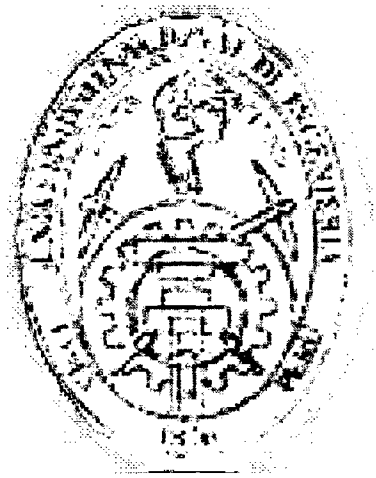
UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS



IDENTIFICACIÓN AUTOMÁTICA DE SIMILITUD FONÉTICA ENTRE MARCAS COMERCIALES USANDO REDES NEURONALES

TESIS

Para optar el Título Profesional de:

INGENIERO DE SISTEMAS

Saucedo Ascona, Edwin Manuel

Shuan Mendez, Alex Eduardo

Lima - Peru

2010

IDENTIFICACIÓN AUTOMÁTICA DE SIMILITUD FONÉTICA ENTRE MARCAS COMERCIALES USANDO REDES NEURONALES

Saucedo Ascona, Edwin Manuel
Shuan Mendez, Alex Eduardo

8 de julio de 2010

DEDICATORIA

De Saucedo Ascona Edwin Manuel:

Esta Tesis se la dedico a mi familia, especialmente a mi madre quien fue mi mayor apoyo y mi soporte. Gracias a ellos es que he podido avanzar en mi formación profesional, su apoyo y aliento que día a día me brindaron fue la motivación para realizar esta Tesis. Fueron ellos quienes me inspiraron, por tanto en estas líneas deseo expresarles mi profundo agradecimiento.

De Shuan Mendez Alex Eduardo:

Quiero dedicar esta tesis, que representa el último esfuerzo para lograr el título profesional de ingeniero de sistemas, a la persona más importante en mi vida: mi madre, cuyo apoyo incondicional ha hecho posible este logro, el cual no es solo mío, sino también suyo. Desde luego reconozco a mi padre y a mis hermanos, por el apoyo que me brindaron durante tantos años de estudio, por su cariño, su comprensión, pero sobre todo por haberme ayudado a asimilar todo lo aprendido.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

DESCRIPTORES TEMÁTICOS

1. Clase NIZA.
2. Codificación IPA.
3. Indecopi.
4. Marcas Comerciales.
5. N-Grams.
6. Nivel de Precisión.
7. Nivel de Reconocimiento.
8. Redes Neuronales.
9. Similitud Fonética.
10. Tokens .

RESUMEN

La investigación pretende establecer como objetivo validar un instrumento para medir el nivel de similitud fonética entre marcas comerciales. Se pretende utilizar las Redes Neuronales para este propósito. Se va a utilizar las marcas registradas en Indecopi para realizar el entrenamiento de la red neuronal y para realizar el Test se utilizará las marcas comerciales presentados en los casos tratados por Indecopi. Se medirá el nivel de reconocimiento y precisión de dicha herramienta, enfatizando el nivel de reconocimiento para así reducir los falsos negativos que vendrían a ser las marcas que son similares pero que no fueron identificadas, pero sin descuidar demasiado el nivel de precisión. Se presenta el modelo de solución en el cual se realiza una serie de tratamientos a las marcas comerciales (nombre en texto) empleando para la investigación dos diferentes categorías la primera la establecida por el algoritmo Phonix y la segunda elaborada con base en el Sistema de trascripción fonética del español IPA.

ABSTRACT

This Investigation pretends to establish as objective to validate a tool to measure the phonetic similarity level between trademarks. It's pretended to use neural networks to this purpose. Trademarks registered in Indecopi will be used to do the neural network training and trademarks presented in the cases treated by Indecopi will be used to do the Test. Precision level and recall level of the tool will be measured, emphasizing the recall level to reduce false negatives which are the trademark that are similar and were not retrieved, but without forget the precision level. A solution model in which a set of treatments are used on trademarks (name in text), using for two phonetic categories for the investigation, the first category is one established by the Phonix algorithm and the second one elaborated based in Spanish phonetic transcription system IPA.

INTRODUCCIÓN

Al hablar de una marca comercial hablamos de un signo distintivo que permite distinguir un producto o servicio de otros, a través de su registro una persona o empresa puede evitar el mal uso de esta por terceros, esto es evitar que estos terceros utilicen el prestigio ya adquirido de una marca para su propio beneficio. Esto se puede dar de dos formas: directa que es cuando el consumidor adquiere un producto confundiendo su nombre con el de una marca prestigiosa, o indirecta cuando el consumidor confunde el origen de un producto. En la actualidad es Indecopi a través de su oficina de signos distintivos quien se encarga de ver estos aspectos de propiedad intelectual de marcas comerciales. Sin embargo, el método que emplean para evaluar la registrabilidad de una marca (ver si hay alguna marca ya registrada con la que la marca a registrase tenga cierto grado de similitud fonética) tiene tan solo un 89 % de eficacia, esto es de las 100 marcas similares que debieron ser detectadas tan solo 89 lo han sido, mientras que las restantes son registradas generando posteriormente denuncias por casos de similitud. En promedio se tratan alrededor de 419 casos de denuncias por infracciones de similitud fonética entre marcas, el tratamiento de estos casos deviene en costos de más de S/131 000 nuevos soles anuales, costos que son asumidos por las empresas las cuales tienen que realizar la solicitud de oposición de estas marcas para defender sus dere-

chos. Lo que se busca con el planteamiento de esta Tesis es encontrar una herramienta que permita incrementar la eficacia hasta llegar a un 97 % en la identificación de similitud entre marcas comerciales con esto se podrá reducir en un 70 % el número de casos de marcas similares registradas así como los costos por su tratamiento.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

I.1. DESCRIPCIÓN DE LA SITUACIÓN PROBLEMÁTICA

Indecopi a través de su oficina de signos distintivos tiene que procesar 419 casos anualmente debido a problemas con imitación de marcas o al alto grado de similitud entre estas, se gasta S/ 131 012 anualmente en resolver casos de este tipo (Ver cuadro 1.2). El número de marcas que se registran mensualmente es alto, son en promedio 1236 solicitudes de registro de marcas mensualmente, de las cuales solo 943 son otorgadas (Ver cuadro 1.1). Además al haber marcas similares en el mercado se corre el riesgo de que se confunda al consumidor, ya que este podría confundir una marca ya registrada con la de un competidor y adquirir esta por equivocación.

En la figura 1.1 se muestra el numero de solicitudes de marcas en Indecopi durante los meses desde el año 2000 hasta el 2009. Siendo los años 2007, 2008 y 2009 proyecciones de los valores de los años anteriores.

Año	Solicitados	Otorgados	Rechazados
2000	16565	12724	3841
2001	14312	13258	1054
2002	14601	11452	3149
2003	14484	11424	3060
2004	15564	11274	4290
2005	14556	10606	3950
2006	14373	10133	4240
2007	14190	9659	4531
2008	14006	9186	4820
2009	13823	8713	5110

Cuadro I.1: Registro de Marcas

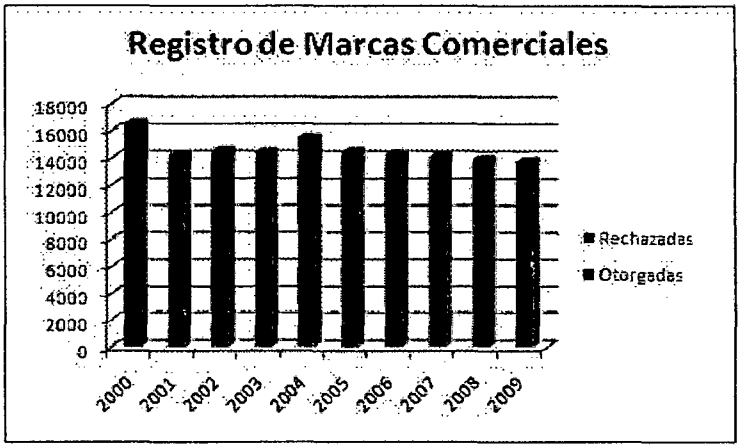


Figura I.1: Registro de Marcas

I.2. DESCRIPCIÓN DEL PROBLEMA

El problema es que el método usado para la evaluación de registrabilidad, esto es evaluar la similitud tanto fonética como grafica de una marca por registrar con una ya registrada, tiene un porcentaje de 89 % de eficacia (Ver cuadro 1.2). Por tanto muchas marcas similares a las ya registradas pasan este examen de registrabilidad y son registradas, estamos hablando de un promedio de 419 marcas similares registradas anualmente. Además este proceso toma gran tiempo, estamos hablando de 25 días hábiles de espera después de la

publicación de la marca en el diario El Peruano y alrededor de 7 días para la finalización del registro en caso la marca haya pasado el examen de registrabilidad y no haya habido un empresa que solicite oposición por similitud con su propia marca.

Año	Casos	% Eficacia	Costo
2000	303	92.69	S/. 85,234
2001	254	80.58	S/. 73,914
2002	323	90.70	S/. 97,126
2003	415	88.06	S/. 124,791
2004	453	90.45	S/. 140,611
2005	488	89.00	S/. 156,209
2006	534	88.81	S/. 176,113
2007	580	88.65	S/. 194,097
2008	626	89.16	S/. 209,557
2009	672	89.28	S/. 227,011

Cuadro I.2: Nivel de Eficacia y Costos

I.3. OBJETIVO DE LA INVESTIGACION

I.3.1. OBJETIVO SUPERIOR

El objetivo superior de esta propuesta de tesis es el de reducir el número de juicios por casos de marcas similares en más de 70 % y con esto reducir los costos que se dan debido a estos juicios también en un 70 %. Cabe especificar que el estudio a realizarse servirá de base para la implementación de una solución que permitirá la consecución del objetivo antes mencionado.

I.3.2. OBJETIVO PRINCIPAL

El objetivo principal que se va a plantear para esta Tesis será el de encontrar una herramienta que permita identificar las marcas que tengan similitud fonética con una marca ya registrada con un nivel de reconocimiento mayor

al 97%, es decir que identifique de 100 marcas con similitud fonética como mínimo 97 de estas.

$$Nivel\ de\ Reconocimiento = \frac{N\ marcas\ similares\ reconocidas}{N\ total\ marcas\ similares} \quad (I.1)$$

I.3.3. OBJETIVOS ESPECÍFICOS

- Identificar y analizar las variables que dan lugar a la similitud fonética entre marcas comerciales.
- Adaptar las herramientas vinculadas a similitud fonética entre palabras para la identificación de similitud fonética entre marcas comerciales.
- Desarrollar una herramienta basada en redes neuronales para la identificación de similitud fonética entre marcas comerciales.
- Evaluar el nivel de reconocimiento y precisión de las herramientas propuestas.

I.4. JUSTIFICACIÓN

El presente estudio brindará una herramienta que permita tratar la creciente demanda de registros distintivos en dos aspectos principales: eficacia (Precisión[1] y Reconocimiento[2]) y velocidad. Debido a la naturaleza acumulativa de las marcas comerciales, cada vez hay un mayor número de estas (ver figura 1.1) en la actualidad hay alrededor de 200 000 marcas registradas, por lo que se hace más difícil el evaluar la registrabilidad de las mismas. Cada vez quedan menos opciones para nombres de marcas por lo que se incrementa

[1]Indicador que permite medir el nivel de eficacia al momento de identificar las marcas, mientras más marcas similares sean identificadas mayor será el valor de este indicador

[2]Indicador que permite medir el nivel de eficiencia al momento de identificar las marcas, mientras menos marcas no similares sean identificadas mayor es el valor de este indicador

el riesgo de similitud con las ya registradas. Lo cual se manifiesta en el incremento de casos que se tienen que tratar anualmente (ver figura 1.2). De no implementarse una solución el número de casos a tratar continuará incrementándose y con esto los costos implicados con la resolución de los mismos.

Figura I.2: Registro de Signos distintivos Acumulados

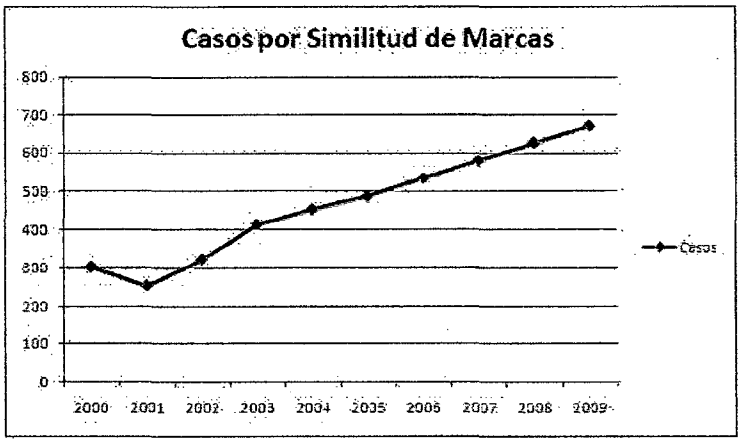


Figura I.3: Casos por infracciones resueltos

I.5. ALCANCES Y LIMITACIONES

I.5.1. ALCANCES DE LA INVESTIGACIÓN

La investigación estará orientada al registro de marcas comerciales que se realiza en el Perú teniendo en cuenta la legislación peruana. Por otro lado la

investigación a realizarse estará limitada al idioma español, debido a que se utilizará como base para el estudio fonético la trascripción fonética con el IPA[3] y debido a que este es el idioma que se utiliza en el Perú. Además para un estudio de similitud fonético se requiere realizar una adaptación al idioma de interés (Fall-GiraudCarrier, [1]).

I.5.2. LIMITACIONES DE LA INVESTIGACIÓN

En primer lugar la identificación de similitud no entrará en la parte gráfica, sino se buscará identificar la similitud fonética como se especifica en el titulo de la investigación. Por tanto no se verá los casos de similitud entre logos o imágenes que se utilizan para las marcas comerciales. La investigación no podrá ser válida para marcas comerciales que utilicen números o signos de puntuación en su escritura. La investigación tendrá mayor prioridad sobre el nivel de reconocimiento que el de precisión, ya que con esto se podrá conseguir identificar el mayor número de marcas similares permitiéndose falsos positivos en la identificación, los cuales vendrían a ser marcas no similares que serían identificadas como si lo fueran.

[3]Esquema de transcripción de fonemas al lenguaje castellano

CAPÍTULO II

REVISIÓN DE LA LITERATURA

II.1. DEFINICIONES BÁSICAS

II.1.1. DEFINICIÓN DE MARCA COMERCIAL

Una marca según la definición de la WIPO[1] (World Intellectual Property Organization) vendría a ser un signo o conjunto de signos que permiten diferenciar los productos o servicios de una empresa de los de las otras.

II.1.2. FUNCIONES DE LAS MARCAS COMERCIALES

Las cuatro funciones principales de las marcas comerciales son:

- Para distinguir los productos o servicios que ofrece una empresa de los que son ofrecidos por las demás, le permite a los consumidores identificar los productos conocidos por ellos o de los que vieron anuncios publicitarios.
- Para identificar la fuente u origen del producto o servicio, esto quiere decir para identificar a la empresa que brinda el producto o servicio, la

[1]Organización Mundial de la Propiedad Intenlectual, vela por el respeto de la propiedad intelectual de una forma más global www.wipo.int

marca comercial va a permitir la diferenciación de fuentes para productos similares.

- Para referirse a la calidad específica de un producto o servicio usado, de tal forma que el cliente pueda confiar en la calidad de un producto porque este está apoyado por la marca bajo la cual este se está produciendo.
- Para promover el marketing y la venta de los productos y que los servicios puedan ser brindados, la marca comercial muchas veces estimula la venta de los productos, creando así interés e inspirando en los clientes un sentimiento de confidencialidad.

II.1.3. REGLAS DE SIMILITUD ENTRE MARCAS COMERCIALES

Existen 18 reglas principales que se utilizan como criterio para establecer la similitud verbal entre marcas comerciales (Bugdahl, [2]). Entre ellas tenemos las reglas para las marcas que:

Son idénticas. Terminan o empiezan de forma similar. Tienen vocales idénticas. Una contiene a la otra. Una tiene una letra o más insertadas. Una tiene una o más letras eliminadas. Tienen las silabas permutadas. Suenan de forma similar. Tienen las mismas consonantes. Tienen consonantes que suenan de forma similar. Tienen las mismas vocales y las primeras consonantes. Tienen las letras permutadas.

II.1.4. FONEMAS

Se utilizarán los fonemas[2] empleados en el habla española, de esta forma se logrará adecuar la investigación a nuestro contexto los cuales están especificados en el cuadro 2.1.

[2]Abstracción mental o abstracciones formales de los sonidos del habla

Fonema
1. /a/: Fonema vocálico de apertura máxima.
2. /B/: Fonema obstruyente bilabial sonoro (grafías: b, v y w, alófonos: [b], [ß]).
3. /c/: Fonema africado palatal (grafía ch).
4. /D/: Fonema obstruyente coronal-alveolar sonoro.
5. /e/: Fonema vocálico palatal de apertura media.
6. /f/: Fonema fricativo labio-dental, en muchas zonas se realiza fricativo bilabial.
7. /G/: Fonema obstruyente velar sonoro.
8. /i/: Fonema vocálico palatal y apertura mínima.
9. /x/: Fonema fricativo velar.
10. /k/: Fonema oclusivo velar sordo (grafías c y qu).
11. /l/: Fonema lateral (coronal-alveolar).
12. /m/: Fonema nasal labial.
13. /n/: Fonema nasal (coronal-alveolar).
14. /ñ/: Fonema nasal palatal.
15. /o/: Fonema vocálico velar de apertura media.
16. /p/: Fonema oclusivo (bi)labial sordo
17. /r/: Fonema vibrante simple (grafía -r-, -r).
18. /rr/(rr): Fonema vibrante múltiple (grafía -rr-, r-).
19. /s/: Fonema fricativo (coronal-)alveolar (grafía s, en algunas variedades z y c).
20. /t/: Fonema oclusivo (coronal-)alveolar sordo.
21. /u/: Fonema vocálico velar de apertura mínima.
22. /y/: Fonema sonorante palatal (grafía y, en las zonas yeístas también corresponde a ll.

Cuadro II.1: Fonemas del habla española

II.1.5. CATEGORIAS FONÉTICAS

Las categorías asignadas a los fonemas estarán basadas en la similitud fonética de los fonemas, viendo aspectos de lugar de articulación y modo de articulación (Ver cuadros 2.2 y 2.3). Así por ejemplo los fonemas /p/ y /b/ se encontrarían dentro de una misma categoría debido a que ambos son articulados en la zona labial con un modo de articulación oclusivo, lo cual los hace similares fonéticamente.

Vocal	**Anteriores**	**Central**	**posteriores**
Altas (Cerradas)	i		u
Medias	e		o
Baja (Abierta)		a	

Cuadro II.2: Fonemas Vocálicos

Lugar de articulación	**Labial**		**Coronal**			**Dorsal**
Modo de articulación	**Bilabial**	**Labio-Den.**	**Dent.**	**Albeolar**	**Palatal**	**Velar**
Nasal	m	n				
Oclusiva	p b	t d		k g		
Fricativa		f (v)		s z	y	
Aproximante				j		
Vibrante múltiple				rr		
Vibrante simple				r		
Aproximante lateral				l		

Cuadro II.3: Fonemas de Consonantes

II.2. ALGORITMOS DE IDENTIFICACIÓN DE SIMILITUD VERBAL

Para la identificación de la similitud verbal entre marcas comerciales se utilizan como herramientas a los algoritmos, los algoritmos que se han utilizado

para la identificación de la similitud entre marcas comerciales son principalmente de dos tipos, los que se basan en las letras dentro de palabra y las que se basan en el aspecto fonético (Fall-GiraudCarrier, [1]).

II.2.1. ALGORITMOS BASADOS EN EL ORDEN DE LETRAS EN UNA CADENA

Edit Distance Es un algoritmo que establece como elemento de medida para la similitud entre palabras a la distancia que existe entre estas, está distancia está definida por número de inserciones, cambios o eliminaciones de letras necesario para que la palabra evaluada sea igual a la palabra blanco (Fischer - Zell , [7]), asignando una distancia de uno por cada inserción, eliminación o cambio de palabra. Por ejemplo para marcas comerciales veamos que la distancia entre la marca Kolinos y la marca Kornos vendría a ser de 2 ya que se eliminaría la letra i y se cambiaría la letra l por la r, el ejemplo se encuentra ilustrado en la figura 2.1.

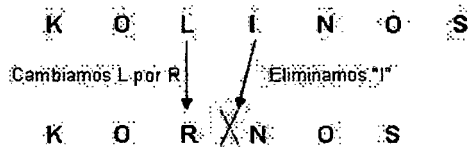


Figura II.1: Ejemplo Edit Distance

Esta basado en la siguiente relación de recurrencia:

$$
\begin{aligned}
&edit(0,0) = 0\\
&edit(i,0) = i\\
&edit(0,j) = j\\
&edit(i,j) = min[edit(i-1,j)+1,\\
&\qquad edit(i,j-1)+1,\\
&\qquad edit(i-1,j-1)+r(s_i,t_j)]
\end{aligned}
$$

Figura II.2: Relacion de recurrencia

Modified Edit Es similar al algoritmo Edit Distance pero la diferencia esta en que la permutación de dos letras contiguas vendría a tener una distancia de

Autores	Año	Método	Título
Fischer I, Zell A [7]	2000	Edit Distance	String averages and self-organizing maps for strings
Wu S, Manber U [8]	1992	Modified Distance	Fast text searching allowing errors
Zobel J, Dart P. [3]	1995	N-Grams	Finding approximate matches in large lexicons
E. Ukkonen [4]	1992	N-Grams	Approximate string-matching with q-grams and maximal matches
Holmes D, McCabe MC [6]	2002	Soundex	Improving precision and recall for soundex retrieval.
Zobel J, Dart P. [3]	1995	Phonix	Finding approximate matches in large lexicons
Zobel J, Dart P [9]	1996	Editex	Phonetic string matching: lessons from information retrieval
Kukich, Karen [15]	1992	Redes Neuronales	Techniques for Automatically Correcting Words

Cuadro II.4: Marco Conceptual Instrumental

1, mientras que para el algoritmo Edit Distance esta vendría a considerarse como 2 intercambios de letras por tanto una distancia de 2 (Wu - Manber, [8]). Por ejemplo la distancia entre las marcas Palmolive y Pamlolive sería de 2 para Edit Distance y 1 para Modified Edit (Ver figura 2.3).

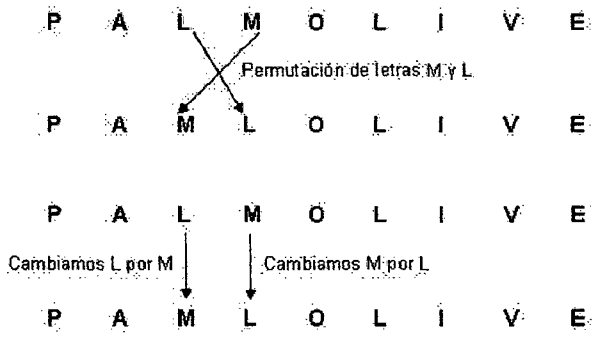


Figura II.3: Ejemplo Modified Edit

N-grams Donde el n-gram de una cadena es cualquier subcadena de esta cuya longitud sea n, y el conjunto de estos n-grams de la cadena s es definido

por Gs (Zobel - Dart, [3]).Por ejemplo Gs de la cadena KIMBO para n=2 sería KI, IM, MB, BO . La similitud entre dos cadenas puede ser medida a través de la siguiente expresión:

$$gram\text{-}count(s,t) = |G_s \cap G_t|$$

Figura II.4: n-gram formula 1

Por ejemplo para las cadenas KIMBO y KUMBOR, donde GKIMBO = KI, IM, MB, BO y GKUMBOR = KU, UM, MB, BO, OR la similitud sería de 2, ya que ambos contienen a los n-grams MB y BO.

$$gram\text{-}dist(s,t) = \sum_{g \in G_s \cup G_t} |s[g] - t[g]|$$

Figura II.5: n-gram formula 2

Pero cuando una cadena contiene a otra como KIMBOMAX y KIMBO, la similitud sería la misma que la que tiene la cadena KIMBO consigo misma. Para esto se desarrolló una nueva medida de distancia (Ukkonen, [4]):

donde s[g] vendría a ser el número de veces que aparece el n-gram g en la cadena s. Así por ejemplo la distancia entre las cadenas KIMBO y KIMBOMAX sería de 3 ya que los n-grams KI, IM, MB, BO aparecen una vez en cada cadena por tanto su suma es cero mientras que los n-grams OM,MA,AX solo aparecen en la cadena KIMBOMAX por lo tanto la suma sería de 3, y al hallar la suma total esta sería exactamente 3.

II.2.2. ALGORITMOS BASADOS EN ASPECTOS FONÉTICOS

Soundex Este algoritmo está basado en el sonido de cada una de las letras para convertir una cadena en una forma canónica (Holmes - McCabe, [6]). Los códigos que utiliza son los mostrados en la tabla 2.5:

Code	Caracteres
0	a e i g o u w y
1	b f p v
2	c g j k q s x z
3	d t
4	l
5	m n
6	r

Cuadro II.5: Codificación Soundex

El algoritmo lo que hace es conservar la primera letra de la cadena y remplazar las siguientes por su correspondiente código, para luego eliminar los códigos que se repiten de forma consecutiva y todas las ocurrencias del código cero. Finalmente toma los cuatro primeros caracteres de la cadena resultante completando con ceros si es necesario. Así por ejemplo la cadena PALMOLIVE estaría representada por P454 y la cadena KIMBO por K510.

Phonix Es similar al algoritmo Soundex, pero como se muestra en el cuadro 2.6 este agrupa a las letras en 8 grupos fonéticos en vez de 6 (Zobel - Dart, [3]), además cuenta con 160 transformaciones para grupos de letras que se utilizan en el Inglés por lo tanto su uso se restringe a este idioma.

Code	Caracteres
0	a e i h o u w y
1	b p
2	c g j k q
3	d t
4	l
5	m n
6	r
7	f v
8	s x z

Cuadro II.6: Codificación Phonix

Editex Este algoritmo combina la medida de distancia del algoritmo Edit Distance con la estrategia de agrupamiento de Soundex y Phonix (Zobel - Dart, [9]). Se utiliza una relación de recurrencia similar a la de edit distance solo que ahora se define una nueva función d(a,b), y los valores para r(a,b) son: 0 si a = b, 1 si a y b pertenecen al mismo grupo y 2 si ocurre de otra forma. Los valores para d(a,b) se dan de igual forma solo que para a igual a h o w y dado que a es diferente de b su valor será de 1.

$$
\begin{aligned}
&edit(0,0) = 0 \\
&edit(i,0) = edit(i-1,0) + d(s_{i-1}, s_i) \\
&edit(0,j) = edit(0,j-1) + d(t_{j-1}, t_j) \\
&edit(i,j) = min[edit(i-1,j) + d(s_{i-1}, s_i), \\
&\qquad edit(i,j-1) + d(t_{j-1}, t_j), \\
&\qquad edit(i-1,j-1) + r(s_i, t_j)]
\end{aligned}
$$

Figura II.6: formula editex

Los grupos fonéticos para este algoritmo son presentados en el cuadro 2.7.

Code	Caracteres
0	a e i o u y
1	b p
2	c k q
3	d t
4	l r
5	m n
6	g j
7	f p v
8	s x z
9	c s z

Cuadro II.7: Codificación Soundex

Redes Neuronales Una de las funciones principales de las redes neuronales es el reconocimiento de patrones (Konohen, [14]) donde el criterio principal de desempeño es la minimización del número de errores de clasificación. Uno de sus principales fundamentos teóricos es la teoría de promedio condicional de pérdida en la toma de decisiones el cual se basa en la teoría de probabilidad

de Bayes. Una red neuronal suele trabajar como una caja negra, la cual recibe una serie de signos de observación constituyendo así el vector de entrada x, y produce una respuesta rh en alguno de sus puertos de salida i, cada puerto es asignado a una diferente clase de objetos observados. Se han hecho aplicaciones de redes neuronales para la determinación de errores de escritura a través del uso de N-Grams[3] (Kukich, [15]), donde los N-Grams pueden ser utilizados como neuronas de entrada y los nodos de salida vendrían a ser todas las palabras en la base de datos y los valores de las capas entre un nodo de entrada y un nodo de palabra de salida serían los índices por cada elemento de la matriz. Estos enlaces podrían ser pesados de forma individual o normalizados a 1. En los test realizados (Kukich, [16])utilizando este método se logró conseguir niveles de eficacia de 75.88 % con 521 palabras del diccionario y 75.29 % con 1 142 palabras del diccionario.

II.3. ESTUDIOS REALIZADOS ANTERIORMENTE

II.3.1. DESARROLLO DE UNA COMPARACIÓN COMPUTARIZADA DE SIMILITUD FONOLÓGICA ENTRE MARCAS COMERCIALES EN SUECIA

Los estudios realizados fueron liderados por Benny Broda durante los años 1960 a 1970 en la oficina de patentes de Suecia, donde se realizaron estudios para determinar distancias fonológicas entre marcas comerciales, se estableció en esta oficina que el nuevo registro de una marca comercial dependía parcialmente de que esta no fuese similar fonológicamente a una ya existente. Posteriormente se llegó a establecer y refinar un procedimiento para determinar estas distancias fonológicas (Brodda, [10]). Además se implementó un test de comprensión a través del uso de una métrica de distancia fonológica

[3]Subcadena de una palabra cuya longitud es menor a la palabra que la origina, es utilizada para la identificación de similitud entre nombres comunes

Autores	Año	Método	Título
Brodda B. [10]	1990	Computational Linguistics	Corpus work with PC beta: a presentation
McAllister R, Brodda B [11]	2002	Brodda algorithm	Development of a new speech comprehension test with a phonological distance metric
Erikson K [12]	1997	Algorithms	Approximate swedish name matching survey and test of different algorithms
Hodge VJ, Austin J. [13]	2001	Phonetex	An evaluation of phonetic spell checkers
C J Fall C. Giraud-Carrier[1]	2005	Algoritmos hibridos	Searching trademark databases for verbal similarities

Cuadro II.8: Antecedentes de Investigación

(McAllister-Brodda, [11])

II.3.2. SISTEMA DE IDENTIFICACIÓN DE SIMILITUD FONÉTICA ENTRE NOMBRES EN DIRECTORIOS TELEFÓNICOS

En esta investigación se busca identificar los nombres que son fonéticamente iguales dentro de un directorio telefónico (Erikson, [12]) a través del uso de algunos de los algoritmos mostrados anteriormente y además de otros y algunas combinaciones de estos. En esta investigación se tienen 3 criterios de evaluación los cuales son: Precisión, Proporción de nombres identificados que son realmente similares. Reconocimiento, Proporción de nombres similares que son realmente identificados. Velocidad, relacionado al tiempo utilizado tanto por el usuario como por el sistema.

II.3.3. CORRECTORES FONÉTICOS DE ESCRITURA.

En esta investigación se desarrolla un corrector fonético de escritura, llamado Phonetex el cual adopta la metodología Phonix y su aproximación a la

combinación de tipos de códigos de Soundex con reglas de transformación fonética para producir un código fonético para cada palabra. Para esto se estudió la etimología del inglés detallada en el Diccionario de Inglés de Oxford. Phonetex llega a tener más grupos fonéticos que los que tienen Soundex, Phonix o Editex, en total 14 grupos:

II.3.4. BUSCANDO EN LAS BASES DE DATOS DE MARCAS COMERCIALES POR SIMILITUD VERBAL

En esta investigación lo que se busca es establecer la variabilidad de resultados que se puede obtener de usar los diferentes algoritmos, especificando que cada uno de estos permite identificar determinadas similitudes, por tanto plantea que se utilice algoritmos híbridos, es decir algoritmos que tomen las características y fortalezas de un algoritmo y las combinen con las de otro algoritmo para así mejorar su nivel de reconocimiento. Para el caso de las marcas comerciales tienen mayor importancia el nivel de reconocimiento (Fall - GiraudCarrier, [1]), esto debido a que es más importante identificar la mayor cantidad de marcas similares aunque con esto se consiga un mayor número de falsos positivos.

CAPÍTULO III

DESCRIPCIÓN DE DATOS

III.1. FUENTE DE DATOS

Los datos de las marcas se encuentran disponibles el portal de Indecopi[1], de donde se puede obtener las jurisprudencias de los casos de marcas que fueron identificadas como similares antes de ser registradas o cuando los dueños de las marcas han detectado la similitud por su propia cuenta.

III.2. ESTRUCTURA DE LOS DATOS

Cada marca comercial tiene la siguiente estructura:

idMarcaComercial.- Que viene a ser el identificador de la marca comercial. Es una variable del tipo número y sirve para identificar y diferenciar a una marca de otra. Nombre.- Viene a ser el nombre en texto de la marca. Es una variable tipo Cadena y puede

contener caracteres diferentes de las letras.

Descripción.- Que viene a ser una breve descripción de la marca comercial que contiene la naturaleza y características del producto, para efectos de

[1]Instituto Naciona de Defensa de la Competencia y de la Protección de la Propiedad Intelectual www.indecopi.gob.pe

estudio este dato no es tan importante.

idClaseNiza.- que viene a ser la clasificación internacional a la que pertenece el producto cuya marca pretende ser registrada. Es una variable tipo número y su rango es de 1 a 45.

III.3. DESCRIPCIÓN DE LOS DATOS

Se utilizaron 1040 marcas obtenidas de los casos de jurisprudencia en Indecopi.

El cuadro 3.1 muestra el número de marcas por cada clase Niza, para efectos didácticos solo se está utilizando las clases que tienen más de 30 marcas. Lo mismo podrá ser visto de forma gráfica en la figura 3.1.

Clase	Cant. Marcas	Porcentaje
1	30	2.9%
3	38	3.7%
5	196	18.8%
9	62	6%
16	30	2.9%
25	196	18.8%
29	30	2.9%
30	63	6.1%
41	48	4.6%
43	30	2.9%

Cuadro III.1: Marcas por Clase

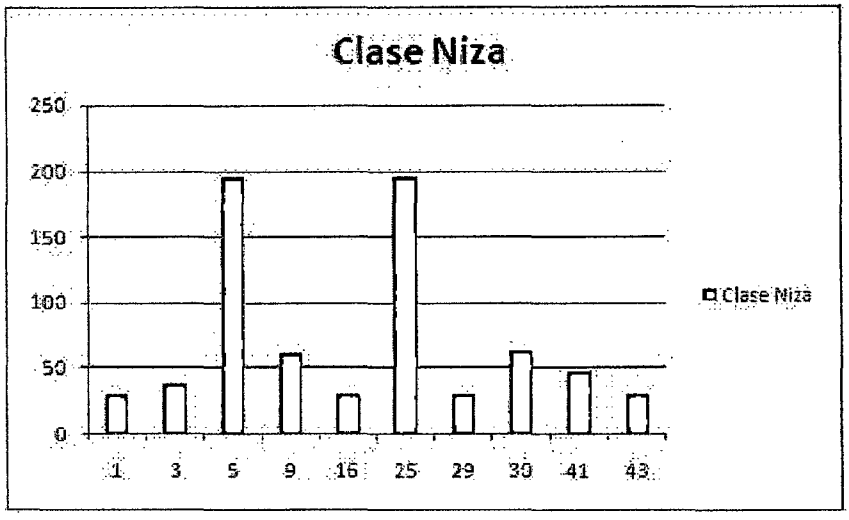


Figura III.1: Marcas por Clase Niza

III.3.1. ESTADÍSTICA UNIVARIADA

Luego de un análisis de las 1040 marcas se obtuvo que 670 de ellas contenían solo a una palabra y 250 a dos (Ver figura 3.2 y cuadro 3.2), es decir el 88 % de las marcas cotienen menos de 3 palabras

Numero de Palabras	Instancias
1	670
2	250
3	74
4	37
5	4
6	3
8	2

Cuadro III.2: Número de palabras por Marca

Luego del análisis de todas las palabras obtenidas de las marcas se obtuvo que el 78 % de las palabras tiene entre 3 y 8 letras (Ver figura 3.3 y cuadro 3.3)

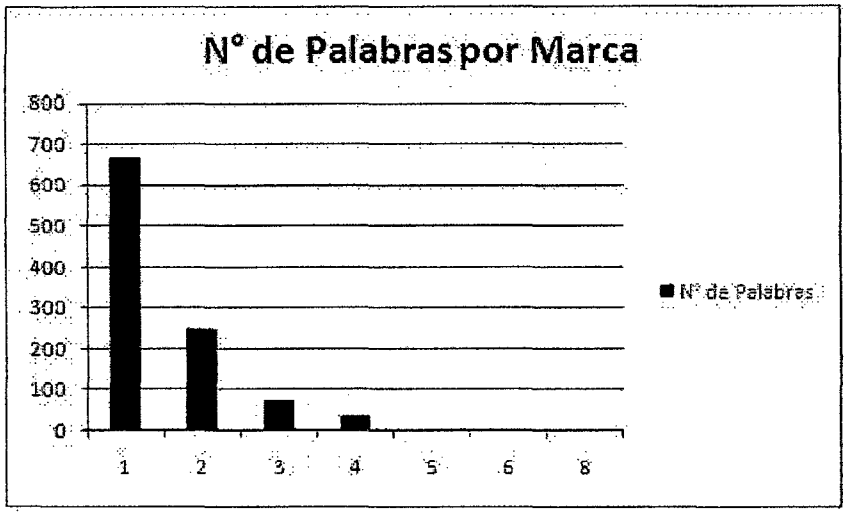


Figura III.2: Palabras por Marca

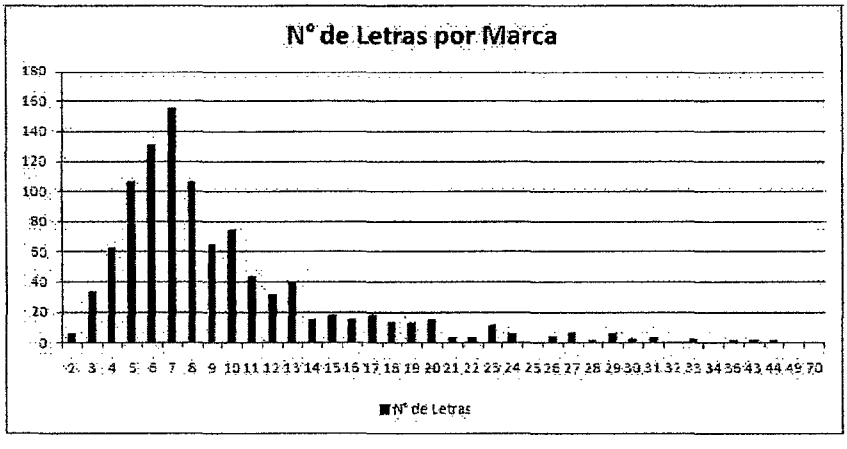


Figura III.3: Letras por Palabra

Número de letras	Instancias
2	6
3	34
4	63
5	107
6	131
7	156
8	107
9	65
10	75
11	44
12	32
13	40
14	16
15	19
16	16
17	19
18	14
19	13
20	16
21	4
22	4
23	12
24	6
25	1
26	5
27	7
28	2
29	6
30	3
31	4
32	1
33	3
34	1
36	2
43	2
44	2
49	1
70	1

Cuadro III.3: Número de letras por palabra

CAPÍTULO IV

MODELO DE SOLUCIÓN

IV.1. MODELO DE SOLUCIÓN

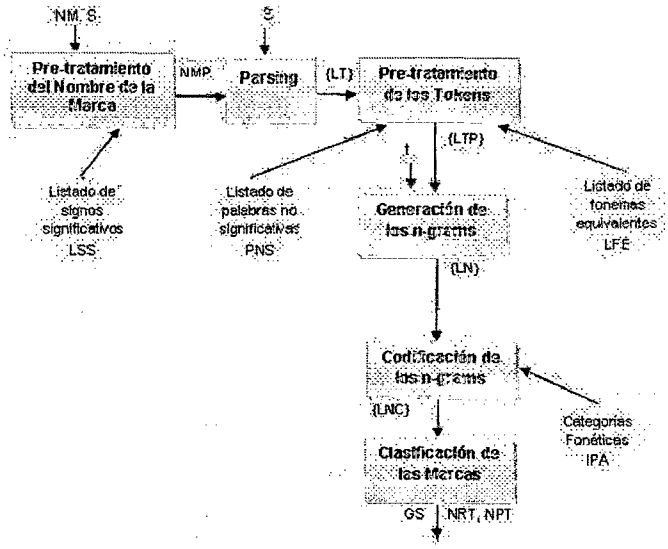


Figura IV.1: Modelo de Solución

El procedimiento general tiene como finalidad principal la obtención del modelo óptimo que permita obtener los valores más altos de reconocimiento y precisión, la entrada principal es el nombre en texto de la marca comercial (NM). Posteriormente el nombre en texto de la marca es procesado en 6 eta-

pas para obtener como resultado el grado de similitud (GS) entre las marcas el cual indica si son fonéticamente similares o no, así como los niveles de reconocimiento (NRT) y precisión (NPT). El procedimiento tiene una lista de parámetros que serán analizados con más detalle en cada etapa del procedimiento.

El procedimiento general está compuesto por 6 etapas: pre-tratamiento del nombre de la marca, parsing, pre-tratamiento de los tokens, generación de los n-grams, codificación de los n-grams y clasificación de los n-grams.

1. Pre-tratamiento del nombre de la marca Esta etapa tiene por finalidad eliminar a aquellos signos que carezcan de significado y puedan distorsionar la identificación de marcas similares a lo largo de sus etapas. Este procedimiento tiene como entrada al nombre en texto de la marca comercial NM. Primero se identifica si cada uno de los caracteres es una letra o un signo, de darse el segundo caso se identifica si el signo tiene significado comparándolo con la lista de signos significativos (Ver anexo A), de tener significado se remplaza el signo por su significado, en caso contrario se elimina dicho signo. Se tiene que tener cuidado de no eliminar el caracter que es utilizado como caracter separador (S), lo que se recomienda para estos casos es ingresar el caracter separador (S) como parámetro y validarlo a fin de no eliminarlo. El resultado es el nombre de la marca pre-tratado NMP, el cual no contendrá signos que puedan causar distorsión a excepción del caracter separador S, el cual será utilizado durante la siguiente etapa.

2. Parsing Esta etapa tiene por finalidad dividir al nombre de la marca en tokens (palabras) los cuales estarán separados por el caracter separador (S). Este procedimiento tiene como entradas al nombre pre-tratado de la marca y al caracter separador (S). Primero se realiza un recorrido a la palabra, caracter

por caracter, hasta que se encuentra al caracter S, entonces se crea un token a partir de los caracteres recorridos antes de encontrar el caracter S. Se procede de la misma forma a partir de la posición siguiente a la ubicación del caracter S. El resultado es la lista de tokens LT de la marca comercial.

3. Pre-tratamiento de los tokens. Esta etapa tiene por finalidad eliminar a aquellos tokens que no posean significado o se encuentren directamente asociados a las definiciones de la clase Niza a la que pertenecen y realizar una simplificación fonética a aquellos tokens que no fuesen eliminados. Este procedimiento tiene como entrada a la lista de tokens LT obtenidos de la etapa anterior. Primero se compara el token con la lista de palabras no significativas (Ver anexo B) la cual contiene a aquellas palabras que no posean significado propio sino sean empleadas para enlazar a otras palabras. Se elimina a aquellos tokens que se encuentren dentro de la lista de palabras no significativas. Posteriormente se procede a identificar a los fonemas o par de fonemas que requieran simplificación, para esto se compara los fonemas de dos en dos con los fonemas de la lista de fonemas equivalentes (Ver anexo C). Se reemplaza a los pares de fonemas que se encuentren dentro de la lista por sus respectivos fonemas equivalentes simplificados. El resultado es la lista de tokens pre-tratados (LTP).

4. Generación de los n-grams. Esta etapa tiene por finalidad generar los n-grams a partir de los tokens. Este procedimiento tiene como entradas a la lista de tokens pre-tratados (LTP) y al tamaño de los n-grams (t1) donde t1=2,3,...t, donde t es el tamaño máximo de los n-grams a emplearse . Primero se verifica si el token tiene una longitud (L) mayor a t1, de ser el caso se procede a recorrer el token hasta la posición (L-t1) y creando a su vez las sub-cadenas de tamaño t1 que vienen a ser los n-grams, de tal forma que se obtendrán

(L-t1+1) n-grams. Para el caso de n-grams que tengan longitud menor a t1 se añadirá (t1-L) caracteres 'a', obteniendo así un n-gram. El procedimiento se realizará para cada token de la LTP. El resultado es la lista de n-grams (LN).

5. Codificación de los n-grams. Esta etapa tiene como finalidad la codificación de los fonemas de los n-grams. La entrada de este proceso es la lista de n-grams (LN). Primero se procede a recorrer todo el n-gram, caracter por caracter, y se remplaza cada fonema por su respectiva categoría fonética de acuerdo la lista de categorías fonéticas (Ver anexo D). El resultado de este procedimiento es la lista de n-grams codificados (LNC).

6.- Clasificación de las marcas. Esta etapa tiene como finalidad obtener el modelo óptimo a través del entrenamiento de la red neuronal y la utilización los pesos obtenidos para clasificar las marcas en dos clases (similares fonéticamente y no similares). Este dato es proporcionado por la base de datos de Jurisprudencia de casos de similitud fonética. El dato es una red neuronal de retro-propagación. Los desempeños alcanzados serán medidos en función al nivel de reconocimiento y precisión, y verificar si se está alcanzando los niveles de desempeño planteados (mayor a 0.95) para validar la hipótesis. El detalle de los procedimientos y algoritmos utilizados en cada etapa serán desarrollados en los siguientes capítulos.

CAPÍTULO V

PRE-TRATAMIENTO DEL NOMBRE DE LA MARCA

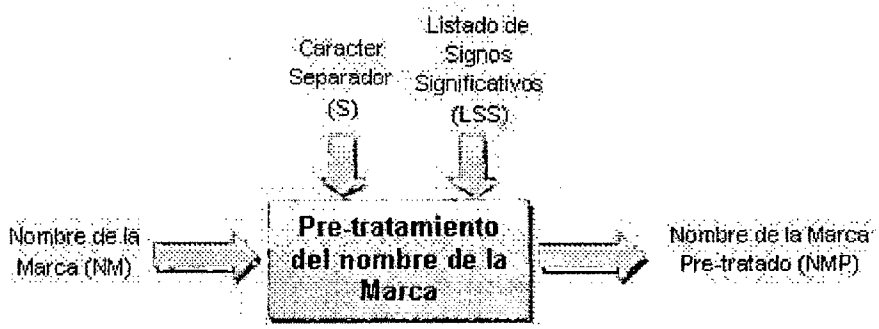


Figura V.1: Pre-tratamiento del nombre de la marca

Esta etapa tiene por finalidad eliminar a aquellos signos que carezcan de significado y puedan distorsionar la identificación de marcas similares a lo largo de sus etapas. Este procedimiento tiene como entrada al nombre en texto de la marca comercial (NM) que vine a ser una variable tipo cadena. Este procedimiento consiste en recorrer toda la cadena NM, a medida que se recorre la cadena NM se evalúa cada caracter (CEVA), evaluando primero si CEVA es una letra, si CEVA no es una letra se evalúa si CEVA tiene significado, mediante la comparación de CEVA con los signos de la lista de signos significativos (LSS), de estas evaluaciones se puede obtener 3 resultados:

1. CEVA es una letra

Este caso se da cuando CEVA se encuentra comprendido entre las letras del alfabeto español, tanto en mayúsculas como en minúsculas, se tiene que tener en cuenta que en muchos lenguajes de programación se usa el alfabeto americano el cual no incluye a la letra 'ñ'. Como CEVA es una letra este es agregado al nombre pretratado de la marca (NMP) mediante concatenación: NMP = NMP + CEVA.

2. CEVA es un signo significativo

Este caso se da cuando CEVA no es una letra, pero se encuentra dentro de la lista de signos significativos de la Base de Datos (BD) (Ver anexo A). La lista de signos significativos con sus respectivos significados se puede ver en la tabla 5.1. Por ejemplo el signo '@' es empleado en muchas marcas para representar la letra 'a' por lo cual si tiene significado.

Signo	Significado
+	mas
á	a
é	e
í	i
ó	o
ú	u
@	a
&	y

Cuadro V.1: Lista de Signos Significativos

Como CEVA es un signo significativo entonces se agrega el significado de CEVA al NMP mediante concatenación: NMP = NMP + obtenerSignificado(CEVA). Donde la función obtenerSignificado(CEVA) nos devuelve el significado de CEVA que se encuentra en LSS.

3. CEVA es un signo no significativo

Este caso se da cuando CEVA no es una letra ni se encuentra dentro de LSS, por tanto CEVA es un signo que necesita ser eliminado del NM.

Como CEVA no tiene significado simplemente este no es agregado a NMP, con lo cual se estaría garantizando la eliminación del mismo.

Se tiene que tener cuidado de no eliminar el carácter que es utilizado como carácter separador (S), lo que se recomienda para estos casos es ingresar a S como parámetro y validarlo a fin de no eliminarlo. El resultado es el nombre de la marca pre-tratado NMP, el cual no contendrá signos que puedan causar distorsión a excepción del carácter separador S, el cual será utilizado durante la siguiente etapa.

Pseudocódigo 5.1:

Procedure: NMP = *pretratamiento*(NM, S)

1. **desde** i = 1 **hasta** *longitud*(NM)
2. CEVA = *caracterEn*(NM,i)
3. si ((*esLetra?*(CEVA) = verdadero) ó (CEVA = S)) entonces
4. NMP = NMP + CEVA
5. sino
6. si (*tieneSignificado?*(CEVA) = verdadero) entonces
7. NMP = NMP + *obtenerSignificado*(CEVA)
8. fin si
9. fin si
10. **fin desde**

NM = Nombre en texto de la Marca.

S = Caracter separador de palabras en la marca.

NMP = Nombre Pretratado.

CEVA = Caracter a ser evaluado.

Durante este procedimientos se utilizan las siguientes funciones auxiliares:

devolverCaracterEn(x, i).- Devuelve el caracter de la cadena 'x' que esta en la posición 'i'.

esLetra?(x).- Devuelve verdadero si 'x' es una letra, de lo contrario devuelve falso.

tieneSignificado?(x).- Devuelve verdadero si 'x' tiene un significado asociado en la LSS de base de datos (BD).

obtenerSignificado(x).- Devuelve el significado de 'x' que se encuentra en la LSS.

Finalmente tenemos que si ingresamos el **NM** = 'L@BNET' al procedimiento obtendríamos como resultado a **NMP** = 'LABNET'.

Se realizará el seguimiento de un mismo caso a modo de ejemplo para poder mostrar una mejor aplicación de los procesos dentro del modelo de solución planteado.

Ejemplo: El caso a seguir consta de 5 marcas comerciales de la clase 5 de NIZA las cuales son: vita salutis, dermaglos, blue drop's, gastroprazol!, prosta teec. La figura V.2 nos presenta la aplicación del pre-tratamiento sobre las marcas brindadas. En el ejemplo se puede apreciar la eliminación de los caracteres «'» y «!» como producto del pre tratamiento, también se puede notar la utilización del caracter separador, el cual se utiliza para informar al proceso que dicho caracter no debe ser eliminado. Así por ejemplo se obtuvo la marca pre-tratada «blue drops» a partir de la marca «blue drop's».

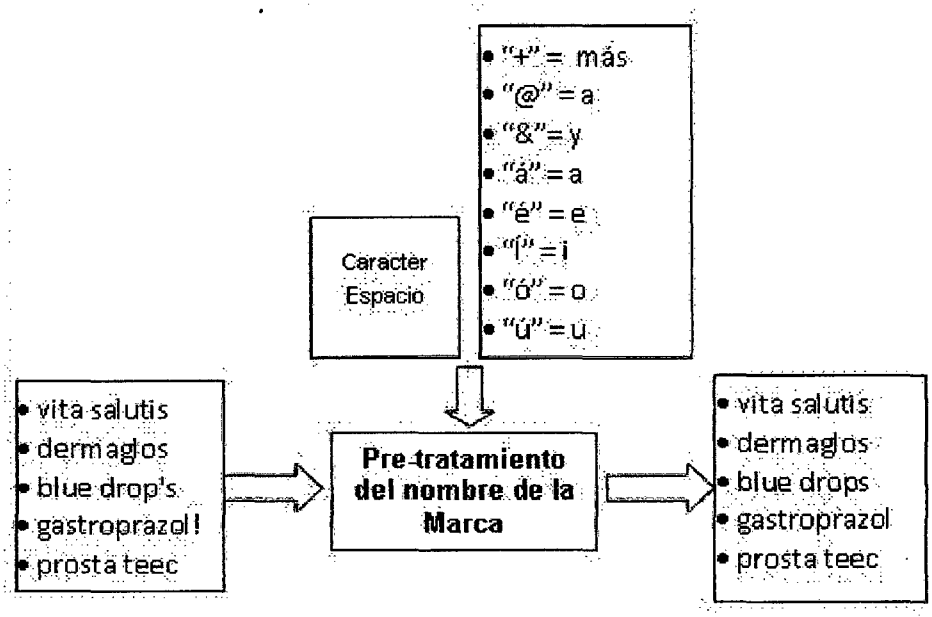


Figura V.2: Ejemplo Pre-tratamiento del nombre de la marca

CAPÍTULO VI

PARSING

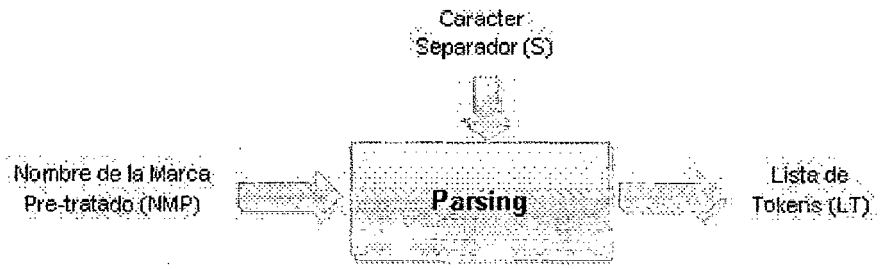


Figura VI.1: Parsing

Esta etapa tiene por finalidad dividir al nombre de la marca en tokens (palabras) los cuales estarán separados por el caracter separador (S). Este procedimiento tiene como entradas a NMP y S. Este procedimiento consiste en recorrer toda la cadena NMP, y evaluar si el caracter (CEVA) es igual a S. Si CEVA es diferente de S, entonces el caracter es agregado a un Token (T) por medio de concatenación. La idea es ir recorriendo la cadena NMP e ir acumulando los caracteres para formar un token, el token será guardado en a lista de Tokens LT cuando se encuentre a S, ya que S nos indica la separación entre palabras, entonces se continuará recorriendo NMP y se irá acumulando un nuevo T hasta encontrar el siguiente S o hasta que se encuentre el final de

NMP. Se obtiene como resultado a LT que será utilizado como entrada en el siguiente procedimiento.

Así por ejemplo para la marca 'sello de oro', los tokens serian: 'sello', 'de', 'oro'.

Pseudocódigo 6.1:

Procedure: LT = *parsing*(NMP, S)

1. **desde** i = 1 **hasta** *longitud*(NMP)
2. CEVA = *caracterEn*(NMP,i)
3. si (CEVA = S) entonces
4. añadir(LT ,T)
5. *reiniciar*(T)
6. sino
7. T = T + CEVA
8. si (i= *longitud?*(NMP)) entonces
9. *agregar*(LT ,T)
10. fin si
11. fin si
12. **fin desde**

S = Carácter separador de palabras en la marca.

NMP = Nombre Pretratado.

T = Token.

LT = Lista de Tokens.

CEVA = Caracter a ser evaluado.

NT = Nombre del Token.

CONS = Contador de separadores S.

Durante este procedimientos se utilizan las siguientes funciones auxiliares

agregar(T, LT).- Esta función añade el Token 'T' a la lista de Tokens 'LT'

Ejemplo: La figura VI.2 nos presenta la aplicación del proceso de parsing sobre las marcas pre-tratadas. En el ejemplo se puede apreciar la utilización del caracter separador para dividir a las marcas pre-tratadas en sus respectivos tokens. Así por ejemplo se obtuvieron los tokens «blue» y «drops» a partir de la marca pre-tratada «blue drops».

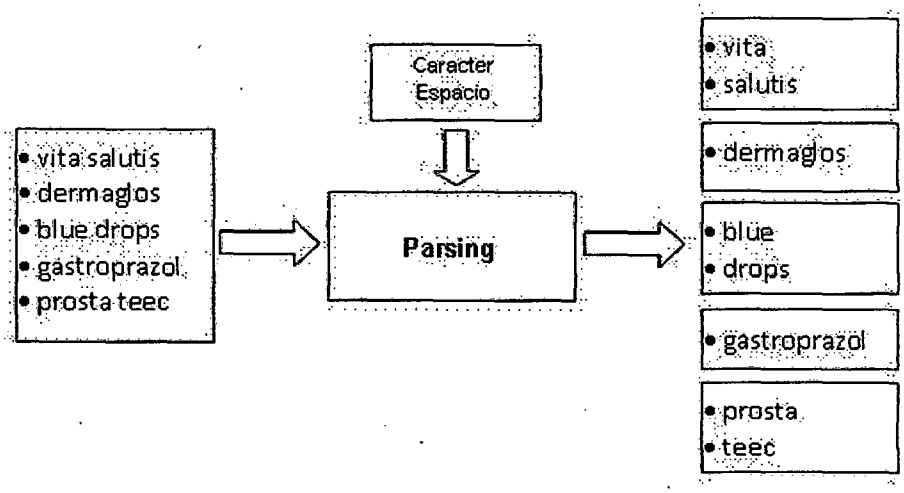


Figura VI.2: Ejemplo Parsing

CAPÍTULO VII

PRE-TRATAMIENTO DE LOS TOKENS

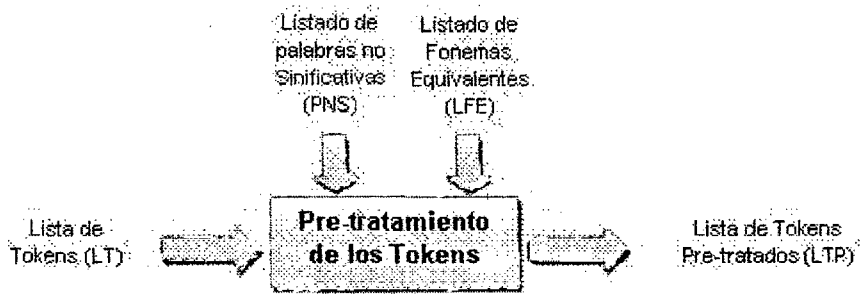


Figura VII.1: Pre-tratamiento del nombre de los Tokens

Esta etapa tiene por finalidad eliminar a aquellos tokens que no posean significado o se encuentren directamente asociados a las definiciones de la clase Niza a la que pertenecen y así mismo realizar una simplificación fonética a aquellos tokens que no fuesen eliminados. Este procedimiento tiene como entrada a la lista de tokens LT obtenidos de la etapa anterior. Este procedimiento consta de dos subrpocedimientos.

VII.1. IDENTIFICACION DE LOS ELEMENTOS NO SIGNIFICATIVOS

La finalidad de este procedimiento es eliminar a aquellos tokens que no tenga significado o que no estén permitidos. Este procedimiento tiene como entrada a la LT obtenidos del procedimiento anterior.

Primero se recorre a LT y se identifica si cada uno de los Tokens T es significativo, para esto se compara a T con la lista de palabras no significativas (PNS) (Ver anexo B). PNS contiene dos tipos de palabras: aquellas que carecen de significado debido a sirven de conectores gramaticales ('y', 'de', 'o', etc) y aquellas que no están permitidas dentro de la clase a la que pertenece el producto a ser registrado. Por ejemplo para a Clase Niza 5 que es de Productos farmacéuticos, veterinarios y venenos para animales entre otros o está permitido que el nombre de la marca contenga las palabras veterinario, farmacéutico o veneno.

Si T no se encuentra dentro de PNS entonces este es añadido a la lista de tokens significativos (LTS), de lo contrario T es eliminado, que vendría a ser lo mismo que no añadir T a LTS.

Pseudocódigo 7.1:

Procedure: LTS = $IdentificacionPalabrasNoSignificativas$(LT,CN)

1. **desde** i = 1 **hasta** $numeroDeTokens$(LT)
2. TEVA = $obtenerToken$(LT, i)
3. si($esPalabraNoSignificativa$?(TEVA, CN) = verdadero) entonces
4. //Nose hace nada
5. sino
6. $agregar$(TEVA, LTS)
7. fin si
8. **fin desde**

CN = Clase Niza a la que pertenece la marca.

LT = Lista de Tokens.

TEVA = Token a ser evaluado.

LTS = Lista de Tokens significativos.

Se utilizan las siguientes funciones auxiliares:

numeroDeTokens(LT).- Número de Tokens que contiene la lista LT.

obtenerToken(LT, i).- Esta función obtiene el Token en la posición 'i' de la lista de Tokens LT.

esPalabraNoSignificativa?(x, CN).- Devuelve verdadero si el Token 'x' es una palabra no permitida de la BD.

VII.2. PRETRATAMIENTO FONETICO DE LOS TOKENS

Este procedimiento tiene como finalidad remplazar los fonemas o par de fonemas por sus equivalentes simplificados. Este procedimiento tiene como entrada a las LTS obtenidos del sub-procedimiento anterior.

Dentro de los nombres de las marcas comerciales se utilizan pares de fonemas que podrían o necesitarían ser simplificados para poder ser codificados sin problemas. Así por ejemplo si tenemos el token queso, los fonemas /q/-/u/-/e/-/s/-/o/ serían los fonemas identificados cuando en realidad los fonemas identificados deberían ser /k/-/e/-/s/-/o/. Por tal motivo es necesario realizar remplazar a los fonemas /q/-/u/ por su equivalente simplificado /k/. Primero se recorre LTS y para cada token T se realiza la simplificación fonética, la cual consiste en recorrer a T de dos en dos fonemas y remplazar el par de fonemas evaluado (FEVA) por su valor simplificado dentro de la lista de fonemas equivalentes (LFE) (Ver anexo C). En caso el par de fonemas tengan una simplificación a un solo fonema, entonces se agrega un espacio en blanco para mantener la longitud de T. Esto debido a que se hacen dos recorridos a T,

un recorrido empezando por los fonemas pares y otro por los fonemas impares. La idea es mantener la misma longitud de T después del recorrido (par) para continuar de forma adecuada con el segundo recorrido (impar). Ya que al eliminarse un fonema dentro de un par se utilizaría al siguiente fonema para formar un nuevo par rompiéndose así la integridad de los pares de fonemas. En la tabla 6.1 se pueden mostrar los pares de fonemas y sus respectivas simplificaciones fonéticas.

Pseudocódigo 7.2:

Procedure: LTP = *pretratamientoFoneticoDeLosTokens*(LTS)

1. **desde** i = 1 **hasta** *numeroDeTokens*(LTS)
2. TS = *obtenerToken*(LTS, i)
3. n = *longitud*(TS)
4. NP1 = n / 2
5. si (*esPar?*(n)) entonces
6. NP2 = n/2
7. sino
8. NP2 = n/2 - 1
9. fin si
10. desde j=1 hasta NP1
11. FEVA = *caracterEn*(TS, 2*j) + *caracterEn*(TS, 2*j+1)
12. TP = TP+ *valorSimplificado*(FEVA);
13. fin desde
14. TS = TP
15. TP = *caracterEn*(TS, 1)
16. si (*esImpar?*(n)) entonces
17. TP = *caracterEn*(TS, n)
18. fin si

19. Desde j=1 hasta NP2
20. FEVA = $caracterEn$(TS, 2*j+1) + $caracterEn$(TS, 2*j+2)
21. TP = TP+ $valorSimplificado$(FEVA);
22. fin desde
23. si ($esPar?$(n)) entonces
24. TP = caracterEn(TS, n)
25. fin si
26. TS = TP
27. Desde j=1 hasta n
28. Si ($caracterEn$(TS)=' ') entonces
29. // no se hace nada
30. sino
31. TP = caracterEn(TS, j)
32. fin si
33. fin desde
34. $agregar$(TP, LTP)
35. **fin desde**

NP1 = Número de pares de fonemas en el Token empezando a contar desde uno.

NP2 = Número de pares de fonemas en el Token empezando a contar desde dos.

TS = Token significativo

TP = Token pretratado

LT = Lista de Tokens.

FEVA = Fonema a ser evaluado.

LTS = Lista de Tokens significativos.

Funciones adicionales **esPar?(n).-** Devuelve verdadero si el número 'n' es par.

esImpar?(n).- Devuelve verdadero si el número 'n' es impar.

valorSimplificado(x).- devuelve el valor simplificado del fonema 'x' que se encuentra en la BD agregando espacios ' ' en caso la simplificación del fonema posea solo una letra (LL pasa a L p.e) , de lo contrario devuelve el mismo fonema.

Ejemplo: La figura VII.2 nos presenta la aplicación del pre-tratamiento sobre los tokens obtenidos a partir de las marcas brindadas. En el ejemplo se puede apreciar la simplificación fonética del token «teec» a su forma simple «tec», esto debido a que dentro de la lista de fonemas equivalentes se especifica que «ee» es equivalente a «e».

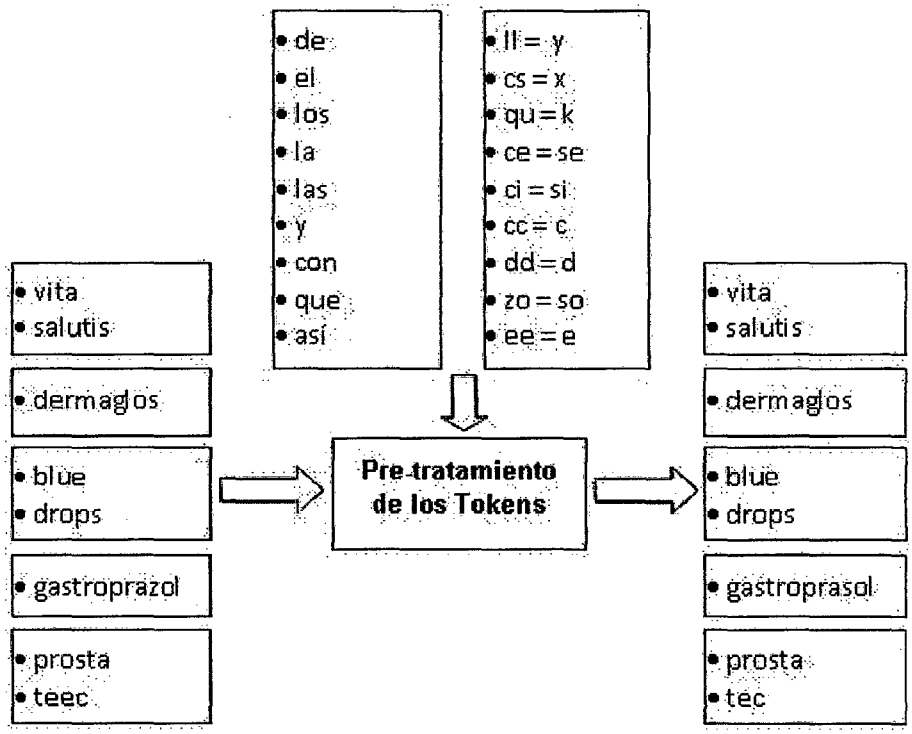


Figura VII.2: Ejemplo Pre-tratamiento del nombre de los Tokens

CAPÍTULO VIII

GENERACIÓN DE LOS NGRAMS

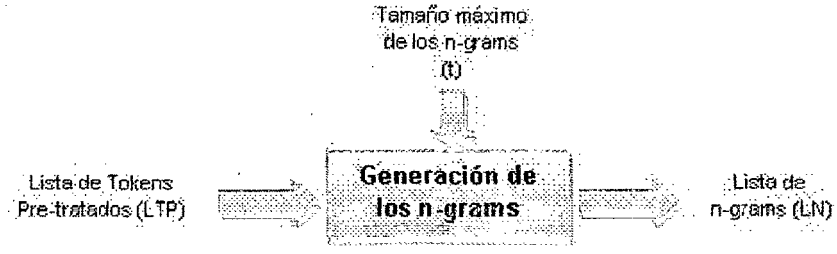


Figura VIII.1: Generación de los n-grams

Este procedimiento tiene por finalidad generar los n-grams a partir de los tokens, donde un n-gram es una subcadena de una longitud fija. Este procedimiento tiene como entradas a la lista de tokens pre-tratados (LTP) y al tamaño máximo de los n-grams (t). Luego se procede a realizar la generación de los n-grams de tamaño t1 = 2, 3, ... t, para cada valor de t1 se verifica si el token tiene una longitud (L) mayor a t1, de ser el caso se procede a recorrer el token hasta la posición (L-t1) y creando a su vez las sub-cadenas de tamaño t1 que vienen a ser los n-grams, de tal forma que se obtendrán (L-t1+1) n-grams. Para el caso de n-grams que tengan longitud menor a t1 se añadirá (t1-L) caracteres 'a', obteniendo así un n-gram. El procedimiento se realizará

para cada token de la LTP. El resultado es la lista de n-grams (LN).

Pseudocódigo 8.1:

Procedure: LNG = *generacionDeLosNGrams*(LTP,t)

```
1. desde i=2 hasta t
2.    t1 = i
3.    si ( longitud(NTP)>t1) entonces
4.       desde j=1 hasta  longitud(NTP)
5.          NG =  subcadena(NTP, j, j+t1)
6.          agregar(LNG,NG)
7.       fin desde
8.    sino
9.       NG=NTP
10.      si( longitud (NTP)<t1) entonces
11.         desde j=1 hasta (t1- longitud(NTP))
12.            NG=NG+'a'
13.         fin desde
14.      fin si
15.       agregar(LNG,NG)
16.   fin si
17. fin desde
```

t = Tamaño máximo de los NGrams.

t1 = Tamaño de los NGrams .

NTP = Nombre del Token Pretratado.

NG = NGram.

Se usan las siguientes funciones auxiliares:

subcadena(C, x, y).- devuelve la subcadena de la cadena 'C' que empieza en 'x' y termina en 'y'. x,y posiciones de la cadena C.

Ejemplo: La figura VIII.2 nos presenta la aplicación del proceso de generación de los n-grams para el caso de t1=2 es decir para generación de n-grams de tamaño 2. El proceso debe ser repetido pero tomando a t1=3 y t1=4 dado que se está especificando que el tamaño máximo de los n-grams es de t=4.

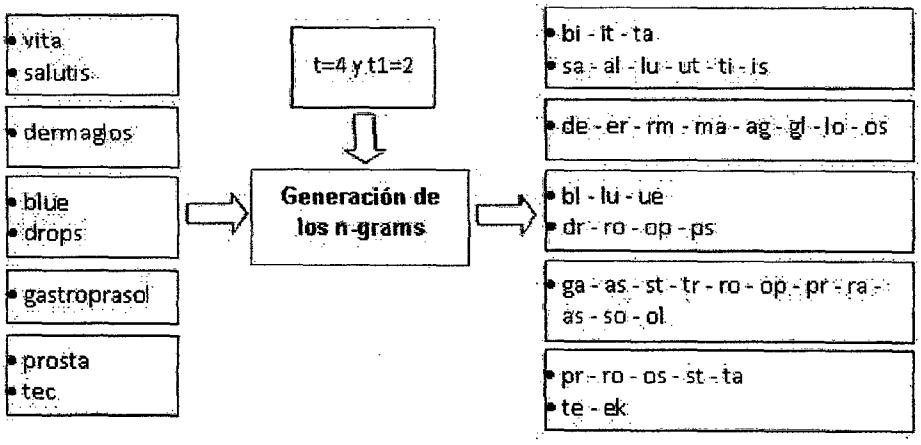


Figura VIII.2: Ejemplo Generación de los n-grams

CAPÍTULO IX

CODIFICACIÓN DE LOS NGRAMS

Figura IX.1: Codificación de los n-grams

Esta etapa tiene como finalidad la codificación de los fonemas de los n-grams. La entrada de este proceso es la lista de n-grams (LN). Primero se procede a recorrer todo el n-gram, caracter por caracter, y se remplaza cada fonema por su respectiva categoría fonética de acuerdo la lista de categorías fonéticas. El resultado de este procedimiento es la lista de n-grams codificados (LNC).

Por ejemplo para los n-grams 'sey,'eyo' del token 'seyo', su codificación de acuerdo a la categoría fonética basada en el IPA, sería:

(sey) (10 - 1 - 12)

(eyo) (1 - 12 - 0)

Código	Fonemas
0	A O
1	E I
2	U
3	B V
4	C K
5	D T
6	L R
7	M N
8	G J
9	F P
10	S Z
11	X
12	Y

Cuadro IX.1: Categorias Fonéticas

Pseudocódigo 9.1:

Procedure: LNGC = *codificacionDeLosNGrams*(LNG)

1. **desde** i=1 hasta *numeroNGrams*(LNG)
2. TNG = *obtener*(LNG, i)
3. desde i=1 hasta *longitud*(TNG)
4. NG = NG + ';' + *devolverCodigo*(caracterEn(TNG, i))
5. fin desde
6. agregar(LNGC, NGC)
7. **fin desde**

TNG = Texto del NGram.

NGC = NGram Codificado.

Se usan las siguientes funcionaes auxiliares:

numeroNGrams(LNG).- Devuelve el número de NGrams que contiene la lista LNG. devolverCodigo(X) = Devuelve el código del fonema 'X'.

Ejemplo: La figura IX.2 nos presenta la aplicación del proceso de codificación de los n-grams de tamaño 2 obtenidos en el proceso anterior. El proceso debe ser repetido para los n-grams de tamaño 3 y 4 del proceso anterior.

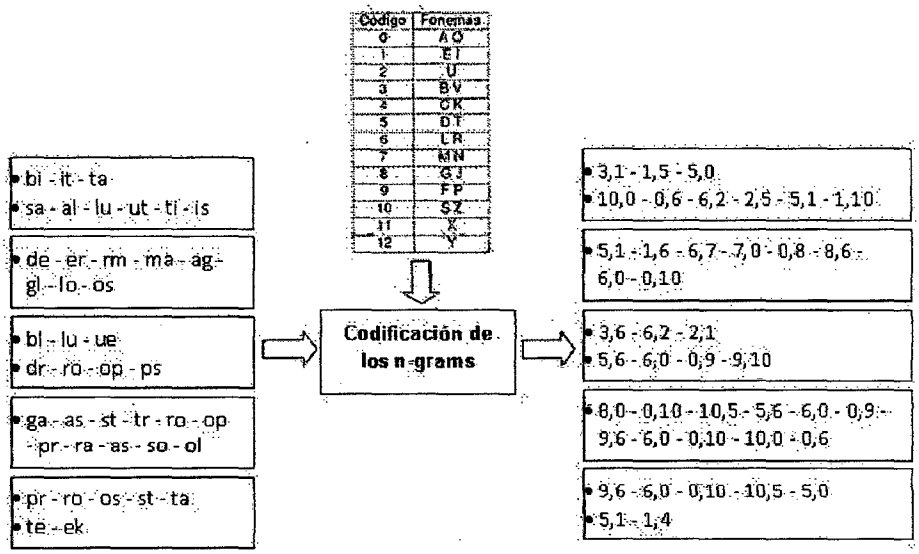


Figura IX.2: Ejemplo Codificación de los n-grams

CAPÍTULO X

CLASIFICACIÓN DE LAS MARCAS

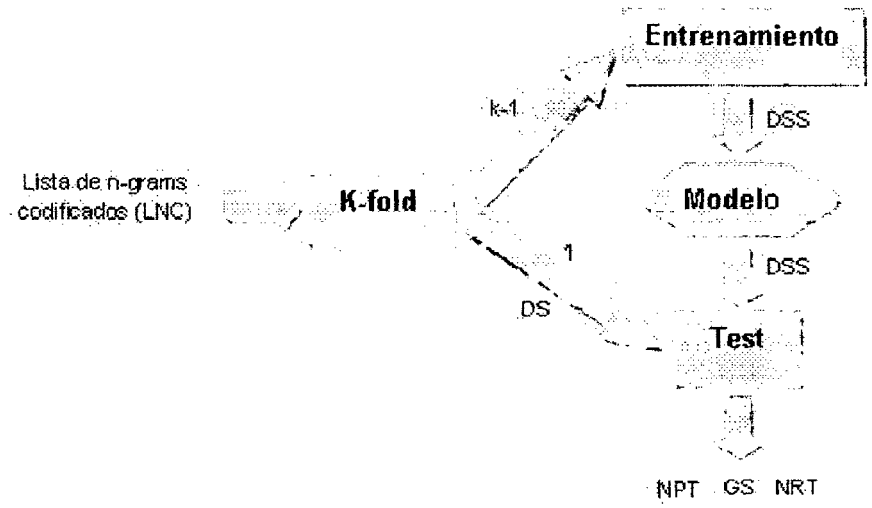


Figura X.1: Clasificación de las marcas

Esta etapa tiene como finalidad encontrar el modelo óptimo que permita la clasificación de las marcas en similares y no similares con los valores más altos de reconocimiento y precisión. Para lo cual se utilizará la técnica de Neuronales Artificiales. Para esto se toma como parámetros de ingreso para la red neuronal la cantidad de n-grams iguales que tienen las marcas a ser comparadas, se tienen 3 inputs que son la cantidad n-grams de tamaño 2 que tienen en

común, la cantidad de n-grams de tamaño 3 que tienen en común y la cantidad de tamaño de n-grams de tamaño 4 que tienen en común.

X.1. ARQUITECTURA DE LA RED NEURONAL

La Red Neuronal utilizada es la red de Retropropagación, se utilizó este tipo de Red Neuronal debido a dos aspectos: primero esta es una de las redes más utilizadas y conocidas por nosotros los autores y segundo que este tipo de Red Neuronal ya ha sido utilizado para la clasificación y detección de nombres similares (Kukich, [16]). Se puede apreciar la Arquitectura de la Red en la Figura X.2.

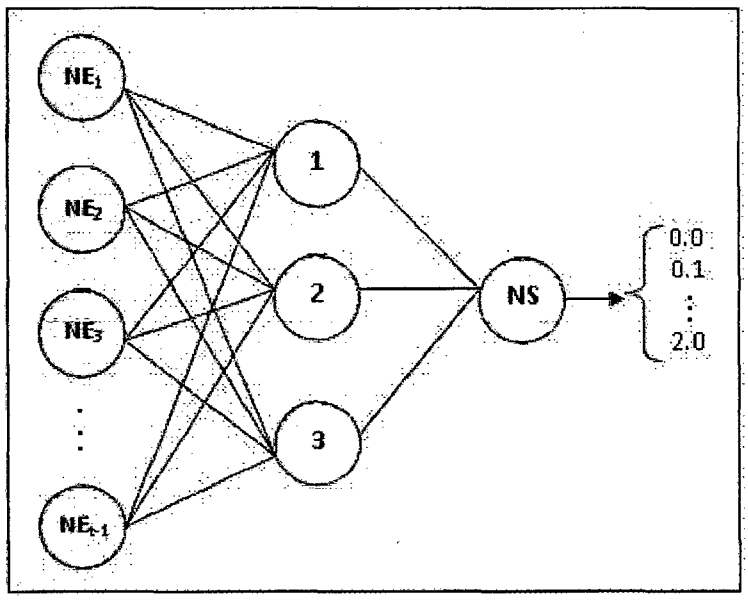


Figura X.2: Arquitectura de la Red Neuronal

X.1.1. ESTRUCTURA DE LAS CAPAS DE ENTRADA

La estructura de la entrada de la red neuronal estará conformada por (t-1) nodos o neuronas de entrada (NE) de tipo de entrada numérica, donde t es la cantidad de n-grams utilizada en el experimento. Para el cálculo del valor

de la neurona de entrada n se realizará la comparación entre los n-grams de tamaño n+1 de las marcas a ser comparadas, el valor será igual a la razón de la cantidad de n-grams similares y la cantidad de comparaciones posibles entre n-grams de dicho tamaño. Así por ejemplo para las marcas 'kuevo' y 'cevo', para la neurona de entrada 1 se compararán los ngrams de tamaño 2, tenemos que 'kuevo' tiene los n-grams 'ku','ue','ev','vo' y cevo tiene los n-grams 'ce','ev','vo' tenemos que en común tienen a 2 n-grams 'ev','vo' y el número de combinaciones posibles es de 12 (producto de la cantidades de n-grams de tamaño 2 de ambas marcas).

$$Parametro\ de\ Entrada\ n = \frac{Nro\ n-grams\ similares\ tam.n+1}{Nro\ de\ comparaciones\ posibles\ tam.n+1} \quad \text{(X.1)}$$

Ejemplo: La figura X.3 nos presenta un ejemplo de la obtención del parámetro de entrada donde los n-grams tienen tamaño 2, se compara los n-grams codificados de de las marcas «blue drops» y «gastroprazol», la primera cuenta con 7 n-grams codificados y la segunda con 11, por lo cual el número de comparaciones posibles es de 7 por 11 que daría un total de 77 comparaciones posibles. Tambien se tiene que ambas marcas tienen en común 3 n-grams codificados («0,5»,«6,0»,«0,9»), para este caso se está comparando la marca «blue drops» contra la marca «gastroprazol» por lo cual no se cuenta el n-gram codificado repetido «6,0». Para este caso el valor del parámetro de entrada sería igual a 0.039.

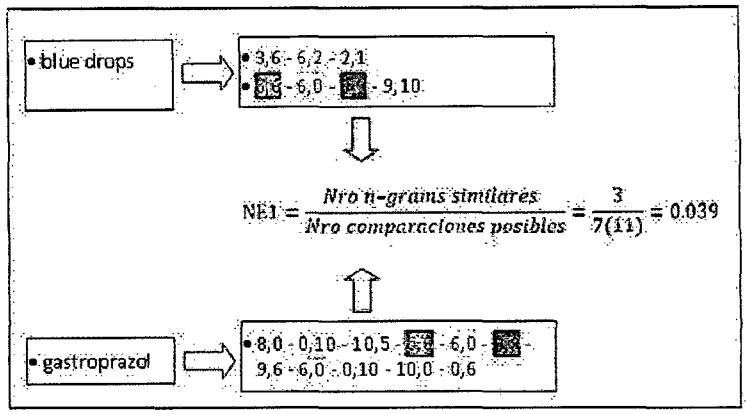


Figura X.3: Ejemplo Parámetro de Entrenamiento

X.1.2. ESTRUCTURA DE LAS CAPAS OCULTAS

Para determinar el número óptimo de neuronas ocultas, nos basaremos en la experimentación. Esta parte es netamente empírica. Se probaron con capas de 1 a 10 siendo el valor de 3 el valor que nos permitía obtener los valores más óptimos de Reconocimiento y Precisión.

Nro NO .	Prom. REC	Prom. PREC	Prom. EFEC	Varianza REC	Varianza PREC	Varianza EFEC
1	0.9522	0.4999	0.90697	0.0034	0.0032	0.0031
2	0.9615	0.503	0.91565	0.0041	0.0023	0.002
3	0.992	0.653	0.9581	0	0.001	0.0015
4	0.9723	0.6311	0.93818	0.0021	0.0019	0.0019
5	0.9701	0.6412	0.93721	0.002	0.0031	0.0025
6	0.9555	0.6534	0.92529	0.0065	0.0192	0.0109
7	0.9478	0.6612	0.91914	0.0013	0.03301	0.0051
8	0.9433	0.6505	0.91402	0.0091	0.0312	0.0612
9	0.9401	0.6721	0.9133	0.0012	0.0021	0.017
10	0.9311	0.7012	0.90811	0.0012	0.0031	0.0041

Cuadro X.1: Estimación Nro de Neuronas Ocultas

Para obtener el número de neuronas ocultas (NO) de la Red Neuronal se realizaron una serie de pruebas para así obtener el número de capas ocultas que nos permitiera obtener los valores de Reconocimiento y Precisión más elevados. Para esto se realizaron 100 pruebas para los valores del 1 al 10 como número de neuronas ocultas, se realizaron 10 pruebas por cada uno de los valores planteados, es decir se realizaron 10 pruebas teniendo como 1 al número de neuronas ocultas, 10 pruebas teniendo como 2 al número de neuronas ocultas y así sucesivamente (Ver Cuadro X.1 y Figura X.4). Del experimento se pudo encontrar que el número óptimo de neuronas ocultas fue el de 3 .

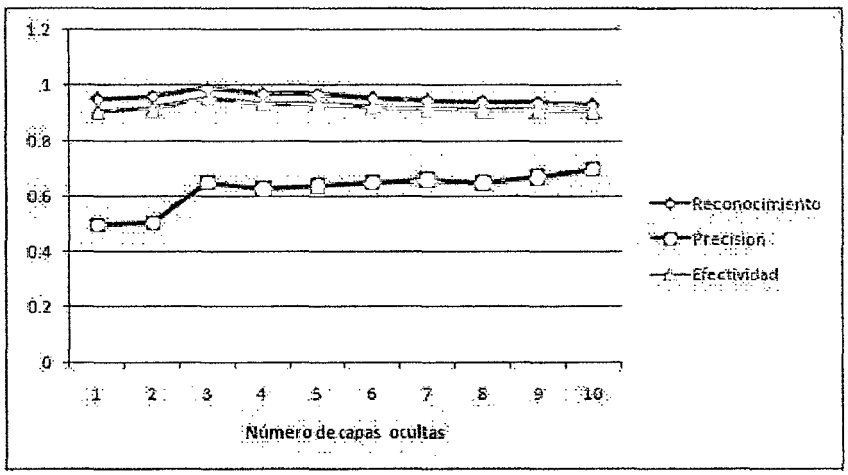


Figura X.4: Capas Ocultas VS Reconocimiento, Precisión y Efectividad

X.1.3. ESTRUCTURA DE LAS CAPAS DE SALIDA

la estructura de la entrada de la red neuronal estará conformada por 1 nodo o neurona de salida (NS) tipo de entrada numérica, que viene a ser el nivel de similitud entre las marcas, su valores oscilan de 0 a 2, obteniéndose el valor de 0 cuando son completamente iguales y el valor de 2 cuando son completamente diferentes.

Para el entrenamiento de la red Neuronal se utilizó el valor de 100 como cantidad de épocas (Ver Anexo 3).

X.2. DIVISIÓN DE DATOS A TRAVÉS DE K-FOLD

Se realiza la separación de los datos a través de la utilización del esquema de experimentación k-Fold, con lo cual la data es dividida en k partes iguales de las cuales k-1 partes son utilizadas para el entrenamiento y 1 parte es utilizada para las pruebas, ver Anexo 1.

X.3. ENTRENAMIENTO

La intención del entrenamiento es la de obtener los pesos óptimos a través de la adaptación que se da durante el entrenamiento, esto permite ir reduciendo el error hasta un valor de error mínimo establecido como meta. Una vez obtenidos los pesos óptimos éstos serán utilizados para la etapa de clasificación.

Para cada neurona de entrada o parámetro de entrada se realiza la comparación de cada uno de los ngrams de la marca registrada con los ngrams de la marca por registrar.

Por ejemplo si tenemos que la primera marca tiene 10 n-grams de tamaño 3 y las segunda marca tiene 5 marcas del mismo tamaño, el parámetro de ingreso será la cantidad de comparaciones posibles como denominador y la cantidad de n-grams similares como numerador. Veamos si en este caso las marcas tuviesen 4 n-grams en común el valor del parámetro de ingreso sería 4/50, lo cual no estaría diciendo que de las 50 comparaciones que se han realizado 4 han sido efectivas, es decir se encontró igualdad.

Para la realización del entrenamiento se utilizaron los casos de jurisprudencia como data de prueba, es decir nosotros ya asumimos que las marcas

encontradas en los casos de jurisprudencia son similares, su valor entonces de clasificación será 0, las otras marcas a utilizarse serán las otras marcas encontradas y su valor de clasificación será 2 ya que estamos asumiendo que son diferentes porque ya han sido registradas y no se encontró similitud alguna.

X.4. CLASIFICACIÓN

Para cada neurona de entrada o parámetro de entrada se realiza la comparación de cada uno de los ngrams de la marca registrada con los ngrams de la marca por registrar.

Por ejemplo usando como tamaño de n-gram 3, para los n-grams 'sey,'eyo','oro' de la marca 'seyo de oro', y los n-grams 'oro,'bel','elt' de la marca 'oro belt' tenemos que se harán en total 9 comparaciones, primero se comparará el n-gram 'sey' con los 3 n-grams de la marcas 'oro belt' y se proseguirá de la misma forma con los otros 2 n-grams de la marca 'seyo de oro'. como podemos apreciar solo el n-gram 'oro' es común para las dos marcas por lo que el valor del parámetro de ingreso será 1/9 = 0.111. Luego de realizar este procedimiento vamos a obtener por cada marca a ser comparada con la marca registrada (t-1) parámetros de entrada, donde t es la cantidad de n-grams utilizada en el expiremento. Para clasificar las marcas similares se tendrá primero que obtener los parámetros de entrada para cada marca producto del procedimiento anterior. Luego de la clasificación y utilizando la data de prueba para el test nosotros obtendremos valores similares producto del entrenamiento DSS. entonces si el valor obtenido es cercano a 0 la marca será considerada como similar y si el valor es cercano a 2 será considerada como diferente. para diferenciar los valores de forma formal se utilizará el parámetro VDF que vendría a ser el valor diferenciador, es decir si el valor obtenido del entrenamiento es menor que

el VDF entonces la marca será considerada similar, y si es mayor al VDF será considerada como diferente. y si además tenemos que el valor de similitud de los datos de prueba es de 0 esta será una marcas similar CS estaríamos ante un caso de una marca reconocida que realmente es similar CSI.

Entonces a partir de la cantidad de marcas identificadas CI, la cantidad de marcas similares identificadas CSI y y la cantidad de marcas similares CS podremos obtener el nivel de reconocimiento y precisión del experimento.

Pseudocódigo 10.1:

Procedure: LNGC = $clasificacionDeLasMarcas$(LNC)

1. **desde** i=1 hasta $numeroMarcas$(LNG)
2. PE = agregar($obtenerParametros$(LNG, i))
3. **fin desde**
4. DSS = entrenarRedNeuronal(PE)
5. **desde** i=1 hasta $numeroMarcas$(LNG)
6. GS = DSS(i)
7. GP = DSS(i)-DS(i)
8. **si** GS menor que VDF
9. CI = CI+CI
10. **si** DS(i) = 0
11. CSI = CSI+CSI
12. **fin si**
13. **fin si**
14. **si** DS(i) = 0
15. CS = CS+CS
16. **fin si**
17. CM = CM+CM
18. **fin desde**

19. NRT = CSI/CS
20. NPT = CSI/CI

LNG = Parámetros de ingreso de las marcas

DSS = Data de Salida obtenida de la Simulación

DS = Data de salida base (Obtenida de la Jurisprudencia)

GS = Grado de similitud de cada marca

GP = Grado de precisión de cada marca (es el error)

VDF = Valor Diferenciador (entre similares y no similares)

CI = Cantidad de Marcas Identificadas

CSI = Cantidad de Marcas Similares Identificadas

CS = Cantidad de Marcas Similares

CM = Cantidad de Marcas

NRT = Nivel de Reconocimiento Total

NPT = Nivel de Precisión Total

Se usan las siguientes funcionaes auxiliares:

entrenarRedNeuronal(LNG).- Realiza el entrenamiento de la red Neuronal y da como resultado la Data de Salida de la Simulación.

Ejemplo: La figura X.5 nos presenta los datos de entrada de la marca «vita salutis» cuyo idMarca es 52, se presentan 3 casos, el primero y el tercero que es la comparación de la marca consigo misma y el segundo que es la comparación con la marca luego de una modificación realizada de forma aleatoria, los otros registros son resultado de la comparación de la marca con las otras marcas comerciales de la misma clase.

Como salida se obtiene el nivel de similitud y en nivel de precisión para cada marca, así como el identificador de la misma y el valor de similitud que debió ser obtenido (Ver Figura X.6). 1041 es el idMarca de la marca tomada de la jurisprudencia, donde es especificada como similar, como podemos apreciar

el valor de similitud es de 0, con lo cual nos dice que si son similares. Los registros restantes son de otras marcas dentro de la misma clase cuyo valor de similitud es de 2, lo cual nos dice que son diferentes.

NE 1	NE 2	NE 3	Clase	IdMarca
0.25	0.3333	0.5	0	52
0.166	0.1666	0.0	0	52
0.25	0.3333	0.5	0	52
0.05	0.0	0.0	2	132
0.035	0.0	0.0	2	175
0.0	0.0	0.0	2	176
0.0	0.0	0.0	2	213
0.0	0.0	0.0	2	236
0.0416	0.0	0.0	2	241
0.0	0.0	0.0	2	282
0.0	0.0	0.0	2	293
0.0	0.0	0.0	2	330

Figura X.5: Datos de Entrenamiento

Similitud	Precisión	idMarca	Clase
0	0	1041	0
1.99	0.00975	339	2
1.995	0.005202	353	2
1.983	0.01685	373	2
1.995	0.005202	409	2
1.988	0.01247	414	2
1.995	0.005202	482	2
1.995	0.005202	567	2
1.995	0.005202	661	2
1.995	0.005202	698	2
1.995	0.005202	703	2

Figura X.6: Datos de Salida

CAPÍTULO XI

EXPERIMENTACIÓN

XI.1. DISEÑO DEL EXPERIMENTO

La investigación será del tipo experimental, se utilizarán redes neuronales para la identificación similitud fonética entre marcas, para introducir el aspecto fonético se habrá de utilizar las categorías fonéticas, se determinará el nivel de precisión y reconocimiento de la herramienta en este proceso.

Se van a utilizar la categorización de los fonemas basada en el IPA en español, Así como el tamaño de las subcadenas de los n-grams, pudiéndose medir con eso el nivel de precisión y reconocimiento de la herramienta en el proceso.

Variables manipuladas	**Tamaño n=4**		**Tamaño n=5**	
Categoría basada en el IPA	**Variables a medir**		**Variables a medir**	
	%	%	%	%
	Precisión	Reconocimiento	Precisión	Reconocimiento

Figura XI.1: Diseño de la Investigación

XI.2. VARIABLES INDEPENDIENTES Y DEPENDIENTES

XI.2.1. VARIABLES INDEPENDIENTES

Variable Independiente	Definiciones Conceptuales	Definiciones Operacionales
Categoría de los fonemas	Agrupaciones de fonemas con sonidos similares.	Sistema de trascripción fonética del español IPA.
Tamaño "n"de los n-grams.	Tamaño definido para las subcadenas de los n-grams.	

Cuadro XI.1: Variables Independientes

Categorías de fonemas Las categorías asignadas a los fonemas estarán basadas en la similitud fonética de los fonemas, viendo aspectos de lugar de articulación y modo de articulación. Así por ejemplo los fonemas /p/ y /b/ se encontrarían dentro de una misma categoría debido a que ambos son articulados en la zona labial con un modo de articulación oclusivo, lo cual los hace similares fonéticamente.

Tamaño "n"de los n-grams Tamaño de las subcadenas en que los tokens de las marcas comerciales van a ser divididas, esta variable determina el número de los n-grams que se van a obtener.

XI.2.2. VARIABLES DEPENDIENTES

Precisión Viene a ser la proporción de marcas reconocidas que realmente son similares, mediante esta variable se va a poder medir la eficiencia de la herramienta presentada.

Reconocimiento Viene a ser la proporción de marcas similares que han sido reconocidas, es decir nos muestra de todas las marcas que han sido

Variable Dependiente	Definiciones Conceptuales	Definiciones Operacionales
Precisión	Proporción de marcas reconocidas que realmente son similares.	Comparación y conteo con casos de Indecopi.
Reconocimiento	Proporción de marcas similares que han sido reconocidas.	Comparación y conteo con casos de Indecopi.

Cuadro XI.2: Variables Dependientes

similares por Indecopi, cuantas han sido reconocidas. Mediante esta variable se va a poder medir la eficacia de la herramienta presentada.

En base al nivel de Precisión y Reconocimiento se realizará el cálculo del Nivel de Efectividad. Este indicador nos permitirá obtener el promedio ponderado de ambao. Además este indicador sería nuestra propuesta planteada para la medición del nivel de efectividad de la herramienta.

$$N.\,Efectividad = (N.\,Reconocimiento)(0{,}9) + (N.\,Precision)(0{,}1) \qquad \text{(XI.1)}$$

XI.2.3. VARIABLES DEL MODELO

Después de una revisión se pudo resolver las siguientes variables que forman parte del Modelo.

Variable del Modelo	Definiciones Conceptuales	Definiciones Operacionales
Fonema	Abstracciones mentales o abstracciones formales de los sonidos del habla.	Sistema de trascripción fonética del español IPA.
Posición de los fonemas	Ubicación determinada de un fonema dentro de una palabra	Conteo de posiciones

Cuadro XI.3: Variables del Modelo

Fonema Estos fonemas nos permiten representar de forma un tanto abstracta la pronunciación de una palabra, así por ejemplo la palabra profesor se podría representar mediante fonemas como /p, /r/, /o/, /f/, /e/, /s/, /o/, /r/. Se utilizarán los fonemas empleados en el habla española, de esta forma se logrará adecuar la investigación a nuestro contexto.

Posición de los fonemas La posición de un fonema en particular está determinado por la ubicación del fonema en la palabra, la posición se da de izquierda a derecha y esta podría empezar tanto en cero como en uno. Para esta investigación se tomará al uno como posición de inicio.

XI.3. EXPERIMENTO 1

Para el experimento se ha tomado en cuenta a las marcas de las clases NIZA con 30 o más marcas. Las cuales son las Clases: 1, 3, 5, 9, 16, 25, 29, 30, 41 y 43. Se han realizado 10 pruebas por cada clase utilizando k=3, donde k es el número de particiones a utilizarse (Ver Anexo 1).

XI.3.1. DESARROLLO DEL EXPERIMENTO

Este experimento tiene por finalidad hallar el valor óptimo del 'Valor Diferenciador' VDF, el criterio para seleccionar dicho valor diferenciador es que este permita hallar los niveles de Reconocimiento y precisión más altos y con una varianza menor.

Se realizaron 100 pruebas para determinar el nivel de Reconociminto, precisión y efectividad de la herramienta para cada uno de los valores del VDF que varia de 0.2 a 2.0 se obtuvieron los resultados mostrados en la tabla 11.4.

Nro.	C. Niza	Valor Dif.	Reconocim.	Precisión	Efectividad
1	1	0.2	0.75	1	0.78
2	1	0.4	0.92	0.92	0.92
3	1	0.6	0.92	1	0.93
4	1	0.8	1	0.92	0.99
5	1	1	1	0.92	0.99
6	1	1.2	1	0.86	0.99
7	1	1.4	1	0.86	0.99
8	1	1.6	1	0.86	0.99
9	1	1.8	1	0.86	0.99
10	1	2	1	0.08	0.91
11	3	0.2	0.83	0.83	0.83
12	3	0.4	0.83	1	0.85
13	3	0.6	0.83	0.91	0.84
14	3	0.8	0.83	0.83	0.83
15	3	1	0.83	0.83	0.83
16	3	1.2	0.92	0.85	0.91
17	3	1.4	0.92	0.79	0.9
18	3	1.6	1	0.67	0.97
19	3	1.8	1	0.57	0.96
20	3	2	1	0.08	0.91
21	5	0.2	0.69	0.69	0.69
22	5	0.4	0.77	0.67	0.76
23	5	0.6	0.77	0.53	0.74
24	5	0.8	0.77	0.83	0.78

Continúa en la siguiente Página

Cuadro XI.4: Estimación del parámetro Valor Diferenciador

Nro.	C. Niza	Valor Dif.	Reconocim.	Precisión	Efectividad
25	5	1	0.85	0.65	0.83
26	5	1.2	0.85	0.61	0.82
27	5	1.4	0.85	0.61	0.82
28	5	1.6	0.92	0.43	0.87
29	5	1.8	1	0.46	0.95
30	5	2	1	0.02	0.9
31	9	0.2	0.83	0.83	0.83
32	9	0.4	0.83	0.83	0.83
33	9	0.6	0.92	0.85	0.91
34	9	0.8	0.83	0.71	0.82
35	9	1	0.92	0.92	0.92
36	9	1.2	0.92	0.61	0.89
37	9	1.4	0.92	0.69	0.89
38	9	1.6	1	0.5	0.95
39	9	1.8	1	0.44	0.94
40	9	2	1	0.03	0.9
41	16	0.2	0.75	1	0.78
42	16	0.4	0.75	0.9	0.77
43	16	0.6	0.67	0.89	0.69
44	16	0.8	0.75	0.82	0.76
45	16	1	0.75	0.82	0.76
46	16	1.2	0.75	0.82	0.76
47	16	1.4	0.75	0.75	0.75
48	16	1.6	0.83	0.63	0.81

Continúa en la siguiente Página

Cuadro XI.4: Estimación del parámetro Valor Diferenciador

Nro.	C. Niza	Valor Dif.	Reconocim.	Precisión	Efectividad
49	16	1.8	1	0.4	0.94
50	16	2	1	0.1	0.91
51	25	0.2	0.67	0.67	0.67
52	25	0.4	0.67	0.67	0.67
53	25	0.6	0.75	0.6	0.74
54	25	0.8	0.75	0.56	0.73
55	25	1	0.75	0.43	0.72
56	25	1.2	0.83	0.53	0.8
57	25	1.4	0.75	0.5	0.73
58	25	1.6	0.83	0.43	0.79
59	25	1.8	0.92	0.46	0.87
60	25	2	1	0.02	0.9
61	29	0.2	0.75	1	0.78
62	29	0.4	0.83	1	0.85
63	29	0.6	0.92	0.79	0.9
64	29	0.8	0.92	0.79	0.9
65	29	1	0.92	0.65	0.89
66	29	1.2	0.92	0.73	0.9
67	29	1.4	1	0.8	0.98
68	29	1.6	1	0.57	0.96
69	29	1.8	1	0.36	0.94
70	29	2	1	0.12	0.91
71	30	0.2	0.67	0.89	0.69
72	30	0.4	0.67	0.67	0.67

Continúa en la siguiente Página

Cuadro XI.4: Estimación del parámetro Valor Diferenciador

Nro.	C. Niza	Valor Dif.	Reconocim.	Precisión	Efectividad
73	30	0.6	0.75	0.56	0.73
74	30	0.8	0.75	0.5	0.73
75	30	1	0.83	0.71	0.82
76	30	1.2	0.92	0.5	0.88
77	30	1.4	1	0.5	0.95
78	30	1.6	1	0.39	0.94
79	30	1.8	1	0.46	0.95
80	30	2	1	0.04	0.9
81	41	0.2	0.92	0.92	0.92
82	41	0.4	1	0.86	0.99
83	41	0.6	1	0.92	0.99
84	41	0.8	1	0.8	0.98
85	41	1	1	0.92	0.99
86	41	1.2	1	0.86	0.99
87	41	1.4	1	0.75	0.98
88	41	1.6	1	0.75	0.98
89	41	1.8	1	0.52	0.95
90	41	2	1	0.05	0.91
91	43	0.2	0.75	1	0.78
92	43	0.4	0.75	1	0.78
93	43	0.6	0.75	1	0.78
94	43	0.8	1	0.92	0.99
95	43	1	1	1	1
96	43	1.2	1	0.86	0.99

Continúa en la siguiente Página

Cuadro XI.4: Estimación del parámetro Valor Diferenciador

Nro.	C. Niza	Valor Dif.	Reconocim.	Precisión	Efectividad
97	43	1.4	1	1	1
98	43	1.6	1	0.92	0.99
99	43	1.8	1	0.86	0.99
100	43	2	1	0.08	0.91

Cuadro XI.4: Estimación del parámetro Valor Diferenciador

En la tabla 11.5 se puede ver el resumen del experimento para la obtención del Valor Diferenciador. Como podrán apreciar para el valor de 1.8 de éste parámetro de obtiene el mayor promedio y la menor varianza, estos son 0.9465 y 0.0007 respectivamente.

Valor Dif. .	Prom. REC	Prom. PREC	Prom. EFEC	Varianza REC	Varianza PREC	Varianza EFEC
0.2	0.7609	0.8831	0.7731	0.0064	0.016	0.0059
0.4	0.8019	0.8507	0.8068	0.0109	0.0195	0.0103
0.6	0.8269	0.8042	0.8246	0.0112	0.032	0.0107
0.8	0.8603	0.7694	0.8512	0.012	0.0197	0.0117
1	0.8846	0.7851	0.8747	0.0094	0.0306	0.0101
1.2	0.9096	0.7218	0.8908	0.0067	0.0213	0.0065
1.4	0.9179	0.7241	0.8986	0.0106	0.0244	0.0101
1.6	0.959	0.6144	0.9245	0.005	0.0343	0.0052
1.8	0.9917	0.54	0.9465	0.0007	0.0312	0.001
2	1	0.0637	0.9064	0	0.0012	0

Cuadro XI.5: Resumen estimación del parámetro Valor Diferenciador

XI.3.2. CONCLUSIONES DEL EXPERIMENTO

Se ha obtenido que el valor más óptimo para el Valor Diferenciador es de 1,8. Este valor será utilizado en la experimentación para obtener el valor de reconocimiento de la herramienta (Experimento 2).

XI.4. EXPERIMENTO 2

Para el experimento se ha tomado a en cuenta a las marcas de las clases NIZA con más de 7 marcas. Las cuales son las Clases: 5, 9, 25, 30, 32 y 41. Se han realizado 5 pruebas por cada clase utilizando k=8, donde k es el número de particiones a utilizarse (Ver Anexo 1).

XI.4.1. DESARROLLO DEL EXPERIMENTO

Se realizaron 30 pruebas para determinar el nivel de Reconociminto, precisión y efectividad de la herramienta. El resumen de los experimentos se muestran en la tabla 11.6 para los N-Grams de tamaño 4.

Medida	Reconocimiento	Precisión	Efectividad
Promedio	100 %	49,76 %	94,97 %
Varianza	0 %	3,95 %	0,03 %

Cuadro XI.6: Resumen del Experimento Tamaño de N-Gram 4

En la tabla 11.7 se muestran el resumen de los resultados.

Medida	Reconocimiento	Precisión	Efectividad
Promedio	100 %	47,61 %	94,76 %
Varianza	0 %	3,68 %	0,03 %

Cuadro XI.7: Resumen del Experimento Tamaño de N-Gram 5

En la tabla 11.8 se puede ver los resultados del experimento con tamaño de N-Gram 4.

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
1	1	4	1	0.72	0.97
2	1	4	1	0.55	0.95
3	1	4	1	0.39	0.94
4	1	4	1	0.56	0.96
5	1	4	1	0.53	0.95
6	3	4	1	0.64	0.96
7	3	4	1	0.39	0.94
8	3	4	1	0.53	0.95
9	3	4	1	0.47	0.95
10	3	4	1	0.44	0.94
11	3	4	1	0.8	0.98
12	3	4	1	0.5	0.95
13	3	4	1	0.5	0.95
14	3	4	1	0.35	0.93
15	3	4	1	0.64	0.96
16	3	4	1	0.57	0.96
17	3	4	1	0.67	0.97
18	3	4	1	0.52	0.95
19	3	4	1	0.43	0.94
20	3	4	1	0.46	0.95
21	3	4	1	0.6	0.96
22	3	4	1	0.43	0.94
23	3	4	1	0.56	0.96
24	3	4	1	0.79	0.98

Continúa en la siguiente Página

Cuadro XI.8: Pruebas realizadas Tamaño N-Gram 4

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
25	3	4	1	0.65	0.97
26	3	4	1	0.71	0.97
27	3	4	1	0.71	0.97
28	3	4	1	0.88	0.99
29	3	4	1	0.75	0.98
30	3	4	1	0.55	0.95
31	3	4	1	0.46	0.95
32	3	4	1	0.63	0.96
33	5	4	1	0.2	0.92
34	5	4	1	0.22	0.92
35	5	4	1	0.2	0.92
36	5	4	1	0.24	0.92
37	5	4	1	0.23	0.92
38	5	4	1	0.23	0.92
39	5	4	1	0.2	0.92
40	5	4	1	0.18	0.92
41	5	4	1	0.18	0.92
42	5	4	1	0.32	0.93
43	5	4	1	0.27	0.93
44	5	4	1	0.32	0.93
45	9	4	1	0.47	0.95
46	9	4	1	0.6	0.96
47	9	4	1	0.6	0.96
48	16	4	1	0.52	0.95

Continúa en la siguiente Página

Cuadro XI.8: Pruebas realizadas Tamaño N-Gram 4

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
49	16	4	1	0.26	0.93
50	16	4	1	0.36	0.94
51	16	4	1	0.33	0.93
52	16	4	1	0.35	0.93
53	16	4	1	0.42	0.94
54	16	4	1	0.35	0.93
55	16	4	1	0.44	0.94
56	16	4	1	0.36	0.94
57	16	4	1	0.35	0.93
58	16	4	1	0.48	0.95
59	16	4	1	0.71	0.97
60	16	4	1	0.4	0.94
61	16	4	1	0.41	0.94
62	16	4	1	0.55	0.95
63	16	4	1	0.67	0.97
64	29	4	1	0.49	0.95
65	29	4	1	0.41	0.94
66	29	4	1	0.49	0.95
67	29	4	1	0.35	0.94
68	29	4	1	0.43	0.94
69	29	4	1	0.45	0.95
70	29	4	1	0.55	0.95
71	29	4	1	0.86	0.99
72	29	4	1	0.69	0.97
Continúa en la siguiente Página					

Cuadro XI.8: Pruebas realizadas Tamaño N-Gram 4

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
73	29	4	1	0.23	0.92
74	29	4	1	0.26	0.93
75	29	4	1	0.23	0.92
76	29	4	1	0.31	0.93
77	29	4	1	0.26	0.93
78	29	4	1	0.26	0.93
79	29	4	1	0.32	0.93
80	29	4	1	0.31	0.93
81	29	4	1	0.29	0.93
82	29	4	1	0.29	0.93
83	29	4	1	0.46	0.95
84	29	4	1	0.32	0.93
85	29	4	1	0.46	0.95
86	29	4	1	0.36	0.94
87	29	4	1	0.55	0.95
88	29	4	1	0.46	0.95
89	29	4	1	0.38	0.94
90	29	4	1	0.3	0.93
91	29	4	1	0.27	0.93
92	29	4	1	0.28	0.93
93	29	4	1	0.52	0.95
94	43	4	1	0.36	0.94
95	43	4	1	0.35	0.93
96	43	4	1	0.3	0.93

Continúa en la siguiente Página

Cuadro XI.8: Pruebas realizadas Tamaño N-Gram 4

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
97	43	4	1	0.3	0.93
98	43	4	1	0.29	0.93
99	43	4	1	0.31	0.93
100	43	4	1	0.28	0.93
101	43	4	1	0.26	0.93
102	43	4	1	0.27	0.93
103	43	4	1	0.64	0.96
104	43	4	1	0.49	0.95
105	43	4	1	0.51	0.95
106	43	4	1	0.62	0.96
107	43	4	1	0.53	0.95
108	43	4	1	0.85	0.99
109	43	4	1	0.64	0.96
110	43	4	1	0.72	0.97
111	43	4	1	0.68	0.97
112	43	4	1	0.64	0.96
113	43	4	1	0.88	0.99
114	43	4	1	1.05	1
115	43	4	1	0.96	1
116	43	4	1	0.92	0.99
117	43	4	1	0.92	0.99
118	43	4	1	0.56	0.96
119	43	4	1	0.56	0.96
120	43	4	1	0.41	0.94

Continúa en la siguiente Página

Cuadro XI.8: Pruebas realizadas Tamaño N-Gram 4

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
121	43	4	1	0.44	0.94
122	43	4	1	0.61	0.96
123	43	4	1	0.61	0.96

Cuadro XI.8: Pruebas realizadas Tamaño N-Gram 4

En la tabla 11.9 se puede ver los resultados del experimento con tamaño de N-Gram 5.

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
1	1	4	1	0.72	0.97
2	1	4	1	0.55	0.95
3	1	4	1	0.39	0.94
4	1	4	1	0.56	0.96
5	1	4	1	0.53	0.95
6	3	4	1	0.64	0.96
7	3	4	1	0.39	0.94
8	3	4	1	0.53	0.95
9	3	4	1	0.47	0.95
10	3	4	1	0.44	0.94
11	3	4	1	0.8	0.98
12	3	4	1	0.5	0.95
13	3	4	1	0.5	0.95
14	3	4	1	0.35	0.93
15	3	4	1	0.64	0.96
16	3	4	1	0.57	0.96

Continúa en la siguiente Página

Cuadro XI.9: Pruebas realizadas Tamaño N-Gram 5

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
17	3	4	1	0.67	0.97
18	3	4	1	0.52	0.95
19	3	4	1	0.43	0.94
20	3	4	1	0.46	0.95
21	3	4	1	0.6	0.96
22	3	4	1	0.43	0.94
23	3	4	1	0.56	0.96
24	3	4	1	0.79	0.98
25	3	4	1	0.65	0.97
26	3	4	1	0.71	0.97
27	3	4	1	0.71	0.97
28	3	4	1	0.88	0.99
29	3	4	1	0.75	0.98
30	3	4	1	0.55	0.95
31	3	4	1	0.46	0.95
32	3	4	1	0.63	0.96
33	5	4	1	0.2	0.92
34	5	4	1	0.22	0.92
35	5	4	1	0.2	0.92
36	5	4	1	0.24	0.92
37	5	4	1	0.23	0.92
38	5	4	1	0.23	0.92
39	5	4	1	0.2	0.92
40	5	4	1	0.18	0.92

Continúa en la siguiente Página

Cuadro XI.9: Pruebas realizadas Tamaño N-Gram 5

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
41	5	4	1	0.18	0.92
42	5	4	1	0.32	0.93
43	5	4	1	0.27	0.93
44	5	4	1	0.32	0.93
45	9	4	1	0.47	0.95
46	9	4	1	0.6	0.96
47	9	4	1	0.6	0.96
48	16	4	1	0.52	0.95
49	16	4	1	0.26	0.93
50	16	4	1	0.36	0.94
51	16	4	1	0.33	0.93
52	16	4	1	0.35	0.93
53	16	4	1	0.42	0.94
54	16	4	1	0.35	0.93
55	16	4	1	0.44	0.94
56	16	4	1	0.36	0.94
57	16	4	1	0.35	0.93
58	16	4	1	0.48	0.95
59	16	4	1	0.71	0.97
60	16	4	1	0.4	0.94
61	16	4	1	0.41	0.94
62	16	4	1	0.55	0.95
63	16	4	1	0.67	0.97
64	29	4	1	0.49	0.95
Continúa en la siguiente Página					

Cuadro XI.9: Pruebas realizadas Tamaño N-Gram 5

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
65	29	4	1	0.41	0.94
66	29	4	1	0.49	0.95
67	29	4	1	0.35	0.94
68	29	4	1	0.43	0.94
69	29	4	1	0.45	0.95
70	29	4	1	0.55	0.95
71	29	4	1	0.86	0.99
72	29	4	1	0.69	0.97
73	29	4	1	0.23	0.92
74	29	4	1	0.26	0.93
75	29	4	1	0.23	0.92
76	29	4	1	0.31	0.93
77	29	4	1	0.26	0.93
78	29	4	1	0.26	0.93
79	29	4	1	0.32	0.93
80	29	4	1	0.31	0.93
81	29	4	1	0.29	0.93
82	29	4	1	0.29	0.93
83	29	4	1	0.46	0.95
84	29	4	1	0.32	0.93
85	29	4	1	0.46	0.95
86	29	4	1	0.36	0.94
87	29	4	1	0.55	0.95
88	29	4	1	0.46	0.95
Continúa en la siguiente Página					

Cuadro XI.9: Pruebas realizadas Tamaño N-Gram 5

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
89	29	4	1	0.38	0.94
90	29	4	1	0.3	0.93
91	29	4	1	0.27	0.93
92	29	4	1	0.28	0.93
93	29	4	1	0.52	0.95
94	43	4	1	0.36	0.94
95	43	4	1	0.35	0.93
96	43	4	1	0.3	0.93
97	43	4	1	0.3	0.93
98	43	4	1	0.29	0.93
99	43	4	1	0.31	0.93
100	43	4	1	0.28	0.93
101	43	4	1	0.26	0.93
102	43	4	1	0.27	0.93
103	43	4	1	0.64	0.96
104	43	4	1	0.49	0.95
105	43	4	1	0.51	0.95
106	43	4	1	0.62	0.96
107	43	4	1	0.53	0.95
108	43	4	1	0.85	0.99
109	43	4	1	0.64	0.96
110	43	4	1	0.72	0.97
111	43	4	1	0.68	0.97
112	43	4	1	0.64	0.96

Continúa en la siguiente Página

Cuadro XI.9: Pruebas realizadas Tamaño N-Gram 5

Nro.	C. Niza	Tam. N-Gram	Reconocim.	Precisión	Efectividad
113	43	4	1	0.88	0.99
114	43	4	1	1.05	1
115	43	4	1	0.96	1
116	43	4	1	0.92	0.99
117	43	4	1	0.92	0.99
118	43	4	1	0.56	0.96
119	43	4	1	0.56	0.96
120	43	4	1	0.41	0.94
121	43	4	1	0.44	0.94
122	43	4	1	0.61	0.96
123	43	4	1	0.61	0.96

Cuadro XI.9: Pruebas realizadas Tamaño N-Gram 5

XI.4.2. CONCLUSIONES DEL EXPERIMENTO

De este primer experimento se puede concluir que los resultados son más óptimos utilizando el tamaño de los N-Grams igual a 4 y se pude alcanzar un alto nivel de reconocimiento (100 %) Pero obtenemos también que el nivel de efectividad disminuye debido a que se tiene un nivel de precisión de 49.76 % obteniendo finalmente un nivel de efectividad del 94.97 %

XI.5. ANÁLISIS DE SENSIBILIDAD

El objetivo de éste análisis es identificar cuan sensibles son los resultados ante variaciones en los valores que puedan tomar los parámetros de ingreso utilizados en la Red Neuronal. Para esto se realizará dicho análisis para los 3

parámetros de ingreso utilizados. Por cuestiones didácticas el valor de similitud que antes era presentado en el rango de 0-2 teniendo a 0 como similares y 2 como diferentes será presentado de forma inversa, es decir si las marcas son similares el valor de similitud será de 2 y si son distintas el valor será de 0. El valor de cada parámetro será disminuido en 0.01 y se realizará la corrida respectiva a fin de obtener el valor de similitud correspondiente para luego identificar su variación.

XI.5.1. ANÁLISIS DE SENSIBILIDAD PARÁMETRO 1

Se realizó el análisis de sensibilidad para el parámetro 1 el cual está determinado por la razón de la cantidad de n-grams de tamaño 2 similares entre las dos marcas a comparar y el número máximo de combinaciones posibles para las comparaciones entre dichos n-grams. Como resultado se obtuvo el gráfico mostrado en la figura XI.2, se tiene que el valor de similitud se mantiene constante para todo valor del parámetro 1 mayor que 0.38, se experimenta un incremento en el rango 0.15 a 0.38, para todo valor del parámetro 1 menor que 0.15 el valor de similitud disminuye de forma acelerada. Esto nos quiere decir que para valores del parámetro 1 menores a 0.15 el valor de similitud es muy sensible a los cambios en el valor de dicho parámetro.

XI.5.2. ANÁLISIS DE SENSIBILIDAD PARÁMETRO 2

Se realizó el análisis de sensibilidad para el parámetro 2 el cual está determinado por la razón de la cantidad de n-grams de tamaño 3 similares entre las dos marcas a comparar y el número máximo de combinaciones posibles para las comparaciones entre dichos n-grams. Como resultado se obtuvo el gráfico mostrado en la figura XI.3, se tiene que el valor de similitud se mantiene constante para todo valor del parámetro 2 mayor que 0.15, a partir de este

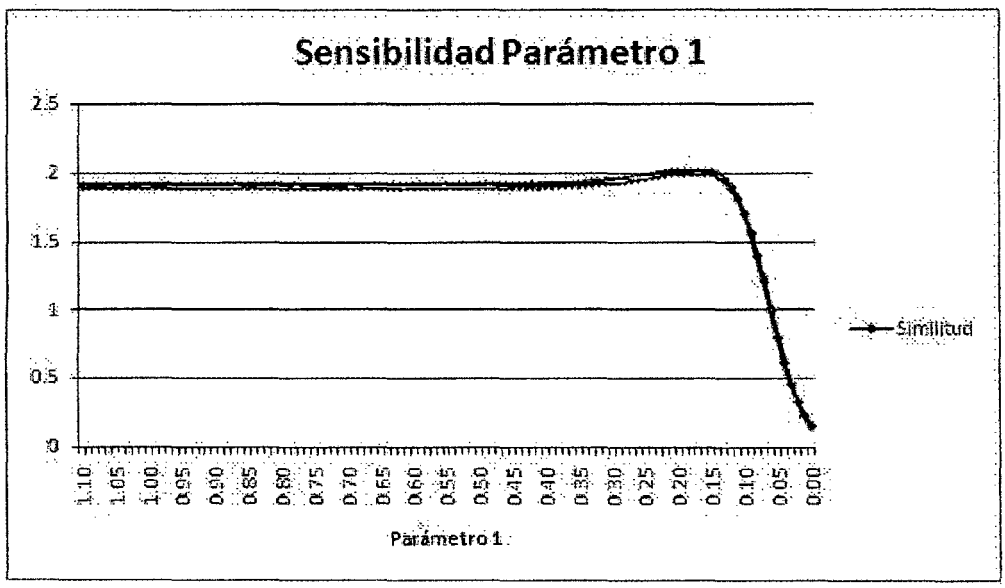


Figura XI.2: Análisis de Sensibilidad Parámetro 1

valor del parámetro 2 el valor de similitud disminuye de forma acelerada. Esto nos quiere decir que para valores del parámetro 2 menores a 0.15 el valor de similitud es muy sensible a los cambios en el valor de dicho parámetro.

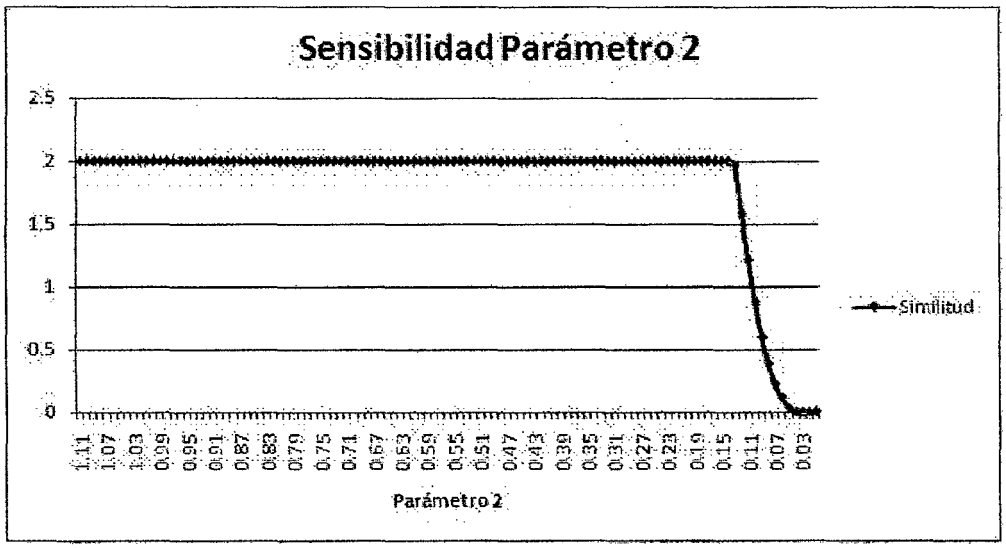


Figura XI.3: Análisis de Sensibilidad Parámetro 2

XI.5.3. ANÁLISIS DE SENSIBILIDAD PARÁMETRO 3

Se realizó el análisis de sensibilidad para el parámetro 3 el cual está determinado por la razón de la cantidad de n-grams de tamaño 4 similares entre las dos marcas a comparar y el número máximo de combinaciones posibles para las comparaciones entre dichos n-grams. Como resultado se obtuvo el gráfico mostrado en la figura XI.4, se tiene que el valor de similitud se mantiene constante para todo valor del parámetro 3 mayor que 0.19, a partir de este valor del parámetro 3 el valor de similitud disminuye de forma acelerada. Esto nos quiere decir que para valores del parámetro 3 menores a 0.19 el valor de similitud es muy sensible a los cambios en el valor de dicho parámetro.

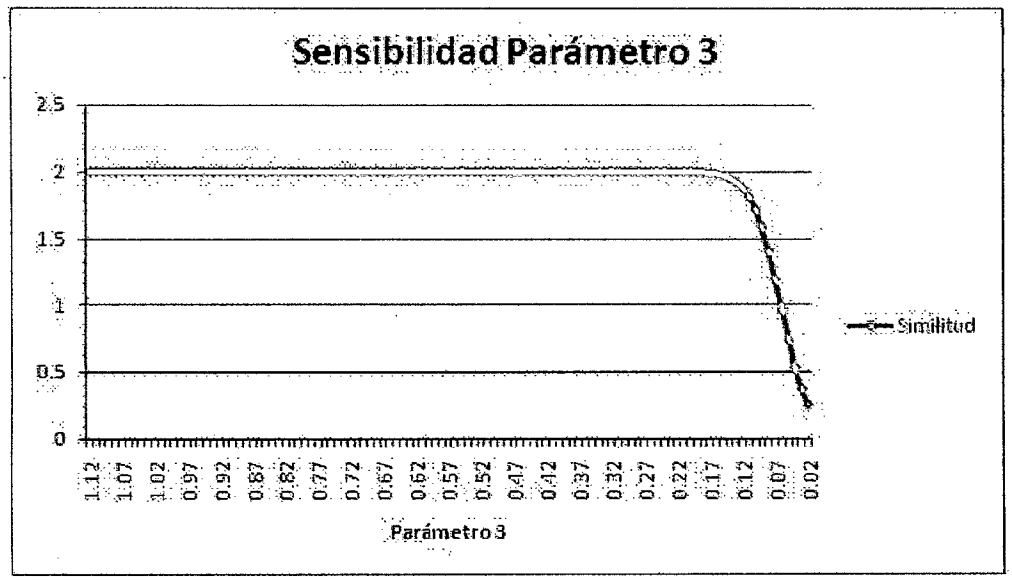


Figura XI.4: Análisis de Sensibilidad Parámetro 3

Del análisis de sensibilidad se pudo obtener que si bien es cierto para un rango amplio de los valores de los parámetros de ingreso el valor de similitud se mantiene casi constante, existe un rango menor (valores menores que 0.15, 0.15 y 0.19 para los parámetros 1, 2 y 3 respectivamente) donde el valor de similitud es muy sensible a la variación de los valores de los parámetros y disminuye notablemente. Para estudios posteriores podría plantearse como

propuesta la utilización de los resultados obtenidos del análisis de sensibilidad como parámetro de ingreso en la red neuronal a fin de mejorar los valores de reconocimiento y precisión.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. La eficacia del proceso manual para la identificacion de marcas similares tiene un nivel de tan solo 89 % (Ver cuadro 2.2).

2. El procentaje de marcas solicitadas que son similares a las ya registradas es de 23.70 % (Ver figura 5.1).

3. En promedio se tratan alrededor de 419 casos de denuncias por infracciones de similitud fonética entre marcas (Ver cuadro 2.2).

4. El tratamiento de estos casos deviene en costos de mas de S/131 000 soles anuales (Ver cuadro 2.2).

5. En el estudio se tiene que dar mayor preponderancia al nivel de reconocimiento sobre el nivel de precisión, esto va a permitir que el mayor número de marcas similares sean identificadas.

6. Luego de realizar el primer experimento se ha obtenido que la herramienta tiene un nivel de reconocimiento del 100 % por lo que hasta el momento podríamos decir que se ha demostrado la hipótesis.

7. Se obtienen mejores resultados utilizando un tamaño de N-Gram igual

a 4, ya que los valores de efectividad obtenidos fueron de 94.97% y 94.76% para los tamaños de N-Gram 4 y 5 respectivamente.

8. A través del análisis de sensibilidad se obtuvo que los valores de similitud son muy sensibles para los rangos menores que 0.15, 0.15 y 0.19 de los parámetros de ingreso de la Red Neuronal 1, 2 y 3 respectivamente, para valores mayores se cumple que el valor de similitud se mantiene constante y cercano a 2.

RECOMENDACIONES

1. Se recomienda hacer un análisis más profundo y buscar la forma de aplicar la herramienta para marcas que se encuentren en otro idioma.

2. Se recomienda utilizar la herramienta en conjunto con otra herramienta que determine el ivel de similitud entre las imágenes que se utilizan en la Marca Comercial.

3. Se recomienda al momento de utilizar la herramienta realizar varias pruebas utilizando el esquema de pruebas K-Fold.

4. Se recomienda utilizar la herramienta inicialmente en conjunto con los procedimientos normales a fin de determinan los niveles de Reconocimiento y Precisión con valores más reales.

5. Se recomienda para estudios posteriores la utilización de los resultados obtenidos del análisis de sensibilidad como parámetro de ingreso en la red neuronal a fin de mejorar los valores de reconocimiento y precisión.

ANEXOS

ANEXO1: ESQUEMA DE EXPERIMENTACIÓN

Se utilizó como Esquema de Experimentación a k-Fold el cual consiste en la creación de k particiones del conjunto de datos. Para cada uno de los k experimentos, k-1 particiones se usan para entrenamiento y una para prueba. Si contásemos con 1200 datos de prueba y utilizaremos k igual a 12 entonces de los 1200, 1100 datos serían utilizados para realizar el entrenamiento de la red neuronal y 100 para las pruebas.

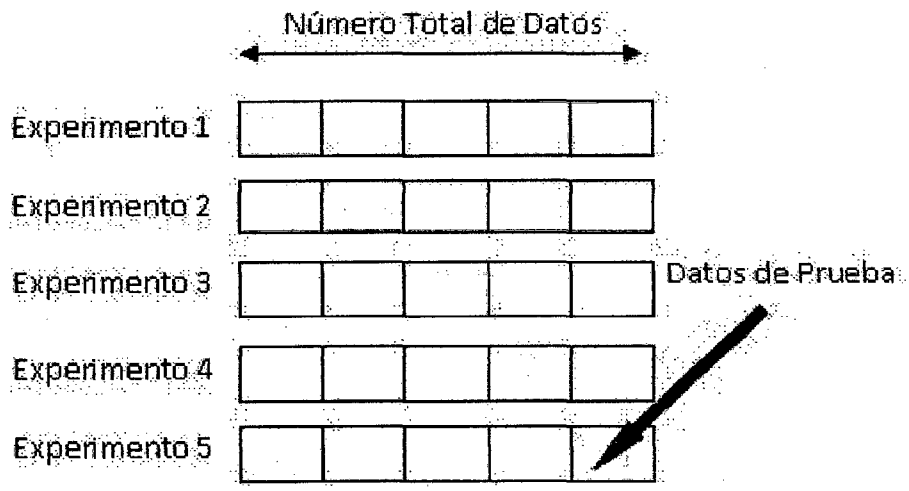


Figura XI.5: Experimentación utilizando K-Fold

ANEXO2: NÚMERO ÉPOCAS DE ENTRENAMIENTO A UTILIZARSE

El número de épocas a utilizarse en un entrenamiento viene a ser el número de iteraciones a realizarse durante el mismo. A medida que se van realizando más iteraciones se van ajustando más los pesos y se reduce el error hasta alcanzar el número máximo de iteraciones establecido o alcanzar el objetivo que es obtener el error mínimo planteado. Durante los distintos experimentos realizados el número de épocas utilizado no sobrepasó las 23 iteraciones, es decir se obtuvo el error mínimo requerido antes de realizar el número de iteraciones establecidas (500 épocas), el número promedio de iteraciones obtenido fue de 5.6764. Por tal motivo se decidió reducir el número de épocas utilizado a 100, de esta forma se reduciría el número de épocas a utilizar, lo cual para algunos casos reduciría el tiempo de procesamiento de la herramienta y al mismo tiempo mantendría los niveles de reconocimiento y eficacia de la misma. En conclusión se utilizarán 100 épocas para el entrenamiento de la Red Neuronal.

```
TRAINLM-calcjx, Epoch 0/100, MSE 4.02138/0.0001, Gradient 6.58026/1e-010
TRAINLM-calcjx, Epoch 2/100, MSE 9.01401e-005/0.0001, Gradient 0.0317737/1e-010
TRAINLM, Performance goal met.

TRAINLM-calcjx, Epoch 0/100, MSE 1.45643/0.0001, Gradient 4.10518/1e-010
TRAINLM-calcjx, Epoch 10/100, MSE 0.00034972/0.0001, Gradient 0.00580691/1e-010
TRAINLM-calcjx, Epoch 16/100, MSE 9.4332le-005/0.0001, Gradient 0.0293332/1e-010
TRAINLM, Performance goal met.

TRAINLM-calcjx, Epoch 0/100, MSE 7.70237/0.0001, Gradient 8.40223/1e-010
TRAINLM-calcjx, Epoch 3/100, MSE 2.58462e-005/0.0001, Gradient 0.0160042/1e-010
TRAINLM, Performance goal met.
```

Figura XI.6: Épocas del Experimento

GLOSARIO DE TÉRMINOS

1. Clasificación Niza: La Clasificación de Niza es una clasificación de los productos y servicios para el registro de las marcas de fábrica o de comercio y las marcas de servicios, esta clasificación asocia a ls productos de acuerdo a su parecido y uso.

2. Codificación IPA: Es la clasificación de los fonemas basada en Transcripción Fonética IPA.

3. Fonema: Son abstracciones mentales o abstracciones formales de los sonidos del habla. En este sentido, un fonema puede ser representado por una familia o clase de equivalencia de sonidos (técnicamente denominados fonos), que los hablantes asocian a un sonido específico durante la producción o la percepción del habla.

4. Marca Comercial: una Marca es un nombre, término, símbolo, diseño o combinación de éstos elementos que identifica los productos de un proveedor y los distingue de los productos de la competencia.

5. N-Gram: Un N-Gram no es más que una subcadena de una palabra de un tamaño menor al de la palabra, ha sido utilizada para la identificación de similitud entre nombres.

6. Nivel de Precisión: Es la medida utilizada para medir el nivel en que una herramienta identifica las palabras similares adecuadas, es decir mientras menos palabras no similares sean identificadas mayor será el valor de esta medida.

7. Nivel de Reconocimiento: Es la medida utilizada para medir el nivel en que un herramienta identifica palabras similares, es decir mientras más palabras similares sean identificadas mayor será el valor de esta medida.

8. Red Neuronal: Las redes neuronales artificiales (RNA) son modelos que intentan reproducir el comportamiento del cerebro. Como tal modelo, realiza una simplificación, averiguando cuáles son los elementos relevantes del sistema, bien porque la cantidad de información de que se dispone es excesiva o bien porque es redundante. Una elección adecuada de sus características, más una estructura conveniente, es el procedimiento convencional utilizado para construir redes capaces de realizar determinada tarea, las Redes Neuronales pueden ser utilizadas para realizar Clasificación y Pronóstico.

9. Similitud Fonética: Es la similitud establecida por el parecido existente entre dos palabras al momento de ser pronunciadas, en el ambiente de las marcas comerciales esta Similitud Fonética podría producir en los consumidores una idea errónea de parecido entre las marcas comerciales que fuesen fonéticamente similares.

10. Token: Un Token viene a ser una palabra que forma parte de una Marca Comercial, para el experimento se utilizan aquellos Tokens que tienen un valor significativo, no tomándose entonces aquellos Tokens que sean palabras que sirven como conexión lingüística.

BIBLIOGRAFÍA

[1] C J Fall, C. Giraud-Carrier. Searching trademark databases for verbal similarities, World Patent Information 27 (2005) 135-143.

[2] Bugdahl V. Markenrecherchen - eine subjective Momen - taufnahme. MarkenR 2003;05/2003:169-80

[3] Zobel J, Dart P. Finding approximate matches in large lexicons. Software-Practice Experience 1995;25:331-45

[4] E. Ukkonen, 'Approximate string-matching with q-grams and maximal matches', Theoretical Computer Science, 92, 191-211 (1992).

[5] World Intellectual Property Organization, Understanding Industrial Property: WIPO Publication No. 895(E) ISBN 92-805-1257-9

[6] Holmes D, McCabe MC. Improving precision and recall for soundex retrieval. Proceedings of the International Conference on Information Technology: Coding and Computing. Las Vegas, Nevada: IEEE Computer Society; 2002. p. 22-7

[7] Fischer I, Zell A. String averages and self-organizing maps for strings. In: Proceedings of the Second ICSC Symposium on Neural

Computation-NCÕ2000. Canada/Switzerland: ICSC Academic Press; 2000. p. 208-15

[8] Wu S, Manber U. Fast text searching allowing errors. Común ACM 1992;35:83-91.

[9] Zobel J, Dart P. Phonetic string matching: lessons from information retrieval. In: Frei H-P, Harman D, Schäble P, Wilkinson R, editors. Proceedings of the 19th International Conference on Research and Development in Information Retrieval. Zurich, Switzerland: ACM Press; 1996. p. 166-72

[10] Brodda B. Corpus work with PC beta: a presentation. Proceedings of the 13th conference on computational linguistics, vol. 3. Mor- ristown, NJ: Association for Computational Linguistics; 1990. p. 405-9

[11] McAllister R, Brodda B. Development of a new speech comprehension test with a phonological distance metric. Proceedings of Fonetik 2002, the XVth Swedish Phonetics Conference, Fysik-centrum, Stockholm, May 29-31, 2002, vol. 44. Stockholm, Sweden: KTH; 2002. p. 149-51.

[12] Erikson K. Approximate swedish name matching survey and test of different algorithms. MasterÕs thesis, Department of Numerical Analysis and Computing Science, KTH, Royal Insti- tute of Technology, Nada, S-100 44 Stockholm, Sweden, 1997

[13] Hodge VJ, Austin J. An evaluation of phonetic spell checkers, 2001.

[14] Teuvo Kohonen. An Introduction to Neural Computing, Helsinki University of Technology, Neural Networks. Vol. 1. pp. 3-16. 1988

[15] Kukich, Karen. Techniques for Automatically Correcting Words in: ACM Computing Surveys, Vol 24(4) Dec. 1992, pp 377-439

[16] Kukich, Karen. A Comparison of Some Novel and Traditional Lexical Distance Metrics for Spelling Correction. In: Proceedings of INNC-90-Paris, Paris, France, July 1990, pp 309-313