, formando bi-gramas, tri-gramas y sus equivalentes en saltos de gramas, *1-skip-grams*, *2-skip-grams*, etcétera. Realizando diferentes experimentos con la técnica de reemplazo y de salto de gramas, obteniendo los mejores resultados al usar simultáneamente uni-gramas y bi-gramas conjuntamente.

El equipo CINVESTAV-IPN, del Instituto Politécnico Nacional de México, integrado por Roberto Hernández Petlachi y Xiaou Li, en su publicación "Sentiment analysis of texts in spanish based on semantic approaches with linguistic rules" (Análisis de sentimiento sobre textos en Español basado en aproximaciones semánticas con reglas lingüísticas) [20] explican su propuesta la cual consiste en aproximaciones semánticas, etiquetación morfológica y orientación semántica con métodos supervisados, realizando

primeramente un preprocesado, seguida de un lematización usando la herramienta Freeling, posteriormente tokenización, segmentación y etiquetación, para finalmente aplicar reglas lingüísticas para obtener la polaridad.

En la etapa de preprocesamiento el equipo realizó una corrección de signos de puntuación, reemplazo de direcciones web por la cadena “enlace”, reemplazo de emoticonos por su equivalente en palabras con existencia en su diccionario de orientación semántica, corrigieron las abreviaturas como “q” (que) y “xq” (porque), corrección de gramática mediante el algoritmo de Levenshtein usando como base el Corpus de Referencia del Español Actual (CREA) de la Real Academia Española, además de la normalización de caracteres repetidos como vocales. Una vez realizado el preprocesamiento, determinan los conceptos con carga emocional de cada oración para identificar su carga emocional de acuerdo a un léxico afectivo llamado SODictionariesV1.11Spa de orientación semántica, realizando la sumatoria de todas las palabras de opinión que se encuentran en los diccionarios. Además, identifican los intensificadores de sentimiento (términos capaces de ampliar o disminuir la intensidad emocional del texto que afectan) asociando un porcentaje al intensificador tanto positivo como negativo, que permite aumentar o disminuir la afección de los elementos involucrados y que afectan la determinación final de polaridad de un texto. Adicionalmente integran la detección de negación, que consiste en invertir la polaridad de las palabras que son exclusivamente afectadas, usando un diccionario de su autoría, construido con las posibilidades de negación en el español.

El equipo SINAI Word2Vec, de la universidad de Jaén en España, integrado por A. Montejo-Ráez, M.A. García-Cumbreras y M.C. Díaz-Galiano, en su publicación “*SINAI Word2Vec participation in TASS 2014*” (Participación de SINAI Word2Vec en TASS 2014) [21] explican su aproximación para la tarea de determinación de sentimiento a nivel global, la cual basaron en un método

supervisado con el uso de Support Vector Machines (SVM) sobre la sumatoria de vectores de palabras con un modelo generado a partir de Wikipedia en español, siendo interesante que no aplican análisis sintáctico ni análisis léxico. El sistema Word2Vec representa las palabras mediante vectores de espacio continuo, basada en el modelo de bosas de palabras o n-gramas, incluyendo skip-gramas.

Capítulo 4. ANÁLISIS Y DISEÑO DE LA APLICACIÓN

El propósito de la aplicación es determinar la clase a la que pertenecen los textos a procesar, para ello se ha desarrollado una aplicación que está basada en aprendizaje automático supervisado. Se hacen necesarias las tareas de preprocesamiento, selección de características que aporten información para la determinación de clases y un sistema de clasificación automática.

En este capítulo se analiza el proceso que sigue la aplicación y la metodología usada para resolver las tareas que involucra.

4.1 Propósito de la aplicación

El propósito de la aplicación es clasificar textos cortos o *tweets* de carácter general, es decir, el tema de cada texto puede ser política, entretenimiento, economía, deportes, tecnología, música, entre otros; y las clases o categorías que se consideran son P+, P, NEU, N, N+ y NONE (muy positivo, positivo, neutro, negativo, muy negativo y sin polaridad). Para la clasificación se adopta una metodología poco convencional que consiste en considerar a los términos como multiclase, es decir, a pesar de algunos términos sean usados ampliamente en textos positivas, también existe una baja frecuencia de uso en textos negativos o sin polaridad, que es útil sumada a la frecuencia de otros términos de la misma clase.

Por otra parte, un producto puede ser representado por un bien o servicio de consumo o incluso la imagen de una persona (políticos, cantantes, actores, etcétera), por lo que en este contexto, es importante que la aplicación no únicamente se centre a algo específico como música, películas o bienes de consumo, sino que abarque distintos ámbitos.

4.2 Preprocesamiento

La solución al problema de la clasificación de textos cortos (*tweets*) mediante aprendizaje automático tiene varias aproximaciones. La mayoría de las propuestas, si no es que todas, realiza un preprocesamiento del texto debido a toda la jerga que se genera en la red social Twitter, y si a eso le sumamos el hecho de que muchos usuarios interactúan en diferentes redes sociales con diferentes convenciones e idiomas, haciendo del lenguaje escrito una deformación y mutación que es necesario corregir para facilitar la aproximación del problema.

Derivado de lo anterior, se hace necesario contar con un módulo que realice la normalización de texto, en nuestro caso y debido a que adoptamos el modelo de bolsa de palabras se incluye una tokenización de las palabras, normalización del formato de las palabras y obtención de la forma canónica de las palabras.

El preprocesamiento del texto es necesario tanto para la parte del corpus de entrenamiento como para la parte de pruebas, por lo que es necesario crear una función que realice dicha función de forma automática y constantemente durante casi todo el proceso de clasificación de polaridad. A continuación se listan las tareas a realizar dentro del módulo de preprocesamiento.

4.2.1 Eliminación de texto no útil

Consideramos necesaria la eliminación de texto que no aporta información relevante para nuestro modelo de clasificación, por lo que direcciones web, nombres de usuario y *hashtags*, serán excluidos de los textos, esto debido a que las direcciones web no aportan información, a menos que se exploren los sitios citados; algunos nombres de usuario podrían ser considerados como características, afectando el resultado por la cantidad de veces que aparecen en los tweets; y los *hashtags* son creaciones espontáneas o modas esporádicas que muchas veces no son trascendentes en el lenguaje.

En los ejemplos de la tabla 4.1 se puede observar como algunos textos pueden mutar al aplicar algún algoritmo de corrección de texto, por ejemplo el hashtag #VeoTV y las direcciones web, que cambiaron a voto y ftp respectivamente en una prueba realizada con el algoritmo de Levenshtein. También se puede observar que algunos textos de tema (*hashtag*) y nombres de usuario, se mantienen y solo pierden el proceso los símbolos que identifica su tipo (# y @), dejando la posibilidad de ser consideradas como características más adelante, lo que en cualquiera de los casos representarían interferencia en el proceso de clasificación.

Tabla 4.1 Ejemplos de eliminación de texto no útil.

Texto original	Texto aplicando eliminación y normalización de términos	Texto sin aplicar eliminación y normalización de términos
Salgo de #VeoTV , que día más largooooo!!...	Salgo de que día más largo	Salgo de voto que día más largooooo...
@MauperezMk Preciosa de verdad - Perdóneme by Pablo Alborán, from	Preciosa de verdad perdóneme by pablo alborán, from	Mauperezmk Preciosa de verdad - Perdóneme by Pablo Alborán, from

#SoundHound http://t.co/u0CAbR3X		soundhound ftp
Doloooo de cabeza para Merkel.? F.Holland http://t.co/WO3MB4VU	Dolor de cabeza para merkel f holland	Doloooo de cabeza para Merkel.? F.Holland ftp

Para implementar la eliminación de términos no útiles se ha programado un buscador que identifica las características de los enlaces, usuarios y *hashtags* que pueden ser fácilmente detectados por los símbolos “http://”, “www”, “@” y “#” y posteriormente son borrados, dando paso a un nuevo texto. De tal forma que un documento d , formado por términos t_i , sería el resultado de la diferencia de d y los términos t_i que tienen similitud con algún elemento de los textos no útiles:

$$d = d - t_i, \text{ si } \{ "http://" \text{ or } "www" \text{ or } "@" \text{ or } "\#" \} \in t_i \quad (4.1)$$

Donde d es el documento de texto corto o *tweet* y t_i es una unidad o palabra que pertenece a d .

4.2.2 Normalización y acentuación de términos

La normalización de términos se aplica a aquellos que han sido modificados por la repetición de caracteres y signos de puntuación, por ejemplo, en la tabla 4.1 se muestran las palabras “largooooo...” y “Doloooo”, de las cuales es fácil deducir que equivalen al adjetivo “largo” y al nombre “dolor”, por lo que la normalización hace necesaria una función que reduzca el número de caracteres repetidos y elimine los signos de puntuación, a excepción de algunas letras, las cuales se usan de forma doble y son comunes en muchas palabras del idioma español, estas letras son “c”, “e”, “l” y “r”.

Debido a que en el idioma español se usa una acentuación gráfica, las vocales acentuadas cambian su valor numérico y por lo tanto cambian al

término, de ahí que una computadora considere diferente la palabra “reelección” de la palabra “reeleccion”, en consecuencia aunque “reelección” se encuentre dentro del vocabulario de la computadora, la computadora establece que “reeleccion” no se encuentra, haciendo necesario acentuar las palabras si en las primeras búsquedas no se tuvo éxito. La figura 4.1 ilustra un diagrama y un ejemplo para las tareas de normalización y acentuación.

Derivado de lo anterior, surge la hipótesis de que el uso de letras y signos repetidos en las palabras, e incluso onomatopeyas, pueden ayudar a determinar del sentimiento o polaridad de un texto, siempre que se manejen como intensificadores de sentimiento, algo similar a las aproximaciones que se proponen en los trabajos [28] y [29], por ejemplo un simple “hola” puede considerarse más intenso cuando se añaden más caracteres o signos de admiración “holaaaaa!!!!”, o un “detestoooo” puede ser más intenso que un “detesto”, o incluso el uso de letras mayúsculas puede ser interpretado como estar gritando, ya que se ha adoptado ese convencionalismo últimamente en redes sociales, pero nosotros no abordaremos esa hipótesis en esta ocasión.

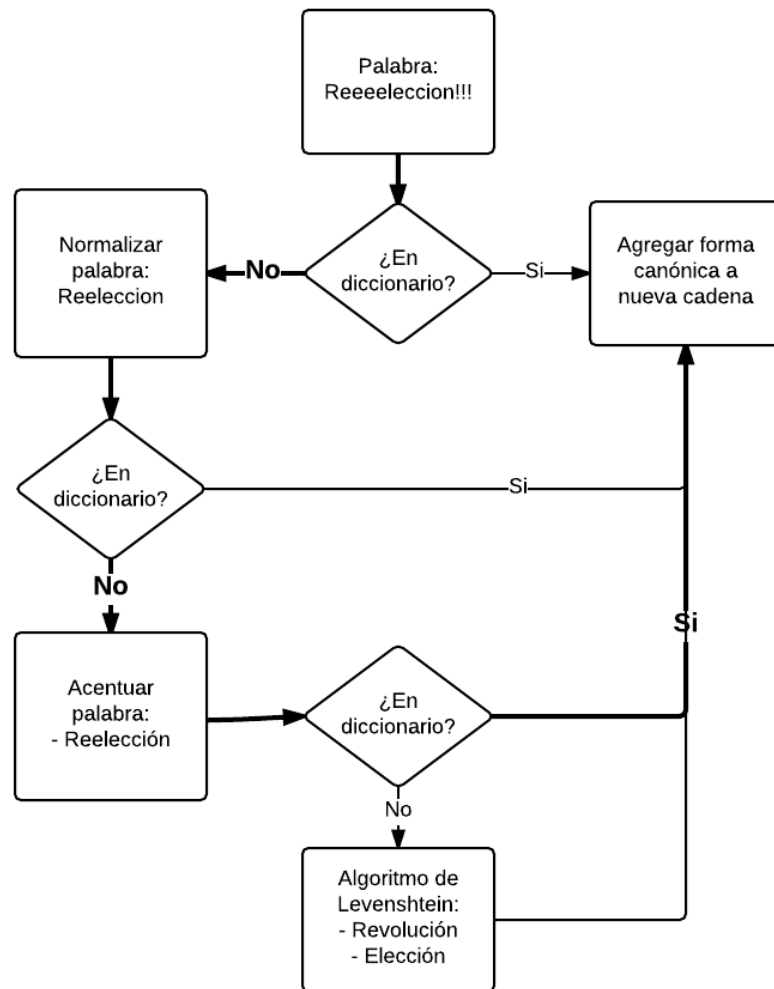


Figura 4.1 Proceso que sigue una palabra en el preprocesamiento.

4.2.3 Algoritmo de Levenshtein

La metodología que desarrolló Vladimir Levenshtein [30] en los años 60 es ampliamente conocida y usada en la corrección de textos. Básicamente consiste en realizar operaciones de sustitución, inserción y extracción de caracteres encontrar el número mínimo cambios o la distancia mínima entre 2 palabras. Para nuestra implementación, se planea tomar como fuente las formas flexionadas de las palabras del diccionario de Freeling [27], así como

la metodología propuesta por Peter Norvig [31], para la corregir las palabras que no superen las fases de búsqueda, normalización y acentuación dentro del preprocesamiento. En la figura 4.2 se ilustra el proceso que sigue una palabra.

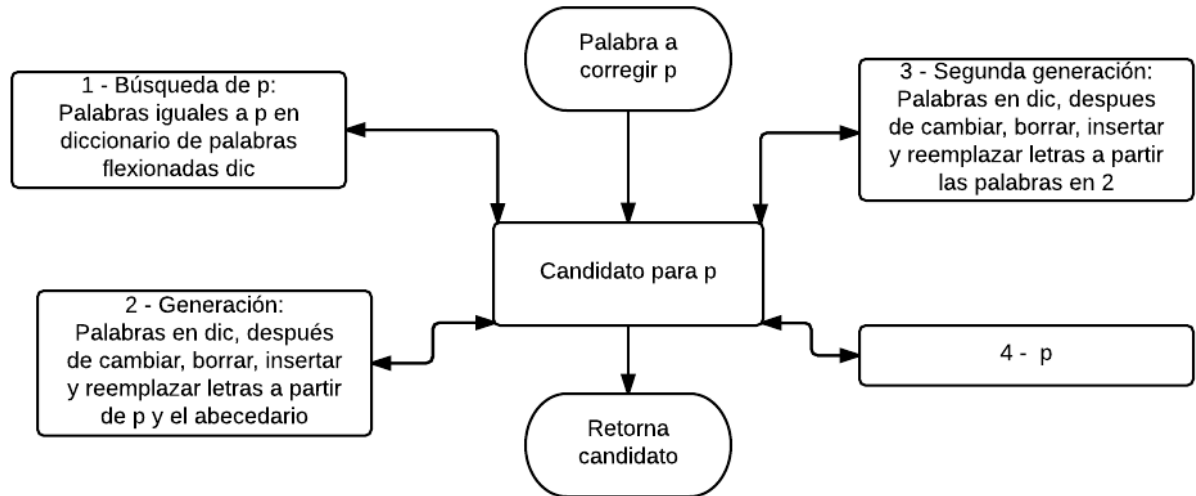


Figura 4.2 Secuencia general del algoritmo de Levenshtein.

4.2.4 Diagrama de secuencia del preprocesamiento

En la figura 4.3 representamos el proceso general para el preprocesamiento de los textos cortos de Twitter. Dicho proceso es necesario tanto para el corpus de entrenamiento como para el corpus de prueba.

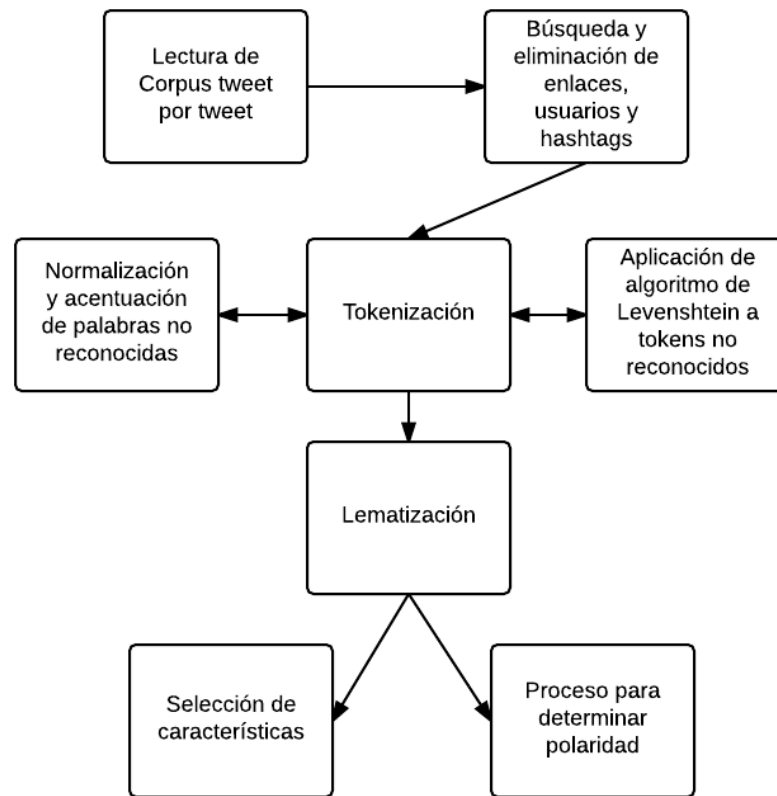


Figura 4.3. Secuencia del preprocesamiento.

4.3 Selección de características

En la etapa de selección de características se remueven los atributos irrelevantes de una representación, esto tiene varias funciones, entre ellas como medida de protección de saturación de atributos, también como mejoramiento de la eficiencia computacional, debido a que un elevado número de características se traduce un mayor número de recursos computacionales para resolver el problema.

Existen dos formas de elegir los atributos [12]: Selección de un Subconjunto de características (*Feature Subset Selection*), donde la nueva representación consiste de un subconjunto de los atributos originales; y construcción de características (*Feature Construcción*), donde las nuevas características son introducidas por la combinación original de características.

Específicamente, la selección de un subconjunto de características no toma en cuenta palabras que no aportan información adicional al documento, también son conocidas como palabras auxiliares o *stopwords*. En el caso del idioma inglés, ejemplos de este tipo de palabras son: “*the*” y “*and*”, que en español son “La” e “Y”, las cuales en ambos idiomas son irrelevantes independientemente de la tarea de clasificación, así que estas palabras serán removidas del diccionario de características útiles. También podemos encontrar dentro de la tarea de selección de características el establecimiento de un umbral de frecuencia en textos o *document frequency thresholding*, en el cual se eliminan las palabras infrecuentes, es decir, todas las palabras que ocurran menos de un determinado número veces (n) en el corpus de entrenamiento. Variando n se puede reducir drásticamente el número de características.

En nuestro caso, optamos por una eliminación de palabras auxiliares (*stopwords*), proceso automatizado que no requiere recursos adicionales, y establecer un umbral de frecuencia de palabras en los textos (*document frequency thresholding*) con algunas modificaciones que se adaptan a nuestras necesidades. Ambas tareas se realizan automáticamente después de construir una matriz de frecuencia de término basado en la polaridad $PTF(t,p)$ y establecer algunos criterios que determinan la utilidad de una palabra.

$$PTF(t, P) = \frac{PTF(t, p_i)}{PTF(t, C)} \quad (4.2)$$

Donde t es un término, P son las polaridades a las que puede pertenecer el término, p_i es una polaridad, $PTF(t,P)$ es la frecuencia del término dividido en polaridades, $PTF(t,p_i)$ es la frecuencia del término en una polaridad, C es el corpus de entrenamiento y $TF(t,C)$ es la frecuencia del término en el Corpus de entrenamiento.

Una vez obtenidos los atributos del corpus de entrenamiento se considera al texto de las publicaciones a evaluar como pequeñas bolsas de palabras que contienen características de diferentes clases en variadas proporciones, es decir, tratamos el problema a nivel de palabras o nivel de morfología léxica, lo que hace al texto más accesible para los algoritmos de aprendizaje automático. De esta forma, cada palabra o término t es tratado como un tipo de atributo. El valor del atributo para un documento d , será el número de veces que ocurre en los documentos de la misma clase dividido por el número total de ocurrencias, algo similar a la frecuencia de término (*Term Frequency* $TF(t, d)$), con la variante de que la frecuencia se distribuye sobre las clases que se evalúan.

Al tratar el problema como un modelo de bolsa de palabras se espera tener un equilibrio entre expresividad y complejidad del modelo, debido a que las palabras son buenas unidades representativas, la razón es porque la que las palabras dan una razonable granularidad en la representación de documentos o textos se puede encontrar en la evolución de los lenguajes. Además, las palabras son los elementos en donde la sintaxis y la semántica se encuentran [12].

Los términos o palabras consideradas como no útiles son descartadas por el sistema, mediante reglas que se aplican al momento de seleccionar características en el corpus de entrenamiento, dichas reglas se explican en el punto siguiente.

4.3.1 Reglas para la selección de características

Para que un término sea parte del conjunto de características, es necesario que supere un proceso y las reglas establecidas. Para ello se considera al corpus de entrenamiento como una enorme bolsa de palabras donde:

1. Se obtiene la frecuencia del término t en el corpus C_{train} sin importar la clase.

$$PTF(t, C_{train}) = \sum t \mid t \in C_{train} \quad (4.3)$$

Donde t es un término, C_{train} es el corpus de entrenamiento, C_{train} es el corpus de entrenamiento y $PTF(t, C_{train})$ es la frecuencia del termino en el corpus en corpus de entrenamiento.

Ejemplo:

Considerando que la distribución de la polaridad en el ejemplo siguiente, y los posteriores, es la siguiente: PALABRA (P+, P, NEU, N, N+, NONE). Tenemos que:

Palabra_1 (5, 15, 2, 4, 1, 0, 1); donde la frecuencia o el número de menciones total del término en el corpus de entrenamiento es 28. Por lo tanto $PTF(t, C_{train}) = 28$.

2. Se neutralizan los términos que tienen una frecuencia menor o igual a 6, pudiendo incrementar o disminuir el valor para variar el número características.

$$PTF(t, C_{train}) = \begin{cases} PTF(t, C_{train}) & \mid PTF(t, C_{train}) > 6 \\ 0 & , \quad i. o. c \end{cases} \quad (4.4)$$

Donde $PTF(t, C_{train})$ es la frecuencia del término en el corpus de entrenamiento, 6 es una constante que se puede modificar y 0 es el valor que se asigna al término, dejando descartado al término del grupo de características.

Tomando el ejemplo anterior, Palabra_1 superaría sin problema la regla, ya que tiene una frecuencia mayor a 6.

3. Obtener la frecuencia del término por clase y dividir cada frecuencia entre la frecuencia total para obtener un peso distribuido en las clases almacenado en un vector. Posteriormente neutralizar los términos que tengan más de 2 pesos en su vector con valores entre .20 y .50. La razón es que si pertenece a muchas polaridades, su utilidad disminuye al poder neutralizarse a sí mismo.

$$PTF(t, P) = \begin{cases} [0,0,0,0,0,0] & | (PTF(t, p_i) \in .20 \sim .50) > 2 \\ PTF(t, P) & , \quad i. o. c. \end{cases} \quad (4.5)$$

Donde $PTF(t, P)$ es el vector de pesos por polaridades o categorías, $[0,0,0,0,0,0]$, es un vector de polaridad vacío para un término, .20 ~ .50 es un umbral preestablecido que se puede variar para admitir mayor o menor número de características y 2 es una constante que se puede variar con el propósito de tener mayor o menor número de características.

Ejemplos:

Palabra_1 (0.1786, 0.5357, 0.0714, 0.1428, 0.0, 0.0358); pasa la prueba

Palabra_2 (0.21, 0.34, 0.0, 0.25, 0.0, 0.20); no pasa la prueba por que aporta información a clases positivas, negativas y sin polaridad.

4. Se suman los pesos P+ y P y los pesos negativos N y N+ para formar 2 categorías, “positivos” y “negativos”. La regla es que los pesos de estas

categorías no deben estar en el umbral de **.30** a **.60** y que la suma de ambos debe ser menor al **.60**. Lo anterior con el propósito de eliminar los términos que se neutralizan entre sí y permitir características potencialmente útiles en las clases NEU y NONE.

$$Positivos = PTF(t, p_0) + PTF(t, p_1) \quad (4.6)$$

Donde $PTF(t, p_0)$ es el peso del termino t en la clase P+, $PTF(t, p_1)$ es el peso del termino t en la clase P.

$$Negativos = PTF(t, p_3) + PTF(t, p_4) \quad (4.7)$$

Donde $PTF(t, p_3)$ es el peso del termino t en la clase N, $PTF(t, p_4)$ es el peso del termino t en la clase N+.

$$(4.8)$$

$$PTF(t, P) = \begin{cases} [0,0,0,0,0,0] , & (positivos, negativos) \in .30 \sim .60 \\ PTF(t, P) , & i. o. c. \end{cases}$$

Donde $[0,0,0,0,0,0]$ es el vector de peso a asignar si positivos y negativos se encuentra en el rango establecido y $PTF(t, P)$ es el vector de pesos de un término.

Ejemplos:

Palabra_1 (7143, 0.25); es válida porque sus valores están fuera del umbral

Palabra_3 (0.05, 0.1, 0.1, 0.475, 0.275, 0.0)

Palabra_3 (0.15, 0.75); es válida porque sus valores no están en el umbral

Palabra_4 (0.125, 0.25, 0.125, 0.0, 0.125, 0.375)

Palabra_4 (0.375, 0.125); es válida porque aunque uno de sus valores está en el umbral, la suma de ambos es menor a .60.

5. Se agrega el término al conjunto de características si su vector de pesos no es neutro, de lo contrario fue descartada anteriormente.

$$Dic_{características}[t] = PTF(t, P) \quad (4.9)$$

Donde $Dic_{características}$ es el diccionario que almacena el conjunto de características validas, t es el término y $PTF(t, P)$ es el vector polaridades del término.

Mediante el proceso y reglas anteriores que establece si un término es válido o no, se realiza una depuración automática de palabras auxiliares (*stopwords*) y de términos que no están dentro del umbral de frecuencia establecido (*document frequency thesholding*), ya que no aportan información relevante para determinar la clase de un texto.

4.4 Procesamiento y determinación de clase

El módulo de procesamiento está destinado a determinar la clase o polaridad de un texto, previa creación del conjunto de características útiles. En esta etapa básicamente se analiza el texto del tweet a clasificar en una polaridad, se preprocesa, se examinan las características presentes en el tweet y se determina la clase de acuerdo a las características presentes o ausentes.

4.4.1 Primera fase de evaluación: Presencia o ausencia de polaridad

En esta fase se forman dos categorías con la sumatoria de los valores de las características del documento d , las categorías son “con polaridad” ($conPol$) y “sin polaridad” ($sinPol$), para ello se suman los pesos $P+$, P , N y $N+$ de los términos en el documento para la categoría de “con polaridad”, mientras que NEU y $NONE$ conforman la categoría “sin polaridad”.

(4.10)

$$conPol = \sum_{i=1}^N PTF(t_i, p_0) + \sum_{i=1}^N PTF(t_i, p_1) + \sum_{i=1}^N PTF(t_i, p_3) + \sum_{i=1}^N PTF(t_i, p_4)$$

Donde $conPol$ es la sumatoria de los valores de las características con polaridad presentes en el documento, t_i es un término que pertenece a un documento, P_0 , P_1 , P_3 y P_4 representan a las polaridades $P+$, P , N y $N+$ respectivamente, N es el número de características que contiene el documento y $PTF(t_i, p_j)$ representa el valor de la clase para cada término.

(4.11)

$$sinPol = \sum_{i=1}^N PTF(t_i, p_2) + \sum_{i=1}^N PTF(t_i, p_5)$$

Donde $sinPol$ es la sumatoria de los valores de las características sin polaridad presentes en el documento, t_i es un término presente en el documento, P_2 , y P_5 representan las clases NEU y $NONE$ respectivamente, N es el número de características que contiene el documento y $PTF(t_i, p_j)$ representa valor de la clase para cada término.

(4.12)

$$OrientaciónGlobal = \begin{cases} NONE & , \quad conPol \leq sinPol(x) \\ 2da \text{ fase} & , \quad i. o. c \end{cases}$$

Debido a que existe un desequilibrio entre las dos nuevas categorías, se multiplica a *sinPol* por un valor *x* que se puede modular para compensar el desequilibrio. De esta forma, si el valor de la categoría *conPol* es menor o igual al valor de *sinPol*, se determina automáticamente que no existe polaridad, por lo que la clase que le corresponde a ese texto es NONE, de lo contrario se pasa a una segunda fase de evaluación.

4.4.2 Segunda fase de evaluación: orientación global positiva, negativa o neutral

En esta fase y después de determinar si existe una polaridad en el texto, se procede a crear 2 nuevas categorías a partir de las 6 disponibles de los términos en el documento, “positiva” y “negativa”, donde “positiva” está compuesta por P+ y P, mientras que “negativa” está compuesta por N+ y N.

$$positiva = \sum_{i=1}^N PTF(t_i, p_0) + \sum_{i=1}^N PTF(t_i, p_1) \quad (4.13)$$

$$negativa = \sum_{i=1}^N PTF(t_i, p_3) + \sum_{i=1}^N PTF(t_i, p_4) \quad (4.13)$$

(4.14)

$$OrientaciónGlobal = \begin{cases} 3ra \text{ Fase} & , \quad positiva > negativa \\ & o \quad positiva < negativa \\ NEU, (positiva|negativa)in (.495 \sim .505) & \end{cases}$$

En esta fase, la condición para pasar a la siguiente fase es que las dos categorías de reciente creación no se encuentren en el umbral establecido, ya que en este caso se considera que el texto es neutral. De lo contrario, se

determina la orientación positiva o negativa del texto que ayudara a evaluar la clase en la tercera fase.

4.4.3 Tercera fase de evaluación: polaridad final

Si el documento en evaluación ha superado las fases anteriores, se procede a determinar la clase que le corresponde (P+, P, N+, N). Si la categoría es positiva se procede a elegir entre P+ y P tomando como base el peso de cada categoría, y lo mismo aplica en caso de que la categoría sea negativa.

(4.15)

$$OrientaciónGlobal = \begin{cases} positiva & \begin{cases} P+ : \sum_{i=1}^N PTF(t_i, p_0) \geq \sum_{i=1}^N PTF(t_i, p_1) \\ P : \sum_{i=1}^N PTF(t_i, p_0) < \sum_{i=1}^N PTF(t_i, p_1) \end{cases} \\ negativa & \begin{cases} N : \sum_{i=1}^N PTF(t_i, p_3) < \sum_{i=1}^N PTF(t_i, p_4) \\ N+ : \sum_{i=1}^N PTF(t_i, p_3) \geq \sum_{i=1}^N PTF(t_i, p_4) \end{cases} \end{cases}$$

4.4.4 Diagrama de secuencia

Las fases para clasificar un documento se resumen en el diagrama siguiente.

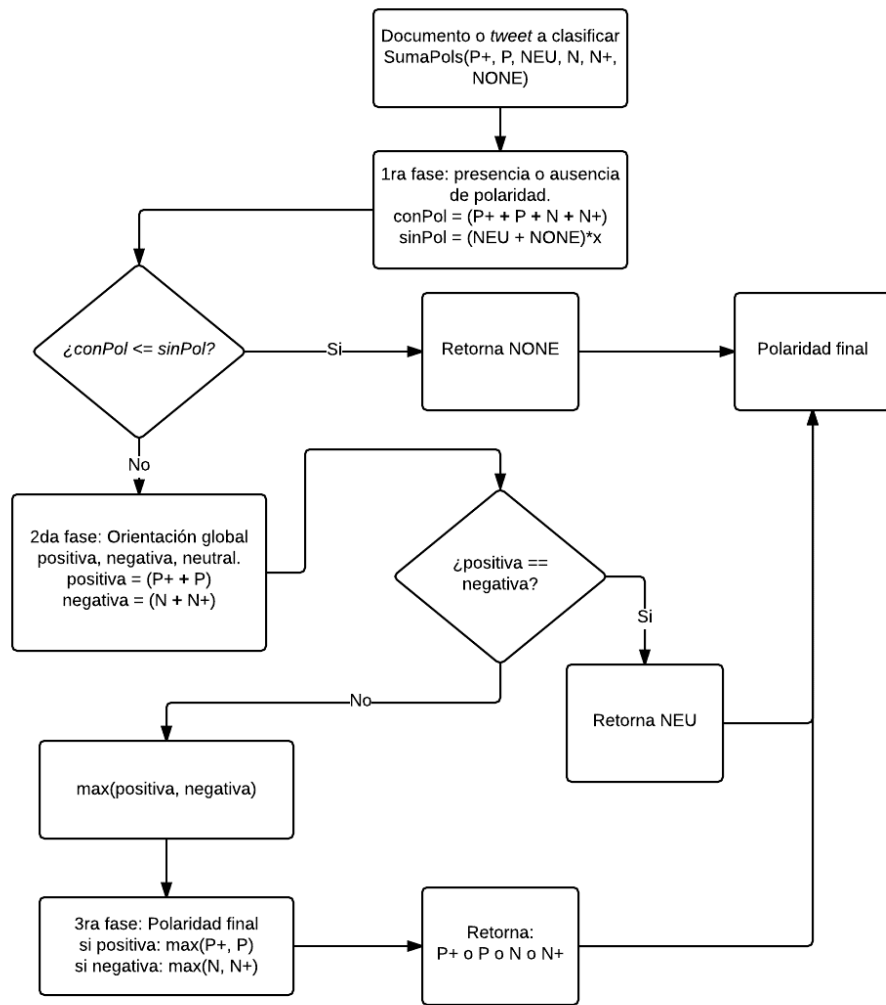


Figura 4.4 Secuencia para clasificación de los textos.

Capítulo 5. IMPLEMENTACIÓN DEL SISTEMA

Con base en lo abordado en el capítulo anterior, el sistema programado está dividido en módulos y submódulos. También se implementan recursos lingüísticos auxiliares en diferentes tareas. En este capítulo se describe la estructura del sistema y los recursos usados.

5.1 Secuencia lógica general del sistema

Con base en el análisis y el diseño de la aplicación, se determinó que la secuencia ideal es la siguiente:

- Iniciar programa
- Crear e importar recursos (diccionarios y listas)
- Si no existe un diccionario de características, se importa el corpus de entrenamiento para crearlo
- Preprocesar texto
 - Eliminación de texto no útil
 - Normalización de palabras
 - Corrección de texto
- Extraer características
 - Aplicar reglas a términos del corpus de entrenamiento
 - Crear diccionario de términos o características
- Clasificar textos cortos (corpus de pruebas)

- Preprocesar texto
- Analizar y aplicar reglas a texto
- Determinar polaridad
- Crear archivo con resultados de la clasificación realizada

El siguiente diagrama resume e ilustra la secuencia que sigue para el sistema:

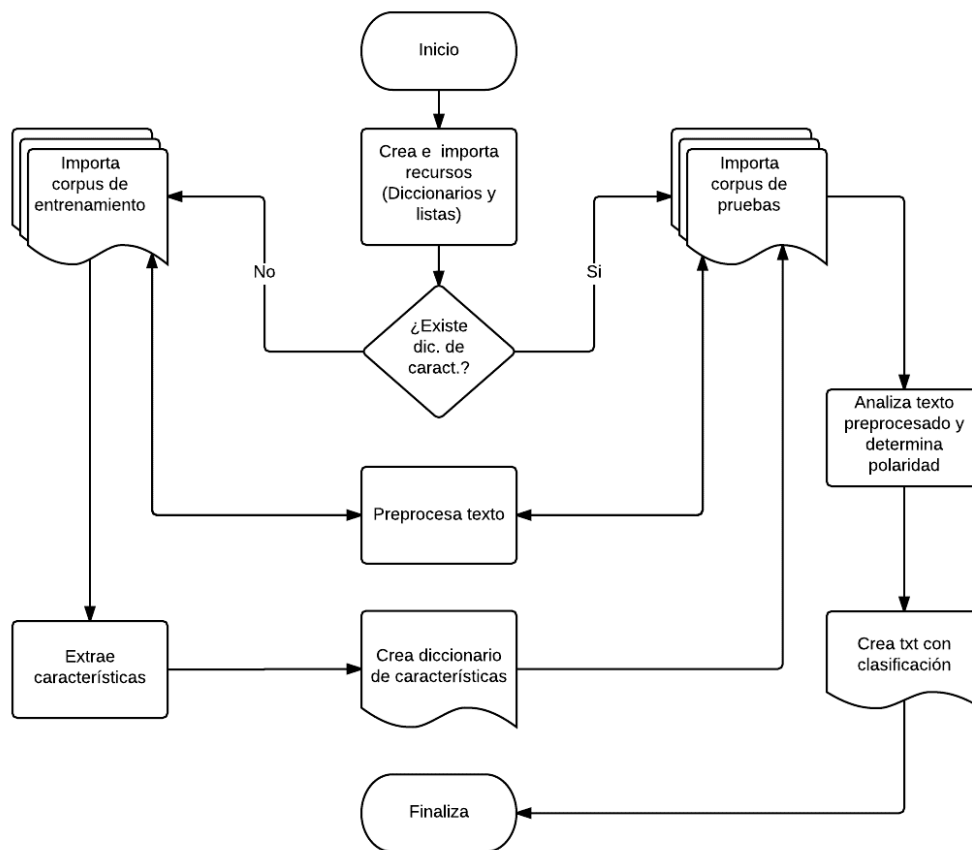


Figura 5.1 Secuencia lógica general del sistema.

Resaltan como módulos principales el preprocesamiento, la extracción o selección de características y la clasificación de los textos, mientras que la eliminación de texto no útil, normalización de las palabras y corrección de palabras son submódulos del preprocesamiento.

5.2 Recursos lingüísticos

Los recursos lingüísticos son un gran soporte para los sistemas dentro del procesamiento de lenguaje natural. Mediante el uso de diferentes recursos, dependiendo del nivel del lenguaje, se facilitan algunas de las tareas.

Algunos sistemas requieren de una gran cantidad de recursos lingüísticos para la solución de problemas, en consecuencia se pueden obtener buenos resultados pero el precio a pagar es tiempo y recursos de cómputo para el procesamiento, por otro lado, realizar un sistema que no se apoye en recursos lingüísticos probablemente sea contraproducente, demandando más recursos computacionales, además de no garantizar buenos resultados. Lo ideal es encontrar un equilibrio en el cual se obtengan buenos resultados sin una enorme dependencia de recursos lingüísticos externos.

En el sistema desarrollado se utiliza un diccionario para la corrección de frecuencia de palabras, un diccionario de lemas y el corpus para el entrenamiento y pruebas.

5.2.1 Diccionario de frecuencia de palabras de la REA

Una primera aproximación para afrontar la clasificación de textos cortos fue el uso de las palabras como características, para ello se utilizó el Corpus de Referencia del Español Actual (CREA) - Listado de frecuencias⁶, el cual fue la fuente para la corrección y normalización de las palabras.

El existen cuatro versiones de este recurso:

⁶ <http://corpus.rae.es/lfrecuencias.html>

- 1000 formas más frecuentes
- 5000 formas más frecuentes
- 10000 formas más frecuentes
- Lista total de frecuencias (737,799 formas)

De las cuales, optamos por implementar una versión ajustada a 105,700 formas, la razón es que las primeras tres versiones omiten formas que aún pueden ser útiles, mientras que las formas subsecuentes a la 105,700 son mucho menos frecuentes.

Con los primeros experimentos, un tanto empíricos, obtuvimos una exactitud promedio de 49%. Posteriormente, mediante la optimización de parámetros que se explican más adelante, alcanzamos una precisión de 54.57%, generando un diccionario de características de 616 palabras o formas. En el siguiente capítulo se muestran con mayor detalle las pruebas y resultados obtenidos.

5.2.2 Diccionario de Freeling

En una segunda aproximación se optó por la utilización de lemas como características, esto es, obtener la forma canónica de las palabras mediante un diccionario de lemas como el de Freeling⁷. La razón de usar lemas en vez de palabras es simple, existen muchas palabras que son formas flexionadas de otras, por ejemplo: “trabajaba”, “trabajé”, “trabajan” y “trabajo”, son formas flexionadas del verbo “trabajar”.

El diccionario utilizado cuenta con 556,210 formas, y después de realizar varias pruebas obtuvimos una exactitud máxima del 56.75%, generando un diccionario de características multipolares de 575 lemas.

5.2.3 Corpus de entrenamiento y de pruebas

Como se mencionó anteriormente, se usó el corpus del TASS 2014 creado por la Sociedad Española para el Procesamiento del Lenguaje Natural. A continuación se da una descripción general del corpus y en el punto 6.1 se amplía esta descripción.

El corpus del TASS 2014 cuenta con 60,017 documentos (textos cortos, comentarios o *tweets*), de los cuales 7,219 tienen el propósito de entrenar al sistema y los 60,798 restantes a probar la exactitud del sistema.

La sección de entrenamiento cuenta con las etiquetas “*tweetid*”, “*user*”, “*content*”, “*date*”, “*lang*”, “*sentiments (value, type)*” y “*topic*”, con el propósito de ser de utilidad para diferentes tareas que proponen en su taller, de las cuales únicamente necesitamos usar “*tweetid*”, “*content*” y “*sentiment (value)*”, esto debido a que solo nos enfocamos en determinar la polaridad a nivel global de los documentos.

La sección de pruebas cuenta con las etiquetas “*tweetid*”, “*user*”, “*content*”, “*date*”, y “*lang*” y únicamente nos son útiles “*tweetid*” y “*content*”, por la razón anterior.

5.2.4 Estándar de oro

Partiendo de que se utiliza un corpus etiquetado tanto para entrenamiento como para pruebas, el estándar de oro o *gold standar* es la clasificación dada a cada documento por la SEPLN. En el caso de la sección de entrenamiento del corpus, las etiquetas se encuentran dentro de archivos XML junto con las otras etiquetas que se mencionan. A continuación un ejemplo del etiquetado

⁷ <http://nlp.lsi.upc.edu/freeling/>

del corpus de entrenamiento:

```
<tweets>
  <tweet>
    <tweetid>142422495721562112</tweetid>
    <user>paurubio</user>
    <content>
      <![CDATA[Conozco a alguien q es adicto al drama! Ja ja
ja te suena d algo!]]>
    </content>
    <date>2011-12-02T02:59:03</date>
    <lang>es</lang>
    <sentiments>
      <polarity>
        <value>P+</value>
        <type>AGREEMENT</type>
      </polarity>
    </sentiments>
    <topics>
      <topic>otros</topic>
    </topics>
  </tweet>
  <tweet>
    <tweetid>142424715175280640</tweetid>
    <user>paurubio</user>
    <content>
      <![CDATA[RT @FabHddzC: Si amas a alguien, déjalo
libre. Si grita ese hombre es mío era @paurubio...]]>
    </content>
    <date>2011-12-02T03:07:52</date>
    <lang>es</lang>
    <sentiments>
      <polarity>
        <value>NONE</value>
        <type>AGREEMENT</type>
      </polarity>
    </sentiments>
    <topics>
      <topic>música</topic>
    </topics>
  </tweet>
</tweets>
```

Figura 5.2 Ejemplo de codificación del corpus en la sección de entrenamiento.

Por otra parte, el estándar de oro para la sección de pruebas del corpus cuenta con un archivo XML sin las etiquetas de las tareas a evaluar, pero además incluye un archivo de extensión “.QREL”, el cual contiene la polaridad de los *tweets*, relacionándose por medio de la etiqueta *idtweet*. A continuación un ejemplo de ambos archivos.

```

<tweets>
  <tweet>
    <tweetid>142391095542816768</tweetid>
    <user>ccifuentes</user>
    <content>
      <![CDATA[#Frailemoroso RT @JorgeNavasGarci .. algun
        alcalde que se haya adelantado a si mismo la paga de
        Navidad en agosto. #Parla si]]>
    </content>
    <date>2011-12-02T00:54:17</date>
    <lang>es</lang>
  </tweet>
  <tweet>
    <tweetid>142393574045126656</tweetid>
    <user>ccifuentes</user>
    <content>
      <![CDATA[Medir las palabras en 140 caracteres:
        http://t.co/s41k07jt]]>
    </content>
    <date>2011-12-02T01:04:08</date>
    <lang>es</lang>
  </tweet>
</tweets>

```

Figura 5.3 Ejemplo de codificación del corpus en la sección de pruebas.

165803549891117056	N+
142393574045126656	NONE
182162138301865984	N+
161114045255139328	P
159953890010337280	N+
170239713171603456	NONE
181166094193672194	N
179851844372283393	P+
142391095542816768	NONE
169397387327062017	N
149110446648066048	P+
144342754691006464	N
186441413871927296	N
163337292424032256	P+
149959551813287936	P+
172764697794314240	NONE
158832646397493249	N+

Figura 5.4 Ejemplo del formato del archivo QREL.

5.3 Estructura del sistema

Tomando como base la secuencia lógica del sistema y usando Python⁸ como lenguaje de programación para desarrollar el sistema, la aplicación consiste en un conjunto de módulos y submódulos que se apoyan en los recursos lingüísticos mencionados con la finalidad de realizar las tareas de preprocesamiento del texto, selección de características útiles y clasificación automática de textos.

5.3.1 Preprocesamiento

El preprocesamiento, como se ha mencionado en capítulos anteriores, tiene la finalidad realizar una depuración, corrección y normalización de texto de los documentos.

Este módulo, requiere de un submódulo de eliminación de texto no útil, otro de normalización de palabras y uno más de corrección de palabras. El módulo de preprocesamiento puede visualizarse como una caja negra que recibe como entrada un documento y da como salida un documento nuevo a partir de los submódulos que contiene.

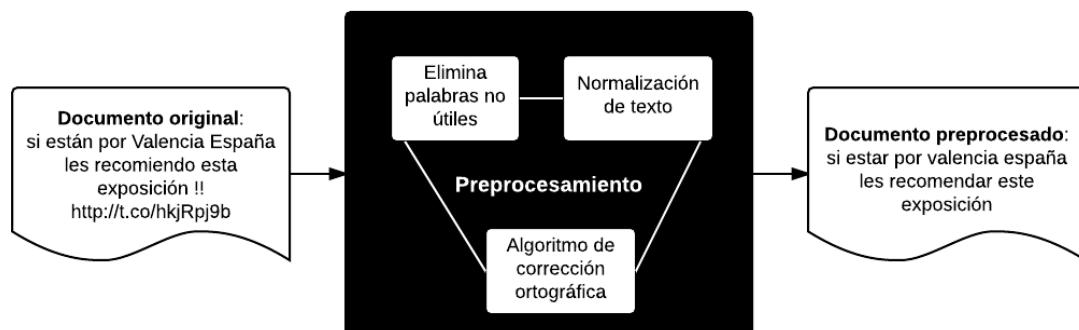


Figura 5.5 Ilustración del preprocesamiento de un documento.

5.3.2 Selección de características

La selección de características se realiza mediante un modelo de bolsa de palabras, donde cada palabra es considerada como independiente y únicamente conserva su PTF o frecuencia de término por polaridad. Que consiste en obtener la frecuencia de cada término en la colección de documentos, después el peso de cada polaridad en el término, para finalmente determinar si es útil mediante las reglas que se explicaron en el punto 4.3. En la figura 5.6 se muestra gráficamente el modulo.

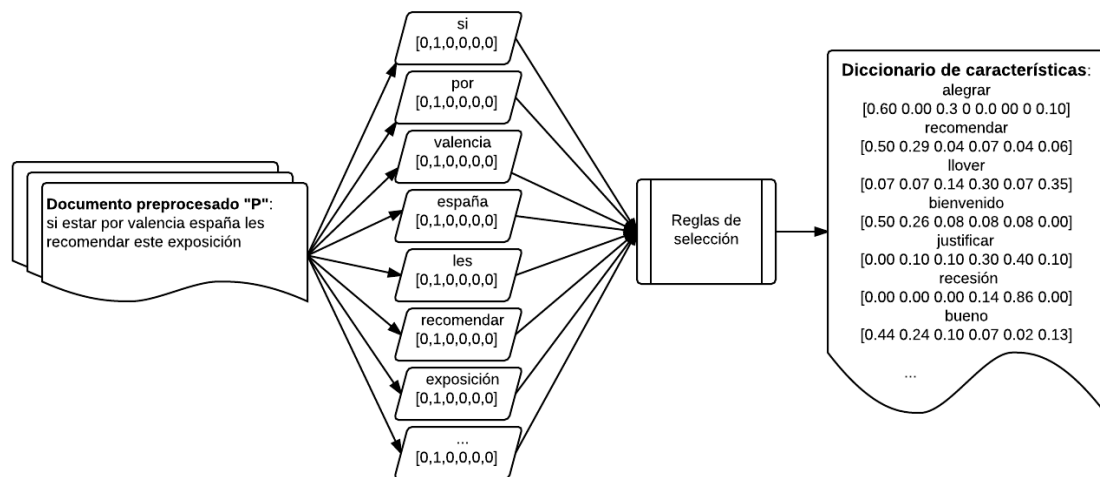


Figura 5.6 Ejemplo del módulo de selección de características.

5.3.3 Clasificación de los documentos

Para el caso de la clasificación de los textos cortos se realiza también la tarea de preprocesamiento, esto con la finalidad de estandarizar los textos, es decir, obtener documentos con palabras o características que pueden ser entendidas por el sistema y por lo tanto se pueden someter a las reglas que se establecieron en el punto 4.4, que en resumen, consisten en sumar todos los pesos de las características, tomando como base el diccionario de palabras o lemas creado a partir del corpus de entrenamiento,

posteriormente se siguen las reglas mencionadas anteriormente. En la figura 5.7 ilustra el proceso de este módulo.

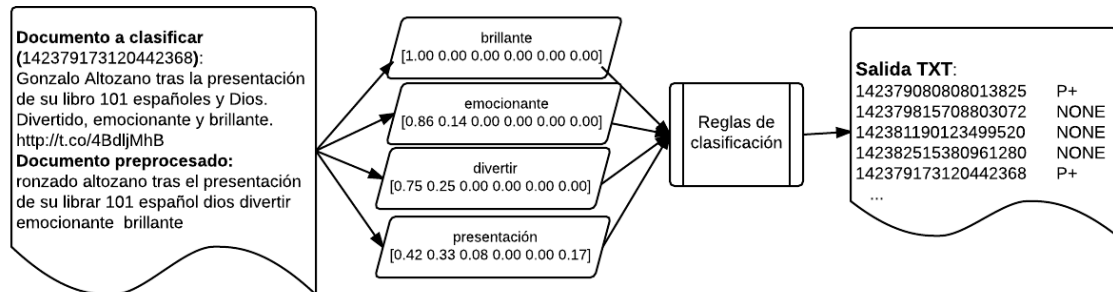


Figura 5.7 Ejemplo del módulo de procesamiento.

5.4 Salidas

El sistema genera salidas a lo largo de su ejecución, las cuales son útiles en diferentes procesos. Estas salidas son archivos planos txt que a continuación se listan:

1. Diccionario de características.
2. Documentos clasificados.
3. Evaluación de resultados y desempeño.

El diccionario de características tiene como propósito, servir como recurso para futuras implementaciones, es decir, evitar que el sistema vuelva a crear un diccionario cada vez que se ejecute. Este diccionario es creado cada vez que el sistema no encuentra este archivo.

El archivo “Documentos clasificados” tiene la función de almacenar el “*idtweet*” y “*polarity*” de cada documento que el sistema clasifica para posteriormente comprobar los resultados obtenidos.

El archivo “Evaluación de desempeño”, tiene como propósito servir como una referencia en la realización de experimentos. En este archivo incluye las medidas de exactitud, exhaustividad, precisión y medida F1.

5.5 Configuraciones alternas

Como muchos sistemas, existen parámetros que pueden ser modificados para obtener diferentes resultados. Estos parámetros pueden afectar al diccionario de características (fundamental para la clasificación) o al proceso de clasificación y se mencionan en el punto 4.3.1.

Parámetros que afectan al diccionario de características:

1. Frecuencia mínima de término.
2. Umbral de pertenencia de término a polaridades.
3. Umbral de diferencia de pesos de polaridades de término.

El parámetro “frecuencia mínima de término” tiene por objetivo ser un filtro para eliminar las palabras que son poco frecuentes y por lo tanto no aportan mucha información, por ejemplo, supongamos que un *tweet* contiene la palabra “*optometrista*” y el texto fue clasificado como positivo, pero el termino no ocurre otra vez, entonces la palabra “*optometrista*” no es de utilidad. La frecuencia mínima que nos dio buenos resultados fue 7, es decir que un término debe tener al menos 7 menciones en el corpus de entrenamiento para tener la posibilidad de ser considerada como característica.

El “Umbral de pertenencia de termino a polaridades” regula que un término contenga únicamente pesos potencialmente útiles, por ejemplo, si un término es de uso común es probable que sea frecuente en todas varias clases, como es el caso de las *stopwords*, por lo que no aportan información relevante para resolver la polaridad de un documento. Si el corpus de entrenamiento fuera balanceado, el peso de cada termino rondaría entre 16.66% para cada polaridad, pero debido a que no lo es, este margen se amplía, permitiendo en el sistema desarrollado hasta 2 pesos individualmente concentren entre 20% y 55% de peso, de esta forma obtenemos pesos contrastantes para los términos, anulando los pesos de las características no útiles.

El “Umbral de diferencia de pesos de polaridades de término” se usa una vez que obtuvimos polaridades contrastantes, y consiste en verificar que estas no se neutralicen, es decir, supongamos que la distribución de pesos de un término fuera la siguiente [0.40 0.05 0.00 0.00 0.45 0.10], podemos observar que la carga de polaridad está presente en polaridades opuestas, por lo que ambas se neutralizan, haciendo del termino no muy útil para determinar la polaridad. De esta forma, si establecemos un umbral de diferencia entre las polaridades opuestas, obtenemos los términos que están mejor polarizados.

Los paramentos anteriores nos ayudan a tener características potencialmente útiles. Por otro lado, los siguientes parámetros afectan al proceso de clasificación y se mencionan en el punto 4.4:

1. Constante de compensación de peso en la ausencia de sentimiento.
2. Umbral de neutralidad de un tweet.

El valor de la “Constante de compensación de peso en la ausencia de sentimiento” tiene la finalidad de balancear el peso entre las categorías de polaridad y la categoría sin polaridad o NONE. Esto debido a que si hacemos

un comparativo en el corpus de entrenamiento (detallado en la tabla y figura 6.1 del capítulo 6) entre el número de documentos con polaridad [P+, P, N, N+] y la sección sin polaridad o con polaridad neutra [NONE, NEU], encontramos que las primeras clases cuentan con el 70.18%, mientras que las segundas únicamente con el 29.82%, por lo que hacer un comparativo entre ambas clases sería desequilibrado. Para poder igualar las condiciones, tendríamos que multiplicar a las segundas por 2.3535, sin embargo, encontramos que multiplicando por 1.6 se obtenían mejores resultados. La razón es que aunque exista un balance en los *tweets*, el número de términos en cada *tweet* es variable, lo que también influye en el balance junto con otros factores.

El “umbral de neutralidad de un tweet” es usado para determinar que un documento es neutral, cuando existe una similitud de peso entre las polaridades, para ello y después de experimentar, llegamos a la conclusión de que si alguno de los pesos positivos [P+, P] o negativos [N, N+] se encuentra entre 0.51 y 0.50, automáticamente se obtiene una neutralidad.

Capítulo 6. PRUEBAS Y RESULTADOS

En el presente capítulo está destinado a describir las pruebas y resultados obtenidos que se realizaron con el sistema implementado. Para ello fue necesario contar con un corpus sobre el cual se realizar los experimentos, teniendo acceso al corpus usado en el TASS 2014.

Para la medición de resultados obtenidos se usó la exactitud (*Accuracy*) en la clasificación y adicionalmente se incluyen las medidas de precisión, exhaustividad y medida-F (*precision, recall, F1-measure*).

6.1 Corpus usado en la evaluación

Es considerado que la selección del corpus es una tarea muy importante, ya que una buena elección permite ahorrar la laboriosa tarea de crear uno nuevo; además, elegir un corpus que sea ampliamente conocido o utilizado por otros investigadores del área permite comparar resultados y aproximaciones abordadas.

Elegimos el corpus que proporciona TASS [10] porque es usado para promover la investigación y desarrollo de nuevos métodos que permitan el análisis de sentimientos en corpus especializados en español; también, porque proporcionan un corpus general de 68,017 publicaciones (hechas por periodistas, políticos y famosos). El corpus está dividido en dos secciones,

una de entrenamiento y otra de pruebas. El corpus de entrenamiento esta etiquetado y fue creado con el propósito de que los participantes lo analicen, diseñen una estrategia y realicen experimentos con la metodología propuesta. El corpus de pruebas no está etiquetado y es usado como la base para evaluar resultados obtenidos, permitiendo comparar la efectividad de los sistemas y determinar qué aproximación obtuvo los mejores resultados.

Adicionalmente, los organizadores del taller proporcionan un archivo con la extensión “.qrel” del corpus de pruebas. Dicho archivo, contiene la relación *idtweet-polaridad* para la comprobación de resultados.

La polaridad de los tweets se distribuye en 6 etiquetas: P+, P, NEU, N, N+ y NONE, las cuales se detalla en las tablas y figuras 6.1 y 6.2.

El corpus fue creado con un formato XML, a continuación se amplía la información sobre las características de etiquetado de la sección de entrenamiento:

- P+ (altamente positivo), P (positivo), NEU (neutral), N (negativo), N+ (altamente negativo) y *NONE* (sin sentimiento).
- *AGREEMENT* (concordancia) y *DISAGREEMENT* (discordancia), usadas para indicar la concordancia de las palabras usadas con el sentimiento expresado. Útil para detectar si la neutralidad de un comentario proviene del uso de palabras neutrales, o el uso de palabras positivas y negativas, también es útil para detectar en cierto nivel el sarcasmo.
- Etiquetado de polaridad a nivel de entidad mencionada en la publicación, usado para determinar el sentimiento que refleja la entidad en la publicación. El etiquetado es igual a lo mencionado en los 2 puntos anteriores.

- Tema (*Topic*), etiqueta usada para determinar el tema o temas al que pertenece el tweet, los temas pueden ser política, economía, literatura, entretenimiento, deportes, música, tecnología, cine, futbol y otros.

Según la documentación de TASS, todo el etiquetado de las publicaciones se realizó de forma semiautomática; donde primero se estableció una línea base de aprendizaje maquina (*machine learning*), fue ejecutada, y posteriormente, las etiquetas fueron revisadas por humanos expertos.

Tabla 6.1 Distribución de tweets del corpus TASS por polaridad en la sección de entrenamiento.

Polaridad	Número de tweets	Porcentaje
P+	1652	22.88 %
P	1232	17.07 %
NEU	670	9.28 %
N	1335	18.49 %
N+	847	11.73 %
NONE	1483	20.54 %
TOTAL	7219	100 %

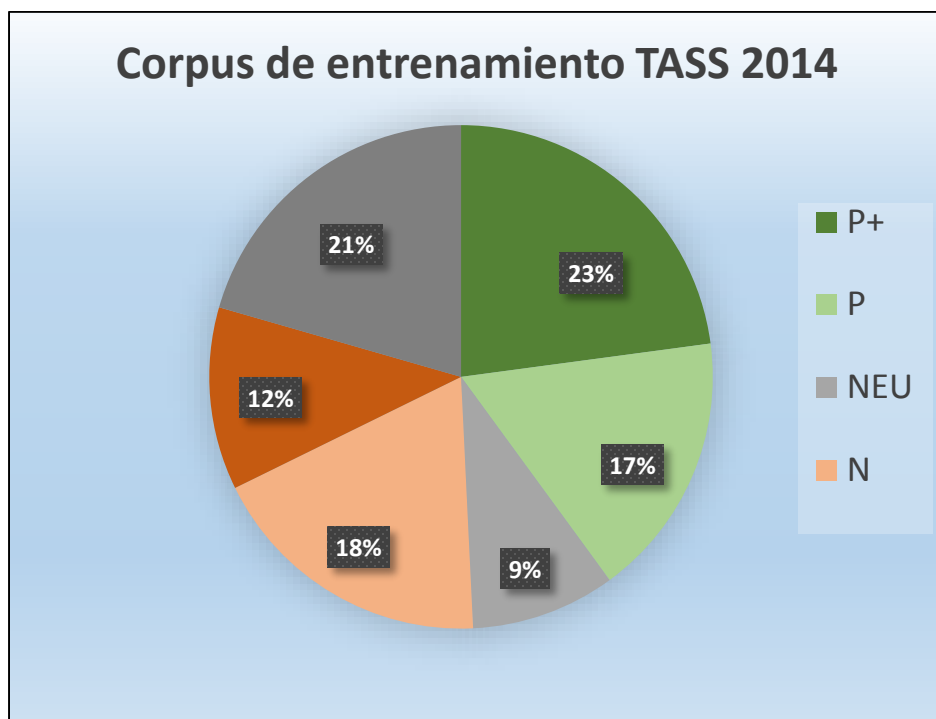


Figura 6.1 Grafica de distribución del corpus TASS en la sección de entrenamiento.

Tabla 6.2. Distribución de tweets del corpus por polaridad en la sección de pruebas.

Polaridad	Número de tweets	Porcentaje
P+	20,745	34.12 %
P	1,488	2.45 %
NEU	1,305	2.15 %
N	11,287	18.56 %
N+	4,557	7.50 %
NONE	21,416	35.22 %
TOTAL	60,798	100 %

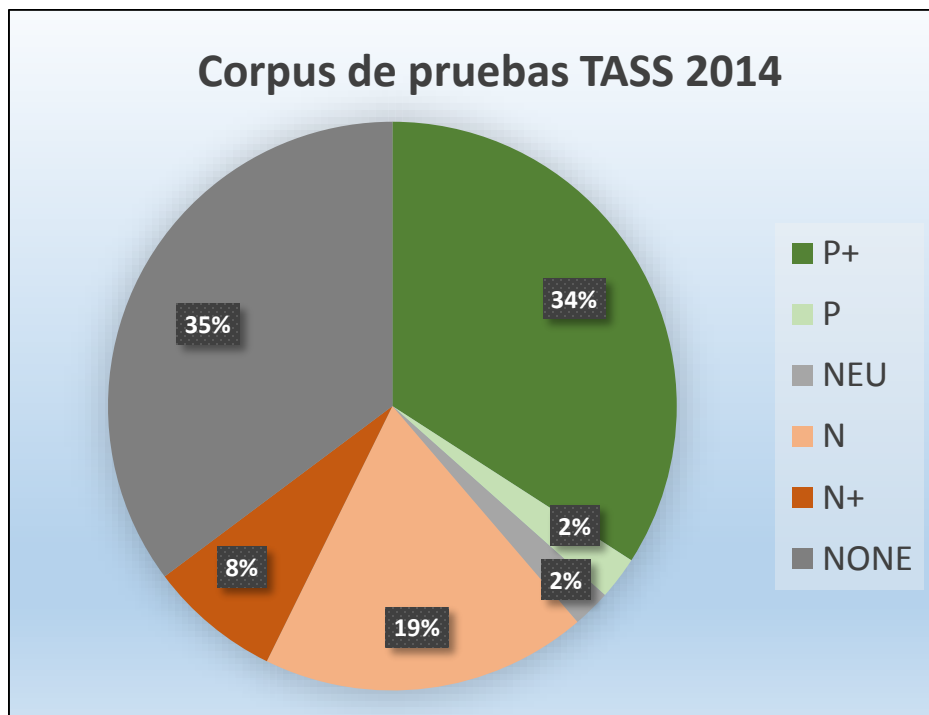


Figura 6.2 Grafica de distribución del corpus TASS en la sección de pruebas.

6.2 Mecanismo de evaluación

El procedimiento usado para verificar la efectividad fue el mismo que se usó en el Taller de Análisis de Sentimientos de la SEPLN 2014, es decir, usamos 7,219 textos etiquetados exclusivamente para entrenar el sistema y 60,798 textos (estándar de oro) para probar la precisión del sistema. Cada texto del corpus, tanto de la sección de entrenamiento como de prueba, está etiquetado con una polaridad global que puede ser medida mediante precisión, exhaustividad y medida F.

6.2.1 Exactitud (*Accuracy*)

La principal medida de evaluación que se usó para clasificar los sistemas de los participantes fue la exactitud o *Accuracy*, es decir, que porcentaje de los textos clasificados estuvieron correctos. Esta medida la obtenemos mediante la siguiente formula:

(6.1)

$$Accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + false\ negatives + true\ negatives}$$

La cual puede ser vista de la siguiente forma aplicada a nuestro contexto:

(6.2)

$$Accuracy = \frac{true\ positives}{size\ of\ corpus}$$

Donde *true positives* es el número de textos clasificados correctamente por nuestro sistema en alguna de las 6 categorías, y *size of corpus* es el número de textos que contiene el corpus de pruebas.

6.2.2 Precisión

La precisión es usada para medir la proporción de los textos clasificados correctamente por el sistema, dicho de otra forma, mide el porcentaje de documentos clasificados correctamente del conjunto total de documentos que el sistema evalúa. Para obtener la medición por clase o etiqueta la evaluación se realiza exclusivamente sobre el número total que el sistema asigno esa etiqueta de forma correcta o incorrecta.

La ecuación para obtener la precisión de los textos evaluados por un sistema que se utilizó en el TASS es la siguiente:

(6.3)

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Donde *true positives* es el número de textos clasificados correctamente en su categoría y *false positives* son los textos clasificados por el sistema de forma incorrecta.

6.2.3 Exhaustividad (*Recall*)

La exhaustividad es usada para determinar qué porcentaje de los textos seleccionados forman parte del conjunto objetivo. Es decir, el porcentaje de documentos que el sistema clasifica correctamente sobre el conjunto total de documentos que se deben clasificar. En el caso de la exhaustividad por polaridad, se comprueban los documentos clasificados correctamente sobre el conjunto de documentos total que pertenecen a polaridad evaluada.

La ecuación para obtener la medida de exhaustividad por los sistemas establecida en el TASS es la siguiente:

(6.4)

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Donde *true positives* es el número de textos clasificados correctamente en la categoría que les corresponde y *false negatives* son los textos clasificados con una etiqueta diferente a las establecidas.

6.2.4 Medida-F (*F1-measure*)

Finalmente, la medida-F combina la precisión y la exhaustividad, es decir, se obtiene una medida de desempeño general, en la que no se sacrifica la

precisión o la exhaustividad. En el TASS se usó la ecuación siguiente para obtener la medida-F de los sistemas:

(6.5)

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Donde *precision* y *recall* fueron obtenidos previamente y 2 es una constante.

6.3 Pruebas realizadas y resultados obtenidos

Se sometió el sistema a las pruebas de exactitud usando la sección de pruebas del corpus; entrenando el sistema con la sección de entrenamiento. Adicionalmente se calculó la medida F1 sobre la clasificación total y sobre cada categoría con el propósito de tener mayor información sobre la efectividad de clasificación en cada una de ellas.

6.3.1 Pruebas y resultados de exactitud del sistema

La principal evaluación la realizamos usando el 100% del corpus de la sección de entrenamiento para extraer las características y probamos el sistema sobre el 100% del corpus en su sección de pruebas. Cabe destacar, que mediante el archivo “*.qrel*” que proporciona TASS en su página se pudo verificar la exactitud del sistema.

Se realizaron 3 experimentos, el primero usando una clasificación aleatoria, el segundo usando palabras como características y el tercero usando lemas como características.

Para el experimento de clasificación aleatoria no fue necesario realizar un entrenamiento, únicamente se programó un clasificador aleatorio en Python usando la función *choice* que se encuentra en la biblioteca *random*⁸. Los resultados obtenidos fueron los siguientes:

Tabla 6.3 Pruebas aleatorias de clasificación.

	Distribución						Exactitud
	P+	P	NEU	N	N+	NONE	
Estándar de oro	34.1%	2.5%	2.2%	18.5%	7.5%	35.2%	100%
	20,745	1,491	1,305	11,284	4,557	21,416	60,798
Aleatorio	5.8%	0.4%	0.4%	3.1%	1.2%	5.9%	16.8%
	3,541	235	212	1892	738	3,591	10,209

Esta prueba se realizó como un primer acercamiento al corpus de pruebas, es decir, establecer una primera línea base a superar. De esta forma, era fácil saber que obtener un resultado más bajo que una clasificación aleatoria haría del sistema una nula aportación.

En el segundo experimento se usaron palabras o formas como características, para lo cual se incorporó el diccionario de frecuencia de palabras de la RAE para la corrección de texto y se siguió la metodología propuesta en el capítulo 4 para la selección de características y clasificación de textos.

⁸ Python Software Foundation. Random - Generate pseudo-random numbers. (2015), de <https://docs.python.org/2/library/random.html>

Tabla 6.4 Resultados obtenidos usando palabras como características.

	Distribución						Exactitud
	P+	P	NEU	N	N+	NONE	
Estándar de oro	34.1%	2.5%	2.2%	18.5%	7.5%	35.2%	100%
	20,745	1,491	1,305	11,284	4,557	21,416	60,798
Palabras como características	21.4%	0.8%	0.1%	5.9%	3.7%	22.8%	54.7%
	12,997	486	30	3,556	2,240	13,867	33,176

Se obtuvieron buenos resultados al utilizar palabras como características, incluso obtuvimos mejores resultados que algunos equipos que participaron en el TASS 2014 con esta aproximación.

El tercer experimento se realizó usando lemas de las palabras como características, de esta forma se redujeron el número de características, esto porque muchas palabras son formas flexionadas de otras, por lo que muchas palabras que antes pudieron ser consideradas como características diferentes ahora comparten solo una. Los resultados se muestran en la tabla 6.5.

Tabla 6.5 Resultados obtenidos usando lemas como características.

	Distribución						Exactitud
	P+	P	NEU	N	N+	NONE	
Estándar de oro	34.1%	2.5%	2.2%	18.5%	7.5%	35.2%	100%
	20,745	1,491	1,305	11,284	4,557	21,416	60,798
Lemas como características	22.5%	0.7%	0.1%	7.4%	3.6%	22.8%	56.8%
	13,493	438	30	4,489	2,173	13,880	34,503

Al usar lemas o la forma canónica de las palabras, notamos que mejoraron notablemente los resultados obtenidos, incluso superamos una propuesta más del TASS 2014 como veremos más adelante.

6.3.2 Pruebas y resultados de la medida F1 sobre las polaridades

Aunque la medida de exactitud fue la que se usó para determinar el ranking de los equipos que participaron en el taller de análisis de sentimientos de la SEPLN en la edición 2014, obtener las medidas *precisión*, *recall* y *F1* nos permite evaluar el desempeño del sistema.

Obtuvimos la medida de exactitud para los 3 experimentos realizados, en la tabla 6.6 se muestran los resultados de la evaluación del sistema a un nivel general, es decir, conjuntamos todas clases y evaluamos el desempeño en la clasificación de las 6 categorías. Mientras que en la tabla 6.7 evaluamos cada polaridad o clase de forma individual, lo cual nos ayudó a determinar las fortalezas y debilidades del sistema a nivel de polaridad en las diferentes experimentaciones.

Tabla 6.6 Precisión, exhaustividad y medida F del sistema aleatorio, por palabras y por lemas.

Sistema	Precisión	Exhaustividad	Medida F
Aleatorio	0.17	1.00	0.29
Palabras como características	0.55	1.00	0.71
Lemas como características	0.57	1.00	0.72

Tabla 6.7 Precisión, exhaustividad y medida F por clase para cada sistema.

Sistema Aleatorio			
Clase	Precisión	Exhaustividad	Medida F
P+	0.34	0.17	0.23
P	0.02	0.16	0.04
NEU	0.02	0.16	0.03
N	0.19	0.17	0.18
N+	0.07	0.16	0.10
NONE	0.17	0.17	0.23
Palabras como características			
Clase	Precisión	Exhaustividad	Medida F
P+	0.67	0.63	0.65
P	0.21	0.33	0.25
NEU	0.08	0.02	0.04
N	0.47	0.32	0.38
N+	0.46	0.49	0.48
NONE	0.53	0.65	0.58
Lemas como características			
Clase	Precisión	Exhaustividad	Medida F
P+	0.70	0.65	0.68
P	0.19	0.29	0.23
NEU	0.07	0.02	0.03
N	0.48	0.40	0.43
N+	0.43	0.48	0.45
NONE	0.57	0.65	0.60

6.4 Comparación de resultados con equipos de TASS 2014

Se estableció una primera línea base a superar de forma obligada; esta primera línea se basó en una clasificación aleatoria, los resultados obtenidos se muestran en la tabla 6.1. Posteriormente, se tomó como meta superar la mayor cantidad de resultados posibles obtenidos por los diferentes equipos que participaron en el TASS 2014.

Los resultados en el Taller de Análisis de Sentimientos de la SEPLN en su edición 2014 se muestran en la tabla 6.8, donde además anexamos los resultados que obtuvimos con nuestra aproximación. Cabe destacar que TASS únicamente proporciona los resultados de exactitud en sus publicaciones por lo de igual forma agregamos el mismo resultado en la tabla general.

Tabla 6.8 Resultados obtenidos en el taller TASS 2014 y resultados obtenidos con las propuestas realizadas.

Id de ejecución	Acc
ELiRF-UPV-run3	0.64
ELiRF-UPV-run1	0.63
ELiRF-UPV-run2	0.63
Elhuyar-Run1	0.61
Elhuyar-Run3	0.61
Elhuyar-Run2	0.61
LyS-1	0.58
MétodoPropuesto_2	0.57
LyS-2	0.56
MétodoPropuesto_1	0.54
SINAIword2vec-1	0.51
SINAI-ESMA-1	0.51
SINAI-ESMA-without_negation	0.51
JRC-run1-ER	0.48

JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams	0.48
JRC-run3-baseline-stop	0.48
IPN-Linguistic_2	0.37
IPN-1	0.37
SistemaAleatorio	0.17

Después de encontrar los parámetros óptimos y realizar experimentaciones usando palabras como características y lemas como características, encontramos que la primer propuesta que se basa en palabras como características rápidamente se posiciono sobre los equipo IPN [20], JRC [19], SINAI [22] y SINAI JCRword2vec [21], cuyas aproximaciones se abordan en el tercer capítulo.

Mientras que en la segunda propuesta al usar lemas como características mejoraron los resultados, superando la segunda aproximación del equipo LyS [18]. Por otra parte, las aproximaciones de los equipos ELiRF-UPV [16], Elhuyar [17] y la primera aproximación del equipo LyS [18], no se pudieron superar con la propuesta realizada, sin embargo, el abordar las tareas desde una perspectiva un tanto simplista e ingenua nos permitió obtener muy buenos resultados, dando pie a probar nuevas metodologías en el futuro, por ejemplo el uso de n-gramas, diccionario de sinónimos que ayude a ampliar la cobertura de características, tratar de obtener una estructura de los *tweets*, entre otras mejoras potenciales.

Capítulo 7. CONCLUSIONES

7.1 APORTACIONES

El trabajo desarrollado aporta una aproximación diferente en cuanto a la metodología para la selección de características y el establecimiento de pesos a cada término, ya que la mayoría de las aproximaciones considera a las palabras como unipolares, es decir, que solo aportan información hacia una dirección (positiva o negativa), en diferentes intensidades y generalmente es invertida cuando se topan con una negación. Sin embargo, nuestra propuesta considera a las palabras como multipolares o multiclase, por lo que una palabra positiva también puede aportar información negativa, excluyendo el uso de negaciones que pueden invertir la polaridad.

Por otra parte, la metodología para clasificar los textos se basa en un conjunto de reglas de clasificación que básicamente conforman un árbol de decisión, considerando que los términos que se involucran son multipolares.

7.2 PRODUCTOS DESARROLLADOS

Los principales productos desarrollados fueron:

- Sistema desarrollado en Python que:
 - Preprocesa textos con formato XML,
 - Extrae términos multipolares de un corpus etiquetado,

- Clasifica textos cortos.
- Un diccionario de lemas con pesos multipolares o multiclase que se extrajo automáticamente
- Un diccionario de palabras con pesos multipolares o multiclase que se extrajo automáticamente

7.3 TRABAJOS FUTUROS

La aproximación que se propuso obtuvo buenos resultados, sin embargo, aún se puede tener mejoras por lo que algunos trabajos futuros incluyen:

- Considerar los emoticonos, onomatopeyas e intensificadores.
- Experimentar con diferentes tamaños de n-gramas
- Realizar un sistema híbrido que considere lemas y palabras

Lo anterior con el propósito de crear un sistema más robusto que resuelva de mejor forma la tarea, sobre todo en clases con bajos resultados como la identificación de neutralidad.

BIBLIOGRAFÍA

1. McLuhan, M. (1994). Understanding media: The extensions of man. MIT press.
2. Nafría, I. (2007). Web 2.0: El usuario, el nuevo rey de Internet. Gestión 2000.
3. Bindé, J. (2005). Hacia las sociedades del conocimiento: informe mundial de la Unesco.
4. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F. Gordon, J. (2013). Empirical study of machine learning based approach for opinion mining in tweets. In Advances in Artificial Intelligence (pp. 1-14). Springer Berlin Heidelberg.
5. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.
6. Twitter, Inc. (2015), Developer Agreement & Policy. Recuperado de <https://dev.twitter.com/overview/terms/agreement-and-policy>
7. Read, J. (2005, June). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop (pp. 43-48). Association for Computational Linguistics.
8. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.
9. Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).

10. Román, J. V., Cámara, E. M., Morera, J. G., & Zafra, S. M. J. (2014). TASS 2014 –Taller de Análisis de Sentimientos en la SEPLN.
11. Corpus de Referencia del Español Actual (CREA), listado de frecuencias, De: <http://corpus.rae.es/lfrecuencias.html>
12. Joachims, T. Learning to classify text using support vector machines, 2002.
13. B Liu, L Zhang, A survey of opinion mining and sentiment analysis, 2012.
14. Sidorov, G. (2013). Construcción no lineal de ngramas en la lingüística computacional: n-gramas sintácticos, filtrados y generalizados. Sociedad Mexicana de Inteligencia Artificial. ISBN 978-607-95367-9-4.
15. Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528). ACM.
16. Hurtado, L. F., & Pla, F. (2014). ELiRF-UPV en TASS 2014: Análisis de Sentimientos, Detección de Tópicos y Análisis de Sentimientos de Aspectos en Twitter. Universidad Politécnica de Valencia. Valencia, España.
17. San Vicente Roncal, I., & Urizar, X. S. (2014). Looking for features for supervised tweet polarity classification. In Proceedings of the TASS workshop at SEPLN.
18. Vilares, D., Doval, Y., Alonso, M. A., & Gómez-Rodríguez, C. (2014). Lys at tass 2014: A prototype for extracting and analysing aspects from spanish tweets. In Proceedings of the TASS workshop at SEPLN.
19. Perea-Ortega, J. M., & Balahur, A. (2014). Experiments on feature replacements for polarity classification of spanish tweets. In Proceedings of the TASS workshop at SEPLN.
20. Hernández Petlachi, Roberto and Xiaoou Li. (2014). Análisis de sentimiento sobre textos en español basado en aproximaciones

semánticas con reglas lingüísticas. In Proceedings of the TASS workshop at SEPLN.

21. Montejo Ráez, Arturo, M. Ángel García Cumberras, and M. Carlos Díaz-Galiano. (2014). Participación de SINAI Word2Vec en TASS 2014. In Proceedings of the TASS workshop at SEPLN.
22. Jiménez Zafra, Salud M., Eugenio, Martínez Cámara, M. Teresa Martín Valdivia, and L. Alfonso Ureña López. (2014). SINAI-ESMA: An unsupervised approach for sentiment analysis in twitter. In Proceedings of the TASS workshop at SEPLN.
23. Castells, Manuel. (1999) "La revolución de la tecnología de la información" del libro La era de la información, Tomo 1, pp 55-87, México.
24. Hill, R. A., & Dunbar, R. I. (2003). Social network size in humans. Human nature, 14(1), 53-72.
25. Boroditsky, L. (2011). How language shapes thought. Scientific American, 304(2), 62-65.
26. A. Gelbukh, G. Sidorov. Procesamiento automático del español con enfoque en recursos léxicos grandes. Segunda edición, ampliada y revisada. IPN, México, ISBN 978-607-414-171-9, printing 1000, 2010, 92-98.
27. Padró, Lluís (2013). Form Dictionary File. De <http://nlp.lsi.upc.edu/freeling/doc/userman/html/node33.html>
28. Arce Castillo, Á. (1999). Intensificadores en español coloquial. Anuario de estudios filológicos.
29. Sánchez, A. J., Rodríguez, E. S. C., & Rodríguez, S. Arreglando y analizando textos. Avances en Informática y Automática, 151-160.
30. Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady (Vol. 10, No. 8, pp. 707-710).

31. Norvig, P. (2007). How to write a spelling corrector. De: <http://norvig.com/spell-correct.html>.
32. Python Software Foundation. (2015). Random — Generate pseudo-random numbers. De <https://docs.python.org/2/library/random.html>

Antes [7]

33. Twitter, Inc. (2015), Twitter Requisitos para la exhibición. Recuperado de <https://dev.twitter.com/terms/api-terms>

Glosario

Accuracy. O exactitud, es una medida que obtiene el porcentaje de documentos clasificados correctamente.

Aldea Global. Término introducido por sociólogo canadiense Marshall McLuhan, que describe a la sociedad como un gran sociedad mundial que supera las limitaciones de distancia para estar interconectados.

Análisis de sentimientos. Identificación y extracción de información subjetiva de un texto por medio del procesamiento del lenguaje natural.

Árbol de decisión. Es un modelo de predicción usado en la inteligencia artificial, que se construye en base a un diagrama lógico. Son similares a los sistemas de predicción basada en reglas.

Características (*features*). Es el nombre que se le da a los términos que se usan para describir algún objeto. En este caso son las palabras con diferentes pesos.

Corpus de entrenamiento (*train*). Es el nombre que se da al conjunto de textos que generalmente están etiquetados y sirven como fuente para entrenar a sistemas de cómputo por medio de aprendizaje automático o *machine learning*.

Corpus de pruebas (*test*). Es el nombre que se da al conjunto de textos que generalmente no están etiquetados y se usa para evaluar los resultados de algún sistema.

Corpus. Dentro del campo de la lingüística, es el conjunto de datos, textos u otros materiales sobre determinada materia que pueden servir de base para una investigación o trabajo.

Diccionario. Libro que contiene y define una o más palabras de algún idioma. Dentro de la programación en python, los diccionarios son una estructura de datos en la cual la indexación clave – valor se realiza por medio de claves alfanuméricas únicas.

Document frequency thresholding. O umbral de frecuencia en el documento, se pueden ver como lo opuesto a las *stopwords*, debido a su baja frecuencia no aportan información.

Emoticon. Neologismo formado por las palabras emoción e icono que usando caracteres ASCII representan de forma simbólica estados de ánimo o formas de mímicas de expresión.

Forma flexionada de una palabra. Son las palabras que se ven afectadas por prefijos, sufijos, género o número.

F-score. O medida F1, es una medida que sirve para medir el desempeño de un sistema que tiene que ver con la recuperación de documentos. Esta medida requiere que se calcule previamente *precision* y *recall*.

Hashtag. Hace referencia al símbolo de numeral, y generalmente se usa junto con una cadena de caracteres que forman una o más palabras para formar etiquetas.

IDF. Del inglés *Inverse document frequency*, frecuencia inversa de documento, es una medida que sirve para identificar que si un término es frecuente o no en una colección de documentos.

Inteligencia Artificial. Área multidisciplinaria que estudia la creación y diseños de sistemas capaces de resolver problemas cotidianos por sí mismos.

Lematización. Tarea del procesamiento del lenguaje natural en la que se identifica la forma canónica o lema de una palabra flexionada.

Lingüística computacional. Área donde converge la lingüística y la computación con especial interés por el estudio y la modelación del lenguaje humano mediante métodos computacionales.

Lista. Matriz unidimensional o vector, que puede contener una secuencia de datos. Las listas generalmente usan una indexación numérica que inicia en 0.

Naive Bayes. Clasificador bayesiano ingenuo, es una metodología que se apoya en la probabilidad fundamentada en el teorema de bayes. Las variables predictoras generalmente adquieren cierta independencia entre sí.

Preprocesamiento. Término usado para indicar la necesidad de realizar tareas previas antes de la tarea principal, en este caso depuración, corrección y normalización de texto.

Procesamiento del Lenguaje Natural. Tiene especial interés por estudiar y modelar el lenguaje mediante métodos computacionales, se le puede considerar sinónimo de la lingüística computacional.

Recall. O exhaustividad, es una medida que obtiene el porcentaje de efectividad en la recuperación de documentos sobre un conjunto objetivo.

Recursos lingüísticos. Conjunto de archivos como diccionarios, tesauros o lexicons que apoyan a la solución de algún problema, ya que contienen

información especializada. El *Spanish Emotion Lexicon* es un ejemplo de ello.

Red social. Término usado para nombrar a los círculos de amistades de una persona. En la actualidad este término se utiliza principalmente referirse a páginas como Facebook, Twitter, Pinterest, entre otras, que sirven para mantener un contacto virtual con otras personas.

Sentimiento. Término usado en la lingüística computacional para nombrar la expresividad subjetiva de un texto, por ejemplo: positivo, negativo, neutral, etcétera.

SEPLN. Sociedad Española para el Procesamiento del Lenguaje Natural, es una asociación científica sin ánimo de lucro con el objetivo de promover todo tipo de actividades relacionadas con el estudio del procesamiento de lenguaje natural.

Stopwords. O Palabras auxiliares o vacías, son aquellas palabras que debido a su alta frecuencia en documentos, independientemente de su temática o propósito, no aportan información.

TASS. Taller de Análisis de Sentimiento de la SEPLN, es un taller de evaluación experimental en el contexto de la Sociedad Española para el Procesamiento del Lenguaje Natural.

TF. Del inglés *Term frequency*, frecuencia de término, es una medida que sirve para determinar la relevancia bruta de un término en un documento; es decir, cuantas veces ocurre un término en documento.

TF-IDF. Del inglés *Term frequency – Inverse document frequency*, combina ambas medidas por medio de las cuales se puede medir que tan común es un término en una colección de documentos.

Tokenización. Tarea del procesamiento del lenguaje natural que consiste en dividir una oración en unidades. Cabe señalar que existen tokens compuestos como es el caso de “Buenos Aires”.

Trending topic. Se usa para describir las tendencias en las redes sociales virtuales, principalmente en Twitter, es decir, cuando un tema es comentado por muchos usuarios de tal forma que se vuelve popular.

Web 2.0. Término usado para describir la etapa en la que se popularizó la producción de contenidos por parte de los usuarios comunes.

Xml. Del inglés *eXtensible Markup Language*, es un lenguaje de marcas utilizado para guardar datos de forma legible, permite utilizar una definición gramática similar a la de html.