INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
LABORATORIO DE PROCESAMIENTO DE LENGUAJE NATURAL

Tesis

# DETECCIÓN AUTOMÁTICA DE HUMOR EN TEXTOS CORTOS EN ESPAÑOL

que presenta el

**Ing. Rigoberto Ocampo Pólito**

Para obtener el grado de:

**Maestro en Ciencias de la Computación**

Director de Tesis:

**Dr. Alexander Gelbukh**

México, D.F., Junio de 2010

SIP-14

# INSTITUTO POLITECNICO NACIONAL

## SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 12:00 horas del día 7 del mes de junio de 2010 se reunieron los miembros de la Comisión Revisora de Tesis designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

***Centro de Investigación en Computación***

para examinar la tesis de grado titulada:

**"DETECCIÓN AUTOMÁTICA DE HUMOR EN TEXTOS CORTOS EN ESPAÑOL"**

| **OCAMPO** | **PÓLITO** | **RIGOBERTO** |
|---|---|---|
| Apellido paterno | materno | nombre(s) |

Con registro:

| B | 0 | 8 | 1 | 4 | 2 | 7 |
|---|---|---|---|---|---|---|

aspirante al grado de:

**MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

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

**LA COMISIÓN REVISORA**

Presidente

Dr. Grigori Sidorov

Secretario

Dr. Marco Antonio Moreno Ibarra

Primer vocal
(Director de Tesis)

Dr. Alexandre Felixovich Guelboukh Kahn

Segundo vocal

Dr. René Arnulfo García Hernández

Tercer vocal

Dr. Héctor Jiménez Salazar

Suplente

Dr. Miguel Jesús Torres Ruiz

EL PRESIDENTE DEL COLEGIO

M. en C. Sergio Sandoval Reyes

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION
EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

# CARTA CESIÓN DE DERECHOS

En la ciudad de México, D. F., el día lunes 31 de mayo del año 2010, el que suscribe **C. Rigoberto Ocampo Pólito,** alumno del programa de **Maestría en Ciencias de la Computación,** con número de registro **B081427,** adscrito al Centro de Investigación en Computación, del Instituto Politécnico Nacional, manifiesta ser autor intelectual del presente trabajo de Tesis, bajo la dirección del doctor **Alexander Gelbukh,** y cede los derechos del trabajo intitulado ***"Detección Automática de Humor en Textos Cortos en Español",*** al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo, sin el permiso expreso del autor y/o director de la tesis. Éste puede ser obtenido escribiendo a la siguiente dirección de correo electrónico: pocampob08@sagitario.cic.ipn.mx. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

***Rigoberto Ocampo Pólito***

# Resumen

A pesar de que el humor ha sido estudiado por ciencias como la filosofía, la psicología, la sociología, etc., desde la época de los grandes pensadores griegos; no es sino hasta la aparición de la computación como herramienta de apoyo a las ciencias, que se ha hecho un paréntesis en la especulación de teorías humorísticas que satisfagan integralmente a las diferentes ramas del saber, para explorar el humor desde el punto de vista estadístico, en búsqueda de características que nos ayuden a comprenderlo mejor y de esa manera, quizás, encontrar una verdadera teoría integral acerca del humor.

Sin embargo, todas esas nuevas exploraciones se han realizado para el idioma inglés, lo cual deja fuera de toda posibilidad el análisis y la detección de expresiones humorísticas como el albur, que es una expresión propia del idioma español, y específicamente propia de los mexicanos.

En este trabajo de tesis se detectan características del humor como: *la rima, la aliteración, el albur* y *el contenido adulto,* en textos cortos (chistes y dichos) en el idioma español y se analizan a través de diversos algoritmos de clasificación contenidos en la aplicación WEKA, en búsqueda de los que mejor resultados presenten para la detección del humor en textos cortos en dicho idioma.

# Abstract

Although humor has been studied since the time of the great Greek thinkers and later by philosophy, psychology, and sociology, among other disciplines, existing theories of humor are still speculative and there is no theory accepted by the experts in all those disciplines. However, now that we have computers as a tool to support science it is a promising option to stop thinking of those theories for a while and first collect hard statistical data, in search of features that could help us better understand this phenomenon. Later this may potentially contribute to the creation of a real comprehensive theory of humor.

However, so far all statistical studies of humor we are aware about have been applied to English. This leaves out some relevant phenomena such as the word play typical for Mexican verbal humor (*albur* in Spanish).

In this thesis we explore such features of humor as rhyme, alliteration, word play (*albur*), and adult contents (including slang) in short texts such as Mexican jokes. We contrast such humorous short texts with Spanish proverbs. We use various classification algorithms implemented in WEKA and determine which one gives the best results on such data.

# Agradecimientos

A mi familia, por compartir junto
conmigo los sacrificios que conllevan
el estudiar una maestría.

A mi padre, por su apoyo celestial.

A mi asesor, Dr. Alexander Gelbukh,
por su infinita paciencia y dedicación.

A mi honorable comité tutorial:
Dr. Héctor Jiménez Salazar,
Dr. René Arnulfo García Hernández,
Dr. Miguel Jesús Torres Ruíz,
Dr. Marco Antonio Moreno Ibarra,
Dr. Grigori Sidorov,
por sus valiosas críticas en la
construcción de esta tesis.

Al Centro de Investigación en Computación - IPN,
por la formidable oportunidad que me brindó
para crecer profesionalmente.

A mi padre Dios, quien nos bendijo
a todos durante este tiempo

# Índice General

# Índice de ilustraciones

# Índice de Tablas

# Capítulo 1. Introducción

En este capítulo se da a conocer de manera detallada las razones principales que motivaron la realización de este trabajo de tesis, sus propósitos; así como las aportaciones del mismo.

## *1.1 Problemática actual*

A pesar de que han pasado ya miles de años desde que el hombre comenzó a tratar de entender el fenómeno humano del humor y de que a lo largo de esa historia diversas ciencias han colaborado en tratar de descifrar este enigma, no se ha logrado establecer una única teoría del humor que sea verdaderamente general; es decir, que sea tomada como una teoría irrefutable y que satisfaga de manera integral a todas la ramas del saber. Las razones son, al menos en esto sí coinciden todas, que el humor es algo tan diverso (en las formas en que se origina, en las formas en que se presenta y en las formas en que se interpreta), que es imposible definirlo en una sola teoría.

Dada esta problemática, y fuera del contexto humanístico-social de las ciencias que más han aportado en el estudio del humor, la computación se presenta como una herramienta muy innovadora de apoyo a estas ciencias, en particular a la lingüística, que propone darle un nuevo matiz al estudio del humor buscando encontrar los

rasgos más distintivos que se presentan en las situaciones humorísticas (delimitadas, claro está, al humor escrito), con el propósito de ofrecer una nueva visión en la búsqueda de la tan soñada Teoría general del humor.

## *1.2 Planteamiento del problema*

Hoy en día existen ya diversas herramientas que intentan detectar el humor escrito, e inclusive generarlo; sin embargo, estas herramientas son muy poco eficientes dado que se especializan en un solo tipo de humor (cabe hacer mención, que el humor escrito se presenta con diversas variantes); además de que no existe una herramienta que detecte textos humorísticos para el idioma español, y por consiguiente, que detecte una de las expresiones humorísticas características de los mexicanos, el albur.

El presente trabajo de tesis pretende subsanar la falta de una herramienta para la detección del humor en el idioma español; así como para la expresión humorística característica de los mexicanos, el albur. Trabajo que no será una tarea sencilla, dado que el albur, como todas las expresiones humorísticas, es una expresión bastante rica - lingüísticamente hablando – pero que será un buen intento por avanzar en materia de estudio y comprensión de la cualidad humana del humor.

## 1.3 Hipótesis

Es posible detectar textos humorísticos en español con los atributos de *rima, aliteración* y *contenido adulto,* propuestos por Rada Mihalcea[20] para el idioma inglés, y por el atributo *palabraalbureable* para el albur, utilizando como herramienta el programa WEKA.

## 1.4 Objetivo General

Detectar algunos atributos característicos del humor como: *rima, aliteración, albur* y *contenido adulto*, en los textos cortos en español: chistes y dichos, para utilizar los algoritmos de la aplicación WEKA a fin de determinar cuál o cuáles son los mejores para la detección de humor en textos cortos de dicho idioma.

## 1.5 Objetivos Particulares

- Crear diccionarios de términos de los atributos humorísticos *albur* y *contenido adulto*.
- Crear un corpus con textos humorísticos en español.
- Crear un corpus con textos no humorísticos (dichos) en español.
- Diseñar un programa de procesamiento de textos que sea capaz de:
    - Detectar los atributos *rima, aliteración, albur* y *contenido adulto*, presentes en los textos.

- Establecer un vector de valores con dichos atributos para el análisis de los textos.
- Crear un archivo .*arff* para poder hacer uso de la aplicación WEKA.

- Experimentar con diversos algoritmos de clasificación de WEKA para determinar cuál o cuáles son los mejores para la detección de humor en textos cortos.

## *1.6 Aportaciones*

Las principales aportaciones de este trabajo de tesis son las siguientes:

Una herramienta de procesamiento de textos en español que obtiene los rasgos humorísticos *rima, aliteración, albur* y *contenido adulto,* a fin de ser procesados y analizados a través del programa WEKA.

La implementación de detección del albur, el cual es un lenguaje muy característico de algunos círculos sociales de los mexicanos, que en ocasiones resulta humorístico.

Un análisis de algoritmos de WEKA para determinar cuál es el mejor para la detección de humor en textos cortos en español.

La confirmación que para el español son similares los rasgos distintivos del humor a los encontrados por Rada Mihalcea[6], para el idioma inglés.

## *1.7 Estructura del documento*

El resto del documento se encuentra organizado de la siguiente manera:

*Capítulo 2. Estado del arte.* Ofrece un análisis de las herramientas existentes actualmente para la detección del humor escrito; así como de los algoritmos utilizados para tal fin.

*Capítulo 3. Marco teórico.* Presenta los conceptos fundamentales para la comprensión del tema desarrollado en el presente trabajo de tesis. Si el lector conoce de estos tópicos, puede omitir leer este capítulo y pasarse directamente al capítulo 4.

*Capítulo 4. Formación de los rasgos.* Muestra una detallada descripción de los pasos previos necesarios para realizar la detección del humor en los textos cortos en español.

*Capítulo 5. Experimentos y resultados.* Presenta los experimentos llevados a cabo; así como los resultados obtenidos y la forma en que se evaluaron y compararon.

*Capítulo 6. Conclusiones.* Finalmente en este capítulo se presentan las conclusiones a las que se llegaron en este trabajo de tesis; así como el trabajo futuro que este tema merece.

# Capítulo 2. Estado del Arte

En este capítulo se describen de manera breve las teorías y los trabajos más importantes y más recientes realizados con respecto a la detección de humor en textos humorísticos.

## *2.1 Ubicación*

El humor es una cualidad humana que sin duda ha intrigado a los hombres de todos los tiempos; al menos desde los tiempos de Aristóteles, quien según filósofos contemporáneos, fue el único filósofo clásico que trató acerca de la comedia[1]; aunque ellos sugieren que fue así, dado que Aristóteles hace referencia a una supuesta Teoría del Humor en su obra *Ars Poética;* que hoy en día consideran incompleta y de la cual piensan que constaba de dos partes, la segunda de ellas precisamente llamada *Ars Comica* [2].

Pero, de entrada, el humor presenta una dificultad muy importante, que es: el ser subjetivo; es decir, que no por todos es advertido y que es percibido con diferentes reacciones por cada uno de nosotros, dado que para ello influyen en demasía factores como la cultura, la religión, el estado de ánimo y el conocimiento que se tenga acerca del mundo (*véase Ilustración 1*).

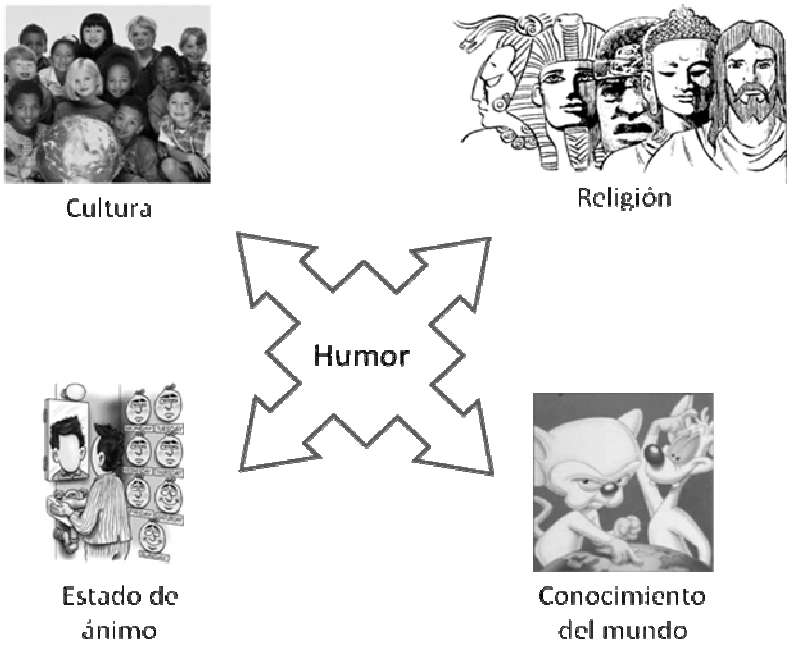


***Ilustración 1. Factores que influyen en la identificación del humor.***

El tratar de entender el humor ha llevado a los estudiosos de las distintas áreas de la ciencia como: la filosofía, la psicología, la lingüística, la sociología, etc., a realizar diversos análisis de las condiciones en que éste se da. Los resultados, son la formulación de numerosas teorías y opiniones que intentan establecer de manera inequívoca los factores que intervienen en el fenómeno del humor y por consiguiente del por qué reímos; aunque ninguna de ellas se ha adoptado como la mejor o la más acertada.

A pesar de tal diversidad, se ha convenido clasificar todas estas aseveraciones en sólo tres teorías del humor: Teoría de la descarga, Teoría de la superioridad y Teoría de la incongruencia las cuales se conocerán en el siguiente capítulo.

## *2.2 Teorías del humor*

A continuación se describen, de manera muy breve, los tipos de teorías del humor que imperan en la actualidad. Cabe señalar que ninguna de estas teorías ha sido particularmente relevante para la metodología de este trabajo de tesis, dado que son teorías que describen el humor a partir de las emociones que éste genera y de algunos elementos lingüísticos propios del humor que no resultan útiles para nuestros propósitos; sin embargo, se dan a conocer como muestra de lo que se ha hecho hasta ahora en materia de humor.

### *2.2.1Teoría de la descarga*

También conocida como Teoría de la liberación. Este tipo de teorías se caracteriza por atribuirle las causas del humor a una necesidad de liberar sentimientos reprimidos.

Uno de los máximos exponentes de esta teoría es el psicoanalista Sigmund Freud, quien dentro de sus aportaciones al análisis del humor opinaba lo siguiente: *"Como los chistes y lo cómico, el humor tiene algo de liberación acerca de algo, pero tiene también algo de un sentido de grandeza y de elevación que suele faltar en las otras dos formas de obtener placer desde una actividad intelectual. Esta grandeza se basa en el triunfo del narcisismo, y en la victoriosa autoafirmación de la invulnerabilidad del ego*" [3].

### *2.2.2 Teoría de la superioridad*

Este tipo de teorías se caracteriza por encontrar en el humor un sentido de insensibilidad, orgullo, arrogancia y humillación por parte de quien se ríe de los fallos o desgracias de los demás.

Uno de sus máximos exponentes es el filósofo francés Henri Bergson, quien sostenía lo siguiente: *"Digiérase que lo cómico sólo puede producirse cuando recae en una superficie espiritual lisa y tranquila. Su medio natural es la indiferencia. No hay mayor enemigo de la risa que la emoción. No quiero decir que no podamos reírnos de una persona que, por ejemplo, nos inspire piedad y hasta afecto; pero en este caso será preciso que por unos instantes olvidemos ese afecto y acallemos esa piedad. En una sociedad de inteligencias puras quizá no se llorase, pero probablemente se reiría, al paso que entra almas siempre sensibles, concertadas al unísono, en las que todo acontecimiento produjese una resonancia sentimental, no se conocería ni comprendería la risa. Probad por un momento a interesaros por cuanto se dice y cuanto se hace; obrad mentalmente con los que practican la acción; sentid con los que sienten; dad, en fin, a vuestra simpatía su más amplia expansión, y como al conjuro de una varita mágica, veréis que las cosas más frívolas se convierten en graves y que todo se reviste de matices severos. Desimpresionaos ahora, asistid a la vida como espectador indiferente, y tendréis muchos dramas trocados en comedia. Basta que cerremos nuestros oídos a los acordes de la música en un salón de baile, para que al punto nos parezcan ridículos los danzarines. ¿Cuántos hechos humanos resistirían a esta prueba? ¿Cuántas*

*cosas no veríamos pasar de lo grave a lo cómico, si las aislásemos de la música del sentimiento que las acompaña? Lo cómico, para producir todo su efecto, exige como una anestesia momentánea del corazón. Se dirige a la inteligencia pura"* [4].

### *2.2.3 Teoría de la incongruencia*

Este tipo de teorías no basan el origen del humor en cuestiones sentimentales propias de los seres humanos; sino en lo inesperado de un acontecimiento que contraviene a una sucesión de eventos que predecimos o suponemos como lógicos acerca de una situación dada.

Al respecto, el filósofo alemán Arthur Schopenhauer afirmaba que: "La causa de lo risible está siempre en la subsunción o inclusión paradójica, y por tanto inesperada, de una cosa en un concepto que no le corresponde, y la risa indica que de repente se advierte la incongruencia entre dicho concepto y la cosa pensada, es decir, entre la abstracción y la intuición. Cuanto mayor sea esa incompatibilidad y más inesperada en la concepción del que ríe, tanto más intensa será la risa" [5].

### *2.2.4 Teoría General del Humor Verbal*

La Teoría General del Humor Verbal o *"General Theory of Verbal Humor (GTVH)"*, de Víctor Raskin y Salvatore Attardo [18], es el resultado de la integración de la teoría semántica *"Semantic Script-based Theory of Humor",* de Víctor Raskin [31],

con un modelo de representación de chistes de 5 niveles, propuesto por Salvatore Attardo. Destaca por ser la teoría más aceptada desde el punto de vista lingüístico y es por ello que no entra en ninguna de las tres clasificaciones anteriores.

Lo que la diferencia de las teorías expuestas *a priori* es que no pretende explicar el humor a través de las emociones o de los procesos fisiológicos que se dan en los seres humanos; sino que explica el humor verbal a través de sus características lingüísticas.

Según esta teoría, un chiste puede ser visto como una 6-tupla:

**(LA, SI, NS, TA, SO, LM)**

Dónde:

**LA:** Es el lenguaje del chiste; es decir, la fonética, fonología, morfo fonética, morfología, léxica, sintáctica, semántica y pragmática en niveles del lenguaje que presenta.

**NS:** Es su estrategia narrativa; es decir, el género del chiste. Éste puede ser representado de manera expositiva o narrativa, acertijo, secuencia de preguntas y respuestas, expandido a un diálogo, etc.

**TA:** Es su objetivo; es decir, a quién va dirigido. Un chiste puede ser dirigido a un individuo o a un grupo étnico, social o político.

**SI:** Es la situación. Consiste en el resto del contenido u otros participantes del chiste; como actividades, objetos, instrumentos, etc.

**LM:** Es el mecanismo lógico; es decir, la manera en que el humor verbal es detectado.

**SO:** Es la oposición semántica; es decir, los dos guiones con los cuales un chiste es compatible y que a través del mecanismo lógico se unen para detectar la situación humorística.

Sin embargo, cabe señalar que esta teoría no pretende ser un modelo para la producción chistes; sino que es simplemente un instrumento para evaluar el grado de similitud entre textos de chistes particulares [31].

## *2.3 Generación de humor, como primeros ensayos antes de la Detección del humor*

Sin duda alguna los estudios acerca del humor no son recientes, dado que datan desde las aportaciones de los grandes pensadores griegos, hasta la fecha. Sin embargo, estos estudios han consistido en la promulgación de diversas teorías acerca de por qué reímos, tratadas desde el punto de vista de ciencias como la filosofía, la psicología, la sociología y la neurociencia.

El surgimiento de la computación como herramienta de apoyo a las diferentes ramas del saber; ha significado una ayuda invaluable en los avances al entendimiento del humor, en particular a la Lingüística (cuya nueva disciplina se denomina Lingüística Computacional o Procesamiento del Lenguaje Natural, PLN), desde un punto de vista estadístico.

Pero hasta ahora la mayor parte de los trabajos realizados con humor, con el apoyo de la computación, tienen que ver con la generación y no con la detección. Esto se debe a que cierto tipo de humor se presenta como plantillas, y es mucho más fácil hacer la sustitución de algunas palabras y de algunas letras de las palabras, que tratar de generar una frase que, además de ser coherente en su estructura, resulte humorística.

Por mencionar sólo algunos trabajos de generación de humor, existen hoy en día programas como: LIBJOG, JAPE, Elmo, WISCRAIC, Ynperfect Pun Selector, HAHAcronym, MSG, Tom Swifties y Jester [19]; de los cuales no es propósito de este trabajo de tesis señalar sus capacidades, pero que distan todavía mucho de generar enunciados humorísticos congruentes.

## 2.4 Trabajos actuales sobre la Detección del humor

Con respecto a la detección del humor, los trabajos más prominentes son los de los doctores Rada Mihalcea y Carlo Strapparava[6][16][24], quienes dentro de sus avances en la materia, han demostrado llevar un orden y una secuencia bien definidos, partiendo de cuestiones tan básicas como de conocer en primera instancia de qué tratan los chistes, enfocándose únicamente al humor escrito. A continuación se hace una breve revisión a los trabajos de detección de humor de estos y otros doctores.

### 2.4.1 "Making Computers Laugh"

"En "Making Computers Laugh" [20], Rada y Attardo ya hacen uso de características de estilo de los textos humorísticos tales como la aliteración, la antonimia y el contenido adulto; así como de características basadas en contenido y de una combinación de ambas características, para hacer una clasificación. Específicamente, comparan los resultados obtenidos con Naïve Bayes y Support Vector Mechines, dos clasificadores regularmente utilizados para la clasificación de textos.

Para tal efecto, ellos consiguieron recolectar 16,000 chistes de un promedio de 15 palabras y una cantidad similar de textos no humorísticos compuestos de encabezados de artículos, proverbios y del British National Corpus.

Los resultados obtenidos en los experimentos de clasificación automática muestran que las aproximaciones computacionales representan una solución viable para la tarea de reconocimiento de humor, que es posible utilizar estas técnicas para distinguir entre textos humorísticos y no humorísticos, y que se alcanza buen desempeño usando técnicas de clasificación basadas en características de contenido y de estilo.

### *2.4.2 "Characterizing Humor: An exploration of Features in Humorous Texts"*

En *"Characterizing Humor: An exploration of Features in Humorous Texts"* [21], Mihalcea y Pulman, analizan de forma detallada otras dos características que, según teorías psicológicas, están presentes de manera muy frecuente en el humor, *human-centeredness* y *negative polarity.* Comprobando con este estudio la veracidad de la suposición y descubriendo además que estas características son consistentes en distintos conjuntos de datos.

### *2.4.3 "Recognizing Humor Without Recognizing Meaning"*

Por su parte, en *"Recognizing Humor Without Recognizing Meaning"* [22], Sjöbergh y Araki, realizan también una aproximación de aprendizaje de máquina para clasificar textos de una línea como chistes o como texto normal, basados también en la combinación de características simples para ver si éstas son suficientes para detectar el humor.

Las características que ellos utilizan son: sobreposición de palabras con otros chistes, presencia de palabras comunes en chistes, ambigüedad y sobreposición de palabras con idiomas comunes. Comprobaron que cuando se prueba y entrena con cantidades similares de chistes y enunciados del British National Corpus (BNC), se alcanza una precisión de clasificación del 85%.

### *2.4.4 "The Impact of Semantic and Morphosyntactic Ambiguity on Automatic Humor Recognition"*

En *"The Impact of Semantic and Morphosyntactic Ambiguity on Automatic Humor Recognition"* [23], Reys, Buscaldi y Rosso, trabajan con la ambigüedad, y confirman que la información obtenida del estudio de la ambigüedad puede ser tomada en cuenta como un conjunto de características para reconocer automáticamente humor; especialmente con medidas tales como *perplexity, mean of senses* y *sense dispersion.*

### *2.4.5 "Computacional Models for Incongruity Detection in Humour"*

Pero los trabajos más innovadores al respecto son siempre los de Rada Mihalcea y Carlo Strapparava, quienes con la colaboración de Stephen Pulman, publicaron en marzo de 2010, *"Computacional Models for Incongruity Detection in Humour"* [24], en donde exploran diversos modelos computacionales para la resolución de

incongruencias (que es una de las teorías del humor más ampliamente aceptadas; la cual sugiere que el humor se debe a la mezcla de dos cuadros opuestos de interpretación posible para un enunciado).

En este trabajo introducen un nuevo conjunto de datos, consistente en una serie de 'set ups' (preparaciones para un punch line o la palabra o conjunto de palabras que detonan el efecto humorístico), cada uno de ellos seguido por cuatro posibles continuaciones coherentes, de las cuales sólo una causa el efecto humorístico. Usando este conjunto de datos, redefinen la tarea como la identificación automática del punch line entre todas las terminaciones convincentes.

Exploran además diferentes medidas de relación semántica, junto con un número de características humorísticas específicas, e intentan comprender lo apropiado de su uso como modelos para la detección de incongruencias.

## *2.5 Discusión*

A pesar de los avances en materia de detección de humor en textos cortos, no se tiene una herramienta realizada para el idioma español; esto por la razones obvias de nuestro mundo globalizado, que nos impone como lenguaje universal el idioma inglés, lo cual trae como consecuencia que los lexicón más completos sean precisamente los de este idioma.

La mayor desventaja de esta situación es que, aunque es cierto que el idioma inglés es un lenguaje estandarizado y utilizado por la gente del mundo que hace ciencia, la mayor parte de la población mundial no domina ni en lo básico este idioma, lo cual representa un problema al dejarlos sin posibilidades de hacer uso en primera instancia de los avances que se vayan logrando en esta y en las demás ramas de la ciencia.

Es por ello que este trabajo de tesis viene a subsanar esta carencia, proporcionando como valor agregado la detección de una expresión tan mexicana como lo es el albur, que es otra variante del tan diverso y tan complejo humor.

# Capítulo 3. Marco teórico

En este capítulo se describen de manera puntual todos los conceptos necesarios para entender el presente trabajo de tesis. Si el lector ya está familiarizado con todos estos conceptos, puede pasar directamente al siguiente capítulo.

## *3.1 Conceptos humorísticos*

Los siguientes son los conceptos humorísticos utilizados en este trabajo.

### *3.1.1 Humor*

(Del lat. *humor, -oris*). 1. m. Genio, índole, condición, especialmente cuando se manifiesta exteriormente. 2. m. Jovialidad, agudeza. 3. m. Disposición en que alguien se halla para hacer algo. 4. m. Buena disposición para hacer algo. 7. m. Psicol. Estado afectivo que se mantiene por algún tiempo. [7].

### *3.1.2 Chiste*

(De *Chistar*). 1. m. Dicho u ocurrencia aguda y graciosa. 2. m. Dicho o historieta muy breve que contiene un juego verbal o conceptual capaz de mover a risa. Muchas veces se presenta ilustrado por un dibujo, y puede consistir sólo en éste. 3. m. Suceso gracioso y festivo. [7].

### *3.1.3 Albur*

El albur es un juego de palabras con un doble sentido. Este doble sentido es por lo general sexual o escatológico. La intención del albur es la de hacer una zancadilla verbal que implique burla o que comunique una afrenta – por lo general amistosa – hacia un interlocutor.

Este juego de palabras emplea principalmente dos elementos verbales. El primer elemento es la asociación de palabras y expresiones aparentemente inocuas con términos sexuales y/o escatológicos. El segundo elemento que se emplea generalmente en el albur es la deconstrucción de palabras inocuas para crear palabras o expresiones completamente distintas, pero que impliquen un mensaje sexual o escatológico. [8].

## *3.2 Atributos humorísticos*

Los atributos del humor que se tratan en este trabajo son los siguientes:

### *3.2.1 Rima*

(Del ant. rimo, éste del lat. *rhythmus,* y éste del gr. *rhythmos*, movimiento concertado). 1. f. Consonancia o consonante. 2. f. Asonancia o asonante. 3. f. Composición en verso, del género lírico. 4. f. Conjunto de consonantes de una lengua. 5. f. Conjunto de consonantes o asonantes empleados en una composición o en todas las de un poeta. [7].

Existen 2 tipos de rimas [35]:

a) Rima Consonante: Es aquella en la cual las terminaciones de las palabras que riman son exactamente iguales, tanto en las vocales como en las consonantes, desde vocal acentuada hacia el final de la palabra.

b) Rima Asonante: Es aquella en la cual las palabras que riman, solo tienen las mismas consonantes, a partir de la ultima vocal acentuada

### *3.2.2 Aliteración*

(Del lat. *littera*, letra). 1. f. Ret. Repetición notoria del mismo o de los mismos fonemas, sobre todo consonánticos, en una frase. 2. f. Ret. Figura que, mediante la repetición de fonemas, sobre todo consonánticos, contribuye a la estructura o expresividad del verso. [7].

### *3.2.3 Antónimo, antonimia*

(De *anti-* y *-ónimo*). 1. adj. Ling. Se dice de las palabras que expresan ideas opuestas o contrarias. [7].

### *3.2.4 Contenido adulto*

Lenguaje informal consistente en palabras y expresiones que no son consideradas apropiadas para ocasiones formales; a menudo peyorativo o vulgar. [9].

## *3.3 Métodos de Aprendizaje Máquina*

Estos métodos tratan de construir, habitualmente a partir de ejemplos etiquetados por un experto en la materia, una máquina (software o hardware) que extraiga toda la información relevante de las muestras disponibles. Una vez entrenada dicha máquina, ésta será capaz de generalizar bien para nuevos ejemplos no etiquetados, de modo que proporcione la respuesta (esto es, la etiqueta) correcta para los mismos. Algunos tipos de estos métodos son las redes neuronales, máquinas de vectores soporte (SVM's), procesos gausianos (GP's), etc. [10].

## *3.4 Minería de datos*

Es la Etapa de Descubrimiento en el proceso denominado Descubrimiento de Conocimiento en Bases de Datos (KDD, por su siglas en inglés). La cual se considera como una tecnología compuesta por etapas que integra diversas áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo; con el fin de reunir sus principales ventajas en la búsqueda de patrones ocultos en la información que pudieran representar un conocimiento extraído de dichas Bases de Datos. [12].

## 3.5 Clasificación Estadística

Una clasificación estadística es una clasificación que tiene un conjunto de categorías discretas, que se pueden asignar a una variable específica registrada en un estudio estadístico o en un archivo administrativo, y es usada en la producción y presentación de estadísticas. [13].

## 3.6 Paradigmas de clasificación

Los siguientes son los paradigmas de clasificación más utilizados.

### 3.6.1 Análisis de regresión

Es un proceso interactivo el cual consiste en el estudio de un conjunto de técnicas que son usadas para establecer una relación entre una variable cuantitativa llamada variable dependiente y una o más variables independientes llamadas variables predictoras. Las variables independientes también deberían ser cuantitativas, sin embargo es permitido que algunas de ellas sean cualitativas. La ecuación que representa la relación es llamada el modelo de regresión. Si todas las variables independientes fueran cualitativas entonces el modelo de regresión se convierte en un modelo de diseños experimentales. [14].

### *3.6.2 Árboles de decisión*

Es método conveniente para presentar y analizar una serie de decisiones que se deben tomar en diferentes puntos de tiempo. En un árbol de decisiones hay nodos y líneas rectas que son las ramas, cuadrados que son los nodos o puntos de decisión y círculos que son nodos o puntos de azar. Las ramas que se extienden de los nodos indican las alternativas que se pueden tomar en el caso de nodos de decisión, o los diferentes resultados de un evento en el caso de los nodos de azar. En este último caso cada rama tiene asociada una probabilidad de ocurrencia. Esta probabilidad es una medida de la posibilidad de que ese evento ocurra. La suma de las probabilidades de las ramas que parten de cada nodo de evento es igual a uno. Es decir, que se supone que los eventos son exhaustivos; a los nodos de decisión no se les asigna probabilidades, ya que en esos puntos el decisor tiene el control y no es un evento aleatorio, sujeto al azar. [15].

### *3.6.3 Máquinas de Soporte Vectorial*

Las máquinas de soporte vectorial o máquinas de vectores de soporte (Support Vector Machines, SVMs) son un conjunto de algoritmos de aprendizaje supervisado que están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio

posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase [28].

### *3.6.4 Naive Bayes*

En términos simples, un clasificador Naive Bayes asume que la presencia (o ausencia) de un rasgo en particular de una clase no está relacionado a la presencia (o ausencia) de cualquier otro rasgo. Por ejemplo, se puede considerar que una fruta es una manzana si es roja, redonda y de unas 4" de diámetro. Aunque estas características dependan unas de otras de la existencia de otras características. Un clasificador Naive Bayes considera todas estas propiedades para contribuir de manera independiente a la probabilidad de que esta fruta es una manzana [29].

### *3.6.5 Red Neuronal*

Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas consisten en una simulación de las propiedades observadas en los sistemas neuronales biológicos a través de modelos matemáticos recreados mediante mecanismos artificiales (como un circuito integrado, un ordenador o un conjunto de válvulas). El objetivo es conseguir que las máquinas den respuestas similares a las que es capaz de dar el cerebro que se caracterizan por su generalización y su robustez.

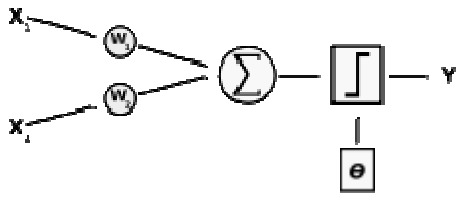


***Ilustración 1. Perceptrón con 2 entradas.***

Una red neuronal se compone de unidades llamadas neuronas. Cada neurona recibe una serie de entradas a través de interconexiones y emite una salida. Esta salida viene dada por tres funciones:

1. Una función de propagación (también conocida como función de excitación), que por lo general consiste en el sumatorio de cada entrada multiplicada por el peso de su interconexión (valor neto). Si el peso es positivo, la conexión se denomina *excitatoria*; si es negativo, se denomina *inhibitoria*.

2. Una función de activación, que modifica a la anterior. Puede no existir, siendo en este caso la salida la misma función de propagación.

3. Una función de transferencia, que se aplica al valor devuelto por la función de activación. Se utiliza para acotar la salida de la neurona y generalmente viene dada por la interpretación que queramos darle a dichas salidas. Algunas de las más utilizadas son la función sigmoidea (para obtener valores en el intervalo [0,1]) y la tangente hiperbólica (para obtener valores en el intervalo [-1,1]). [32].

.

### *3.6.6 Sistemas Basados en Reglas de Producción*

Las reglas de producción son un método procedimental de representación del conocimiento, es decir, pone énfasis en representar y soportar las *relaciones inferenciales* del dominio, en contraposición a los métodos declarativos (énfasis en la representación de los hechos). [33].

La estructura de una regla es:

SI <*antecedentes*>

ENTONCES <*consecuentes*>

Los antecedentes son las *condiciones* y los consecuentes las *conclusiones*, *acciones* o *hipótesis*.

Cada regla por si misma constituye un gránulo completo de conocimiento. La inferencia en los Sistemas Basados en Reglas se realiza mediante *emparejamiento*. Hay dos tipos, según el sentido:

- *Sistemas de encadenamiento hacia adelante:* una regla es *activada* si los antecedentes emparejan con algunos hechos del sistema.
- *Sistemas de encadenamiento hacia atrás:* una regla es *activada* si los consecuentes emparejan con algunos hechos del sistema.

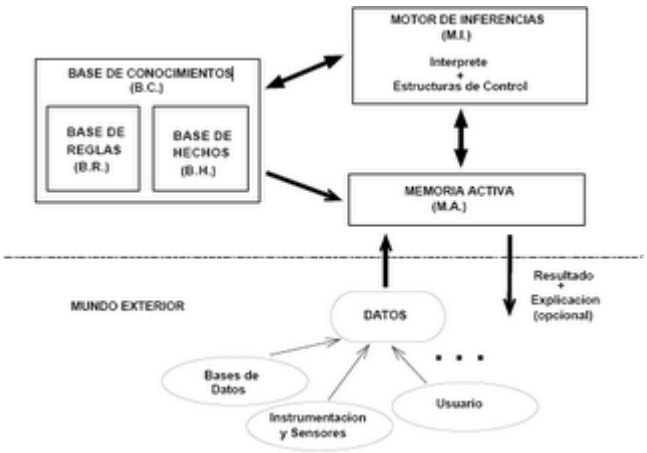


***Ilustración 2. Arquitectura de los Sistemas Basados en Reglas:***

- *Base de Conocimientos:* reúne todo el conocimiento del sistema (Hechos + Reglas).

- *Memoria Activa:* contiene los hechos que representan el *estado actual del problema* (iniciales + inferidos a posteriori) y las *reglas activadas* (en condiciones de ser ejecutadas).

*Motor de Inferencias: decide* que reglas activadas se ejecutarán.

## *3.7 WEKA*

WEKA es una colección de algoritmos de aprendizaje máquina para tareas de minería de datos. Los algoritmos pueden ser aplicados a un conjunto de datos ya sea directamente o llamados desde el propio código del usuario. WEKA contiene herramientas para pre-procesamiento, clasificación, regresión, aglomeración, reglas de asociación y visualización. Puede ser también utilizado para el desarrollo de

nuevos sistemas de aprendizaje máquina. WEKA es un programa de código abierto distribuido bajo la licencia pública general GNU. [10].

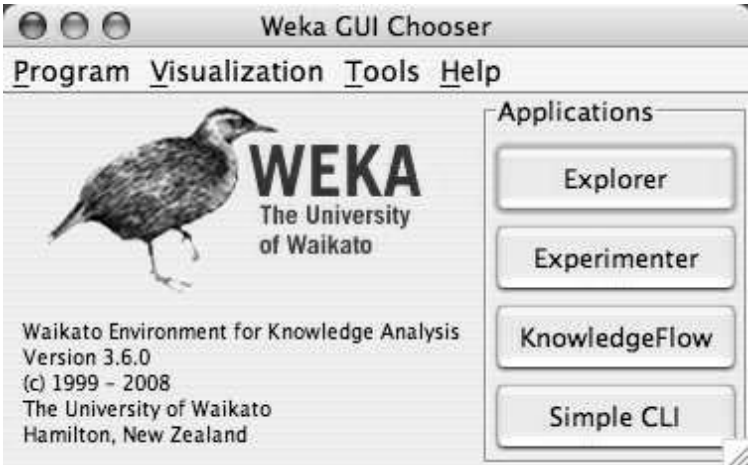


***Ilustración 3. Ventana de inicio de la aplicación WEKA.***

## *3.8 Medidas que se analizan*

Son las medidas devueltas por WEKA que serán analizadas.

### *3.8.1 Accuracy*

En los algoritmos de clasificación, indica el número o el porcentaje de ejemplos clasificados correctamente de todas las clases. [30].

### 3.8.2 False Positive Rate (FP_Rate)

En los algoritmos de clasificación es la proporción de ejemplos que fueron clasificados en la clase *x*, pero pertenecen a una clase diferente, entre todos los elementos que no pertenecen a la clase *x*. [30].

### 3.8.3 Precision

En los algoritmos de clasificación, es la proporción de ejemplos que realmente pertenecen a la clase *x* entre todos aquellos que fueron clasificados como pertenecientes a la clase *x*. [30].

### 3.8.4 Recall

En los algoritmos de clasificación, es la proporción de ejemplos que fueron clasificados correctamente en la clase *x*, entre el total de ejemplos que realmente pertenecen a la clase *x*. Es similar a *True Positive Rate*. [30].

### 3.8.5 True Positive Rate (TP_Rate)

En los algoritmos de clasificación, es la proporción de ejemplos que fueron clasificados correctamente en la clase *x*, entre el total de ejemplos que realmente pertenecen a la clase *x*. Es similar a *Recall*. [30].

## *3.9 Validación Cruzada*

Traducción de *Cross Validation*. Es un método utilzado por WEKA cuando no se le pasan dos archivos como parámetros. Para mandar llamar a las funciones de la aplicación WEKA, se requiere, como parte de los parámetros, que se especifiquen dos archivos de extensión *.arff,* uno de los cuales será utilizado como instancias de entrenamiento y el otro como instancias de prueba. Para el caso de que sólo se le especifique un solo archivo (que es nuestro caso), la aplicación hará automáticamente el uso del la Validación Cruzada por 10, el cual consiste en reordenar aleatoriamente el conjunto de datos y dividirlos en 10 pliegues de igual tamaño. En cada iteración, un solo pliegue se utiliza para las pruebas y los restantes 9 pliegues se utilizan para el entrenamiento del clasificador. Los resultados de las pruebas se recogen y se promedian con los demás pliegues. Esto le da a la *Validación Cruzada* una estimación de la precisión en cuanto a la clasificación. [30].

# Capítulo 4. Metodología para la formación de los rasgos

En este capítulo se describe de manera detallada los pasos seguidos para la extracción de los rasgos de los textos tanto humorísticos como no humorísticos necesarios para su posterior análisis y detección de los textos humorísticos.

## *4.1 Esquema de la metodología*

La metodología llevada a cabo para la detección del humor se puede apreciar en la siguiente ilustración:

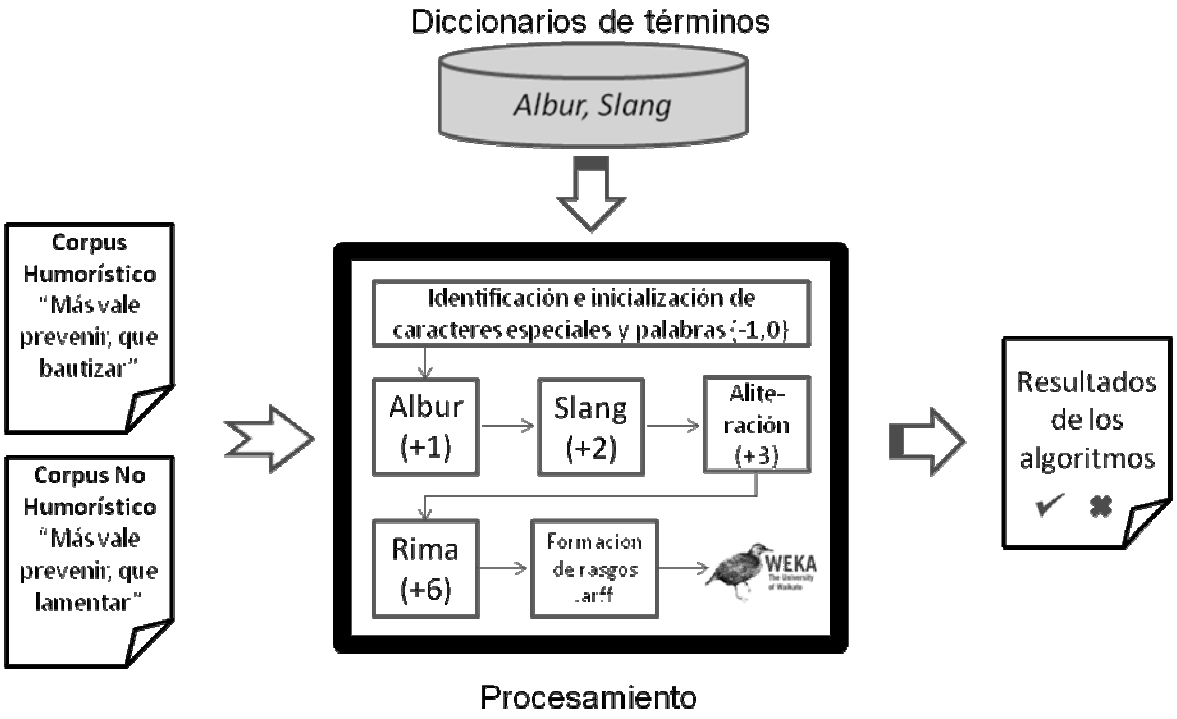


***Ilustración 1. Metodología para la detección de humor.***

## *4.2 Requerimientos previos*

Antes de poder hacer la detección del humor es necesario contar con dos diccionarios de términos, uno con palabras utilizadas en el lenguaje del albur (*véase Anexo 4*), colectados del libro "Antología del Albur" [8]; y otro con palabras de contenido sexual y otras palabras comunes en los chistes (*véase Anexo 3*), colectados de diversos sitios de internet [25][26]. Esto con la finalidad de que sean utilizados como soporte para detectar esas palabras en los textos. Para nuestro trabajo, se cuenta con un diccionario de términos de albur con 203 elementos y un diccionario de palabras de contenido sexual y otras palabras comunes en los chistes de 93 elementos.

Es necesario también contar con al menos dos archivos de texto para el análisis de su contenido, asegurándose de que uno de ellos estará compuesto, al menos en su mayoría, por textos humorísticos; esto para garantizar el mejor desempeño del detector automático de humor. Para nuestro trabajo, se cuenta con un corpus de textos humorísticos de 3,676 elementos y un corpus de textos no humorísticos (dichos) de 10,000 elementos.

## 4.3 Procesamiento

El procesamiento se hace a través de la herramienta de Detección Automática de Humor en Textos Cortos en Español (DAHTCE), desarrollada específicamente para este trabajo de tesis (*ver Ilustración 2*), la cual realiza los siguientes pasos:

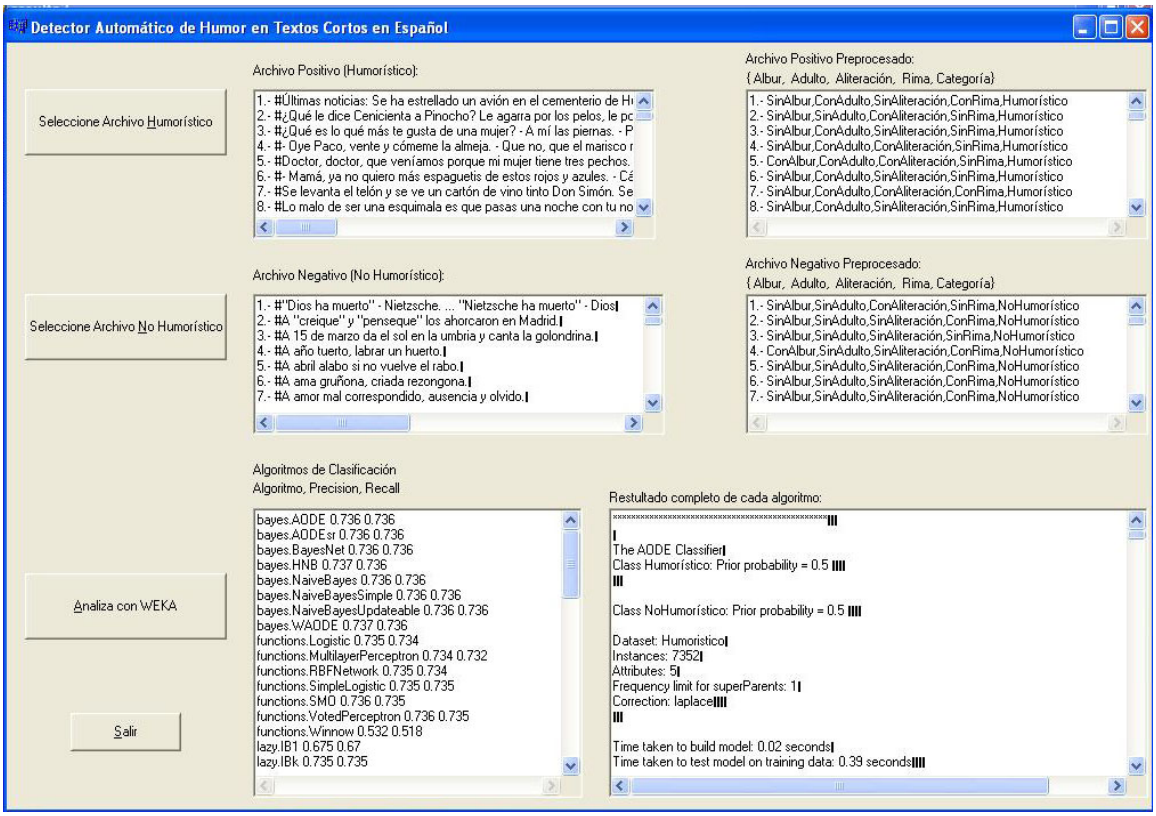


***Ilustración 2. Herramienta de procesamiento DAHTCE.***

### 4.3.1 Recepción de archivos

Recibe dos archivos, a suponerse, uno con textos humorísticos y otro con textos no humorísticos, los cuales procesa por separado identificando cada una de las palabras y caracteres especiales, almacenándolos en una estructura de datos

conformada por una cadena de caracteres y una variable entera donde se inicializan con un valor, (-1) para los caracteres especiales y (0) para las palabras.

Una vez identificadas todas la palabras y los caracteres, DAHTCE utilizará los módulos de *albur, contenido adulto, rima* y *aliteración*, para identificar, por cada renglón del archivo (cada renglón es un chiste o un dicho), los atributos que se encuentran en sus palabras.

### *4.3.2 Módulo de Albur*

El módulo de *albur* identifica, a través del diccionario de términos de albur, si alguna de las palabras del texto (mayores a tres caracteres, para evitar la búsqueda con artículos indefinidos, conjunciones, etc.) presenta esta característica. En dado caso que así sea, se modifica el valor inherente a la palabra incrementándolo en una unidad, para denotar este rasgo en ella. (*véase Tabla 1*).

Cabe mencionar que el albur tiene la dificultad de que es una expresión bastante rica, lingüísticamente hablando; presenta juegos de palabras bastantes complejos que lo hacen muy diverso; por lo tanto, resulta fuera del alcance de este trabajo de investigación intentar detectarlo en todas sus variantes.

Sin embargo, se tiene una lista bastante amplia de las palabras que con mayor frecuencia se encuentran presentes en este lenguaje (*véase Anexo 4*).

### *4.3.3 Módulo de Contenido Adulto*

El módulo de C*ontenido Adulto* identifica, a través del diccionario de términos de contenido adulto, si alguna de las palabras del texto (igualmente mayores a tres caracteres, por las razones arriba expuestas) presenta esta característica. En dado caso que así sea, se modifica el valor inherente a la palabra incrementándolo en dos unidades, para denotar este rasgo en ella. (*ver Tabla 1*).

### *4.3.4 Módulo de Aliteración*

El módulo de *Aliteración* identifica en los textos esta característica tomando en cuenta únicamente las tres primeras letras de cada palabra (de igual forma, para palabras mayores a tres letras), porque son suficientes para identificar este rasgo; y se comparan con las tres primeras letras de las demás palabras. En caso de encontrar igualdad entre ellas, se suman 3 unidades al valor que trae originalmente la palabra (recordando que la palabra puede estar relacionada a un valor de 0, 1 ó 2; dependiendo si resultó ser una palabra con albur o contenido adulto), para indicar que también forma parte de una aliteración, siendo características no exclusivas entre las palabras. (*ver Tabla 1*).

### *4.3.5 Módulo de Rima*

El módulo de *rima* identifica en los textos esta característica analizando únicamente las 3 últimas letras de cada palabra. En caso de encontrar igualdad entre ellas, se suman 6 unidades al valor que trae originalmente la palabra (debido a que en esta ocasión las palabras pueden contener ya valores entre 0 y 5 producto de los análisis anteriores) para indicar que también forma parte de una rima. (*ver Tabla 1*). Cabe aclarar, que en esta tesis únicamente se identifica la rima consonante.

### *4.3.6 Formación de los Rasgos*

El módulo *formación de rasgos* tiene por objetivo la formación del vector de los rasgos, el cual consiste en el análisis final del número asociado a las palabras. Éste indicará los atributos que tiene cada una de ellas (*véase Tabla 1*).

De cada frase se realiza un conteo de cuántas palabras resultaron con cada atributo para determinar los atributos generales de dicho texto; de modo que al finalizar el conteo se tiene para cada frase los valores: *SinAlbur, ConAlbur, SinAdulto, ConAdulto, SinAliteracion, ConAliteracion, SinRima, ConRima*, dependiendo de la presencia o no de cada atributo en la frase.

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

***Tabla 1. Valores de las palabras de los textos al finalizar el procesamiento.***

Por último, se concatenan todos estos valores junto con una etiqueta de *Humoristico* o *NoHumoristico,* dependiendo si las frases provienen del archivo con ejemplos positivos o del archivo con ejemplos negativos. Los resultados son unos vectores de datos como los siguientes:

{SinAlbur, ConAdulto, SinAliteracion, ConRima, Humoristico}

{SinAlbur, SinAdulto, ConAliteracion, SinRima, NoHumoristico}

El resultado final de este módulo es la creación de un archivo *.arff (Attribute-Relation File Format)*, que es el tipo de archivos que procesa WEKA, y cuyo formato se puede apreciar en el ***Anexo 2***.

### *4.3.7 Módulo WEKA*

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

1. ayes.AODE
2. ayes.AODEsr
3. ayes.BayesNet
4. ayes.HNB
5. ayes.NaiveBayes
6. ayes.NaiveBayesSimple
7. ayes.NaiveBayesUpdateable
8. ayes.WAODE
9. ules.ConjunctiveRule
10. ules.DecisionTable
11. ules.OneR
12. ules.Prism
13. ules.ZeroR
14. azy.IB1
15. azy.IBk
16. azy.KStar
17. azy.LBR
18. azy.LWL

19. rees.DecisionStump
20. rees.Id3
21. rees.J48
22. rees.RandomForest
23. rees.REPTree
24. unctions.Logistic
25. unctions.MultilayerPerceptron
26. unctions.RBFNetwork
27. unctions.SimpleLogistic
28. unctions.SMO
29. unctions.VotedPerceptron
30. unctions.Winnow
31. isc.HyperPipes
32. isc.VFI

Esto es con la finalidad de tener una amplia gama de posibilidades en resultados, para así poder analizar cada uno de ellos con el propósito de determinar cuál o cuáles son los mejores algoritmos que nos ayuden a hacer una detección de humor en textos cortos en español, dados los atributos que extraemos de dichos textos.

Cabe señalar, que no es propósito de este trabajo de tesis entender y explicar a fondo las características particulares de cada uno de los algoritmos, sus semejanzas y diferencias con todos los demás; sin embargo, se hace una breve especificación en el ***Anexo 1***.

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

# Capítulo 5. Resultados Experimentales

En este capítulo se muestran algunos de los experimentos más importantes realizados con la herramienta de DAHTCE y con la aplicación de WEKA; así como un análisis de los resultados obtenidos más relevantes.

## *5.1 Metodología experimental*

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

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

## 5.2 Recolección de los datos

Para poder hacer ejercicios relacionados con la detección de humor a través de los algoritmos de clasificación de WEKA, es necesario tener una amplia colección de textos previamente clasificados como humorísticos y no humorísticos. La forma más rápida y barata de obtener dicha información es a través del internet.

El proceso de colección de los textos cortos se hizo a mano; es decir, sin ningún programa de por medio que hiciera la búsqueda y la depuración automática. Por el contrario, simplemente se utilizó el buscador Google para ubicar algunas páginas que ofrecieran textos humorísticos (chistes) [25], y posteriormente otras que ofrecieran textos no humorísticos en un formato similar al de los chistes, siendo estos los llamados dichos [27].

Después de varios meses de buscar y depurar, se logró colectar 3675 chistes y 10000 dichos, y se estuvo en posibilidades de iniciar con los experimentos

### 5.2.1 Datos positivos: chistes

Los textos humorísticos fueron obtenidos de varias páginas de internet [25]. El filtro para que fuesen considerados como textos cortos, fue que tuviesen un máximo de 50 palabras, la razón de esta medida es que muchos de los chistes se presentan como pequeñas narraciones, y no son muchos los que quedan definidos en menos

de 20 palabras; además, se tomó el criterio de las propias páginas que ofrecen los textos humorísticos en el sentido de que ya los tienen clasificados como chistes cortos.

Una dificultad que se encontró a la hora de buscar los textos humorísticos es que en su gran mayoría los chistes vienen representados en narraciones de más de 50 palabras. Además, se tuvo que hacer una depuración bastante minuciosa, dado que cada sitio que ofrece chistes a los usuarios funciona como blog, lo cual significa que todas las personas que así lo deseen, pueden subir sus chistes; lo que trae como consecuencia que muchos de los textos sean narraciones, con algunas pequeñas variantes, de un mismo chiste. Aún así se logró colectar un total de 3,675 textos humorísticos. [25].

### *5.2.2 Datos Negativos: dichos*

Para los datos negativos se buscó un tipo de textos cortos que tuvieran semejanza en cuanto a su estructura y longitud con los chistes; teniendo la posibilidad de utilizar algunos encabezados de las noticias de periódicos o de dichos. Se decidió trabajar con los dichos, ya que son frases populares cortas que tienen como intensión la de transmitir experiencias de vida y sabiduría, con una longitud que rara vez rebasa las 30 palabras; y por su mayor similitud a los chistes en comparación con los encabezados de noticias de periódicos.

Por el momento y para los propósitos de este primer intento de detección de humor en textos cortos en español, fue suficiente trabajar con dichos, ya que se logró colectar un total de 10,000 de ellos [26].

## *5.3 Preparación de los datos*

La única preparación importante que se requirió antes de iniciar con la fase experimental, fue la depuración minuciosa, y a mano, de los textos tanto humorísticos como no humorísticos, principalmente la de los primeros; ya que fue necesario determinar para cada uno de los chistes, si no era demasiado grande para los propósitos del proyecto o determinar si no estaba duplicado o no era muy semejante a alguno ya existente.

Para el caso de los dichos, el de máxima longitud registrada fue de 35 palabras, aún mucho menor que la longitud de palabras del más grande de los chistes, que es de 50 palabras; además de que la depuración fue un poco más sencilla, dado que todos ellos se consiguieron de un mismo sitio web y de un mismo autor [27], lo que supuso un colección ya pre-depurada.

Y lo previo a cada experimentación es determinar la cantidad de ejemplos tanto positivos como negativos que nos interesa analizar; y si es posible, definir previamente también algunos conjuntos de elementos que nos interese

experimentar. De ahí en fuera, el programa DAHTCE es bastante sencillo de entender y de manejar, y no requiere de una configuración previa, lo cual no representa mayor complicación para los usuarios del mismo.

Cabe señalar que para este trabajo de tesis, se analizan los algoritmos para tres casos particulares que considero de suma importancia para encontrar el mejor de ellos en la detección del humor en textos cortos en español:

El caso para cuando se analizan las 3,675 instancias positivas y las 10,000 instancias negativas como un solo conjunto de datos. En lo subsecuente le llamaremos ***CasoTodasInstancias.***

El caso para cuando se analizan 3,675 instancias positivas y 3675 instancias negativas (estas últimas tomadas al azar), como un solo conjunto de datos. En lo subsecuente le llamaremos ***Caso-3675-Heterogéneo***

Y por último el caso para 2833 instancias positivas y 2833 instancias negativas que presentan cuando menos un rasgo distintivo del humor. En lo subsecuente le llamaremos ***Caso-2833-Homogéneo.***

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

## *5.4 Evaluación de los algoritmos*

A continuación se hace la evaluación de los algoritmos de WEKA a través de las variables Accuracy, *Precision, FP_Rate, TP_Rate y Recall.*

## 5.4.1 Accuracy

A continuación se presentan tablas ordenadas con base al ***desempeño general*** mostrado por cada algoritmo en cada caso. (*véase* ***Anexo 8*** *para más detalles*).

### *CasoTodasInstancias*

| Algoritmos | Accuracy (%) |
|---|---|
| bayes.AODE | 80.60 |
| bayes.AODEsr | 80.60 |
| bayes.BayesNet | 80.60 |
| bayes.HNB | 80.60 |
| bayes.NaiveBayes | 80.60 |
| bayes.NaiveBayesSimple | 80.60 |
| NaiveBayesUpdateable | 80.60 |
| bayes.WAODE | 80.60 |
| rules.DecisionTable | 80.60 |
| lazy.IBk | 80.60 |
| lazy.LBR | 80.60 |
| trees.Id3 | 80.60 |
| trees.J48 | 80.60 |
| trees.RandomForest | 80.60 |
| functions.Logistic | 80.60 |
| functions.SimpleLogistic | 80.60 |
| trees.REPTree | 80.58 |
| functions.RBFNetwork | 80.49 |
| functions.MultilayerPerceptron | 79.87 |
| functions.VotedPerceptron | 79.64 |
| rules.ConjunctiveRule | 79.09 |
| rules.OneR | 79.09 |
| lazy.LWL | 79.09 |
| trees.DecisionStump | 79.09 |
| functions.SMO | 79.09 |
| lazy.KStar | 78.35 |
| misc.VFI | 73.61 |
| rules.ZeroR | 73.11 |
| lazy.IB1 | 70.83 |
| functions.Winnow | 39.16 |
| rules.Prism | 26.88 |
| misc.HyperPipes | 26.88 |

Como podemos observar claramente, para el caso en que el número de las instancias positivas es menor a las instancias negativas (para el **CasoTodasInstancias** 3676 instancias positivas por 10000 instancias negativas) los mejores clasificadores son sin duda todos los de bayes. Sin embargo, también muestran buen desempeño sólo algunos otros basados en reglas, en árboles de decisión y en aprendizaje al momento de hacer predicción.

***Tabla 1. Resultados de los algoritmos de WEKA para el CasoTodasInstancias***

### *Caso-3675-Heterogéneo*

| Algoritmos | Accuracy (%) |
|---|---|
| misc.VFI | 73.64 |
| bayes.AODE | 73.62 |
| bayes.AODEsr | 73.62 |
| bayes.BayesNet | 73.62 |
| bayes.HNB | 73.62 |
| bayes.NaiveBayes | 73.62 |
| bayes.NaiveBayesSimple | 73.62 |
| NaiveBayesUpdateable | 73.62 |
| bayes.WAODE | 73.62 |
| lazy.LBR | 73.62 |
| lazy.LWL | 73.46 |
| functions.SMO | 73.46 |
| functions.MultilayerPerceptron | 73.45 |
| functions.VotedPerceptron | 73.42 |
| lazy.IBk | 73.38 |
| trees.Id3 | 73.38 |
| functions.RBFNetwork | 73.30 |
| functions.SimpleLogistic | 73.28 |
| trees.REPTree | 73.26 |
| functions.Logistic | 73.24 |
| lazy.KStar | 73.23 |
| rules.DecisionTable | 73.22 |
| trees.RandomForest | 73.22 |
| trees.J48 | 73.09 |
| rules.ConjunctiveRule | 65.78 |
| rules.OneR | 65.78 |
| trees.DecisionStump | 65.78 |
| lazy.IB1 | 60.02 |
| rules.ZeroR | 50.00 |
| rules.Prism | 50.00 |
| misc.HyperPipes | 50.00 |
| functions.Winnow | 49.21 |

Para el **Caso-3675-Heterogéneo,** que es cuando el número de instancias positivas y negativas es igual (3675 para nuestros ejemplos), todos los algoritmos de bayes siguen siendo muy buenos. Sin embargo, sorpresivamente es aún mejor un algoritmo que no entra en ninguna de las clasificaciones de los demás algoritmos de acuerdo a su filosofía; se trata del algoritmo VFI. También mostró buen desempeño el algoritmo lazy.LBR. (*véase Ilustración 1 para mayor claridad*).

***Tabla 2. Resultados de los algoritmos de WEKA para el Caso-3675-Heterogéneo***

### *Caso-2833-Homogéneo*

| Algoritmos | Accuracy (%) |
|---|---|
| trees.J48 | 72.50 |
| lazy.KStar | 72.48 |
| lazy.LBR | 72.36 |
| trees.RandomForest | 72.27 |
| lazy.IBk | 72.26 |
| trees.Id3 | 72.26 |
| trees.REPTree | 72.22 |
| functions.VotedPerceptron | 72.10 |
| rules.DecisionTable | 72.01 |
| functions.Logistic | 72.01 |
| functions.SimpleLogistic | 71.96 |
| bayes.AODE | 71.94 |
| bayes.AODEsr | 71.94 |
| bayes.WAODE | 71.76 |
| functions.SMO | 71.44 |
| functions.MultilayerPerceptron | 71.28 |
| bayes.BayesNet | 71.11 |
| bayes.NaiveBayes | 71.11 |
| bayes.NaiveBayesSimple | 71.11 |
| NaiveBayesUpdateable | 71.11 |
| misc.VFI | 71.11 |
| functions.RBFNetwork | 70.37 |
| bayes.HNB | 70.35 |
| rules.OneR | 68.71 |
| rules.ConjunctiveRule | 68.06 |
| trees.DecisionStump | 68.06 |
| lazy.LWL | 67.56 |
| lazy.IB1 | 63.33 |
| functions.Winnow | 53.76 |
| rules.ZeroR | 50.00 |
| rules.Prism | 50.00 |
| misc.HyperPipes | 50.00 |

Para el **Caso-2833-Homogéneo,** que es cuando el número de instancias positivas es igual al de las instancias negativas (2833 para nuestro ejemplo), pero a diferencia del Caso-3675-Heterogéneo, todas las instancias tanto positivas como negativas presentan al menos un atributo humorístico; el resultado es bastante sorpresivo al percatarnos que los algoritmos de bayes, que habían mostrado buen desempeño en los casos anteriores, ahora se quedan muy lejos y por debajo de todo los demás clasificadores. Siendo ahora los mejores uno basado en árboles de decisión y otro en aprendizaje al momento de hacer predicción. (*véase Ilustración 2 para mayor claridad*).

***Tabla 3. Resultados de los algoritmos de WEKA para el Caso-2833-Homogéneo***

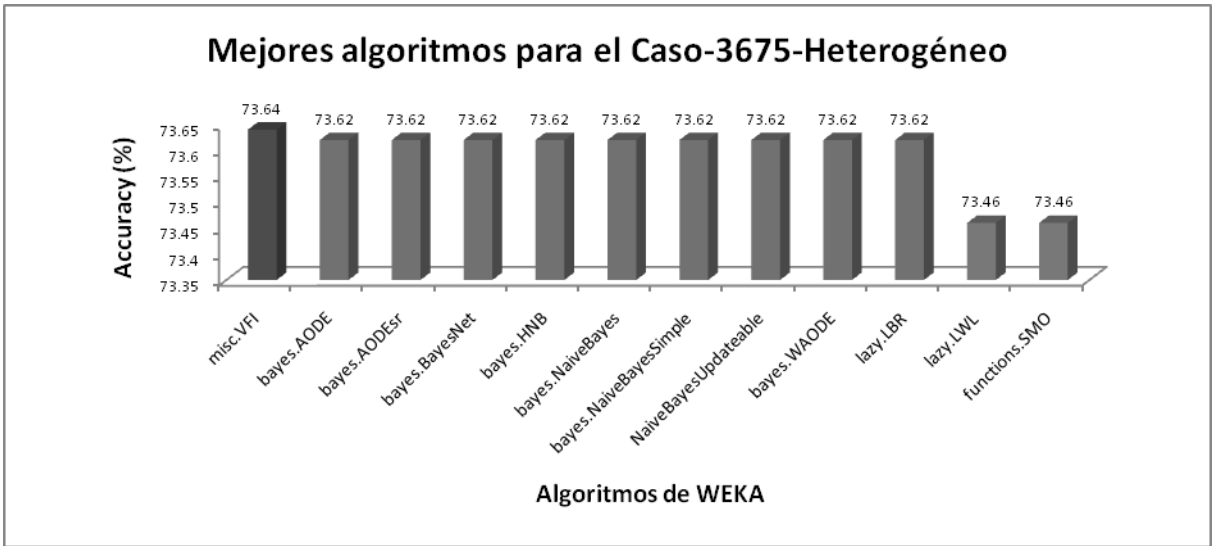


***Ilustración 1. Gráfica de mejores algoritmos de WEKA para el Caso-3675-Heterogéneo.***

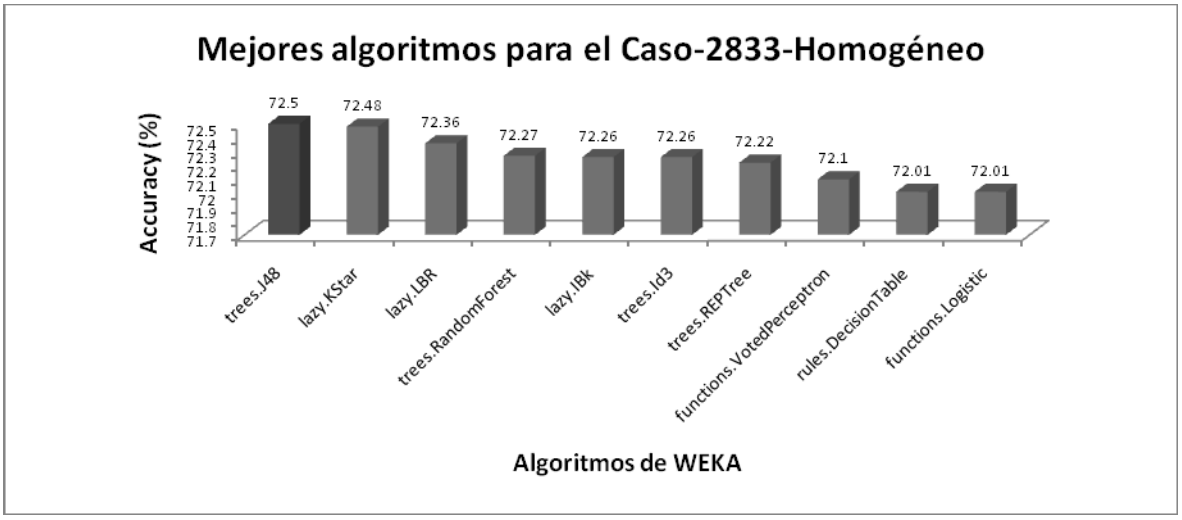


***Ilustración 2. Gráfica de mejores algoritmos de WEKA para el Caso-2833-Homogéneo.***

Sin embargo, estos resultados de los algoritmos sólo nos muestran su precisión en términos generales. Si deseamos analizar cómo le va a cada uno clasificando humor, tenemos que prestar atención a las demás medidas.

## *5.4.2 Precision*

**CasoTodasInstancias:** *(Ver Anexo 5)*:

| Algoritmos | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| misc.VFI | 2665 | 0.725 | 0.26 | 0.507 |
| functions.Winnow | 2378 | 0.647 | 0.702 | 0.253 |
| lazy.IB1 | 1721 | 0.468 | 0.203 | 0.458 |
| ... | ... | ... | ... | ... |
| lazy.KStar | 865 | 0.235 | 0.015 | 0.853 |

***Tabla 4. Resumen de Anexo 5 para la variable Precision.***

Resulta interesante observar que el algoritmo con mayor *Precision* (*lazy.KStar, con 0.853*) resultó ser el que menos instancias positivas pudo clasificar; sin embargo, lo alto del valor obedece a que también en la clase Humorística de ese algoritmo se clasificaron pocos elementos negativos. Por el contrario, el valor de P*recision* más baja (*functions.Winnow*) resultó ser el segundo mejor en la lista; esto se debe a que si bien logró clasificar correctamente una gran cantidad de instancias positivas, también clasificó de manera errónea la mayoría de las instancias negativas. Curiosamente, el algoritmo que presentó el promedio en *Precision* (*misc.VFI, con 0.507*), fue el que más acertó en la clasificación de instancias positivas; seguido también por otro cuyo valor está cercano al promedio (*lazy.IB1, con 0.458*).

**Caso-3675-Heterogéneo:** *(Ver Anexo 6)*:

| Algoritmo | Clase Humorístico | TP_Rate | FP_Rate | Precision |
|---|---|---|---|---|
| functions.Winnow | 3189 | 0.867 | 0.883 | 0.496 |
| ... | ... | ... | ... | ... |
| rules.ConjunctiveRule | 1361 | 0.37 | 0.054 | 0.872 |
| rules.OneR | 1361 | 0.37 | 0.054 | 0.872 |
| trees.DecisionStump | 1361 | 0.37 | 0.054 | 0.872 |

***Tabla 5. Resumen Anexo 6 para la variable Precision***

En este caso también se observa que el algoritmo con *Precision* promedio (*functions.Winnow, con 0.496*), es el que tuvo más aciertos en la clasificación de instancias positivas, en tanto que nuevamente los algoritmos con mayor *Precision* fueron los menos acertados.

**Caso-2833-Homogéneo:** *(Ver Anexo 7)*:

| Algoritmos | Clase Humorístico | TP_Rate | FP_Rate | Precision |
|---|---|---|---|---|
| functions.SMO | 2408 | 0.85 | 0.421 | 0.669 |
| ... | ... | ... | ... | ... |
| functions.Winnow | 1773 | 0.626 | 0.55 | 0.532 |
| ... | ... | ... | ... | ... |
| rules.ConjunctiveRule | 1361 | 0.48 | 0.119 | 0.802 |
| trees.DecisionStump | 1361 | 0.48 | 0.119 | 0.802 |

***Tabla 6. Resumen de Anexo 7 para la variable Precision***

Para este caso, el algoritmo con *Precision* más cercano al promedio (*functions. Winnow, con 0.532*) no fue el mejor clasificando. Sin embargo se repite la constante en cuanto a *Precision* más alta, es peor clasificando.

## *5.4.3 False Positive Rate*

**CasoTodasInstancias:** *(Ver Anexo 5)*:

| Algoritmos | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| misc.VFI | 2665 | 0.725 | 0.26 | 0.507 |
| functions.Winnow | 2378 | 0.647 | 0.702 | 0.253 |
| ... | ... | ... | ... | ... |
| lazy.KStar | 865 | 0.235 | 0.015 | 0.853 |

***Tabla 7. Resumen del Anexo 5 para la variable False Positive Rate.***

Para este caso y para esta variable los algoritmos con mayor *FP_Rate* resultaron los que mejor clasificaron las instancias positivas; en tanto que el que tuvo menor valor de *FP_Rate* fue el que acertó en menos instancias a la hora de clasificar.

**Caso-3675-Heterogéneo:** *(Ver Anexo 6)*:

| Algoritmo | Clase Humorístico | TP_Rate | FP_Rate | Precision |
|---|---|---|---|---|
| functions.Winnow | 3189 | 0.867 | 0.883 | 0.496 |
| misc.VFI | 2651 | 0.721 | 0.248 | 0.744 |
| ... | ... | ... | ... | ... |
| lazy.IB1 | 2378 | 0.647 | 0.446 | 0.592 |
| rules.ConjunctiveRule | 1361 | 0.37 | 0.054 | 0.872 |
| rules.OneR | 1361 | 0.37 | 0.054 | 0.872 |
| trees.DecisionStump | 1361 | 0.37 | 0.054 | 0.872 |

***Tabla 8. Resumen de Anexo 6 para la variable FP_Rate.***

Para este caso también se observa una alta proporción del *FP_Rate* con respecto a la acertada clasificación de las instancias positivas; ocurriendo una situación extraña, para estos comportamientos de la variable, con respecto al algoritmo (*lazy.IB1, con 0.446*), que a pesar de tener un *FP_Rate* alto, su precisión al momento de clasificar no fue bueno.

**Caso-2833-Homogéneo:** *(Ver Anexo 7)*:

| Algoritmos | Clase Humorístico | TP_Rate | FP_Rate | Precision |
|---|---|---|---|---|
| functions.SMO | 2408 | 0.85 | 0.421 | 0.669 |
| functions.Logistic | 2383 | 0.841 | 0.401 | 0.677 |
| ... | ... | ... | ... | ... |
| lazy.IB1 | 1906 | 0.673 | 0.406 | 0.624 |
| ... | ... | ... | ... | ... |
| rules.ConjunctiveRule | 1361 | 0.48 | 0.119 | 0.802 |
| trees.DecisionStump | 1361 | 0.48 | 0.119 | 0.802 |

***Tabla 9. Resumen del Anexo 7 para la variable FP_Rate.***

Para este último caso, se vuelve a corroborar que el *FP_Rate* alto en los algoritmos representa una mejor clasificación de las instancias positivas.

## *5.4.4 True Positive Rate y Recall*

En todos los algoritmos las variables TP_Rate y Recall, están representadas por los mismo valores, de modo que se tratarán de manera conjunta.

**CasoTodasInstancias:** *(Ver Anexo 5)*:

| Algoritmos | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| misc.VFI | 2665 | 0.725 | 0.26 | 0.507 |
| functions.Winnow | 2378 | 0.647 | 0.702 | 0.253 |
| lazy.IB1 | 1721 | 0.468 | 0.203 | 0.458 |

***Tabla 10. Resume del Anexo 5 para las variables TP_Rate y Recall***

**Caso-3675-Heterogéneo:** *(Ver Anexo 6)*:

| Algoritmo | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| functions.Winnow | 3189 | 0.867 | 0.883 | 0.496 |
| misc.VFI | 2651 | 0.721 | 0.248 | 0.744 |
| bayes.AODE | 2646 | 0.72 | 0.247 | 0.745 |

***Tabla 11. Resumen del Anexo 6 para las variables TP_Rate y Recall.***

**Caso-2833-Homogéneo:** *(Ver Anexo 7)*:

| Algoritmos | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| functions.SMO | 2408 | 0.85 | 0.421 | 0.669 |
| functions.Logistic | 2383 | 0.841 | 0.401 | 0.677 |
| functions.VotedPerceptron | 2379 | 0.839 | 0.397 | 0.679 |

***Tabla 12. Resumen del Anexo 7 para las variables TP_Rate y Recall.***

De manera resumida, podemos observar que las medidas *TP_Rate* y *Recall,* reflejan claramente y de manera inequívoca, el número de aciertos en la clasificación de las instancias positivas. Cabe señalar, que con estas medidas no importa el número de instancias negativas que se clasifican en la clase Humorístico, dado que al final de cuentas lo que a nosotros nos interesa es saber cuáles algoritmos clasifican de manera correcta a nuestras instancias positivas.

A continuación, en el último capítulo, las conclusiones finales de este trabajo.

# Capítulo 6. Discusión y Conclusiones

En este capítulo se presentan las conclusiones finales producto de esta tesis, así como sus alcances, limitaciones, aportaciones y trabajo futuro.

## *6.1 Discusión*

Hoy en día existen pocos trabajos de detección de humor en textos utilizando herramientas estadísticas computacionales. Los existentes han sido realizados para el idioma inglés; y por consiguiente, los corpus usados son totalmente diferentes a los ocupados para este trabajo en cuanto a tamaño y estructura de su contenido; aunado a que el humor es tan diverso, que es prácticamente imposible tener dos corpus en dos idiomas diferentes que se puedan considerar homogéneos. Además de que no se tiene la certeza de los tipos de algoritmos que se utilizaron para tales investigaciones. Por tal motivo, resulta inconveniente pretender comparar los resultados de esta tesis con los ya existentes hasta hoy. No obstante, estos resultados se presentan como un nuevo parámetro para nuevas investigaciones.

Tras haber experimentado con los diferentes algoritmos de WEKA y de haber analizado los diversos resultados; es posible decir ahora, que la diversidad en cuanto a número y características de los conjuntos de datos con los que se entrena a WEKA influye de manera directa en la eficiencia que muestran los algoritmos.

Esto es razonable, dado que al cambiar las características del conjunto de datos se tienen también diferentes vectores, y como cada grupo de algoritmos tiene su propia filosofía, se desempeñan de manera diferente en cada situación.

En la siguiente tabla se muestra de manera resumida los doce mejores algoritmos para las categorías de Desempeño General y Cantidad de textos humorísticos clasificados correctamente, en los tres casos. Cada grupo de algoritmos está resaltado con colores diferentes:

| **Desempeño General** | | | **Cantidad de textos humorísticos clasificados correctamente** | | |
|---|---|---|---|---|---|
| **CasoTodas Instancias** | **Caso-3675-Heterogéneo** | **Caso-2833-Homogéneo** | **CasoTodas Instancias** | **Caso-3675-Heterogéneo** | **Caso-2833-Homogéneo** |
| bayes.AODE | misc.VFI | trees.J48 | misc.VFI | functions.Winnow | functions.SMO |
| bayes.AODEsr | bayes.AODE | lazy.KStar | functions.Winnow | misc.VFI | functions.Logistic |
| bayes.BayesNet | bayes.AODEsr | lazy.LBR | lazy.IB1 | bayes.AODE | functions. VotedPerceptron |
| bayes.HNB | bayes.BayesNet | trees. RandomForest | functions.Multilayer Perceptron | bayes.AODEsr | bayes.WAODE |
| bayes. NaiveBayes | bayes.HNB | lazy.IBk | functions. VotedPerceptron | bayes.BayesNet | functions. SimpleLogistic |
| bayes.NaiveBayes Simple | bayes. NaiveBayes | trees.Id3 | functions. RBFNetwork | bayes.HNB | bayes.AODE |
| NaiveBayes Updateable | bayes.Naive BayesSimple | trees.REPTree | bayes.AODE | bayes.NaiveBayes | bayes.AODEsr |
| bayes. WAODE | NaiveBayes Updateable | functions.Voted Perceptron | bayes.AODEsr | bayes.Naive BayesSimple | trees.REPTree |
| rules. DecisionTable | bayes.WAODE | rules. DecisionTable | bayes.BayesNet | bayes.NaiveBayes Updateable | trees.J48 |
| lazy.IBk | lazy.LBR | functions. Logistic | bayes.HNB | bayes.WAODE | rules. DecisionTable |
| lazy.LBR | lazy.LWL | functions. SimpleLogistic | bayes.NaiveBayes | lazy.LBR | bayes.BayesNet |
| trees.Id3 | functions. SMO | bayes.AODE | bayes.Naive BayesSimple | lazy.IBk | bayes.NaiveBayes |

***Tabla 1. Desempeño de los algoritmos de WEKA en los diferentes casos.***

La categoría *Desempeño General,* de la tabla anterior, hace alusión a las instancias que fueron clasificadas correctamente tanto de los textos humorísticos como de los no humorísticos. En tanto que la categoría *Cantidad de textos humorísticos clasificados correctamente,* sólo hace alusión a las instancias positivas que fueron clasificadas correctamente como textos humorísticos.

De dicha tabla podemos hacer tres observaciones importantes:

1. Que los algoritmos de bayes son los más consistentes, ya que aparecen en las seis clasificaciones. De ellos, *bayes.AODE* es el único algoritmo de bayes que aparece en dichas 6 categorías y lo hace además como primer lugar de los algoritmos de su grupo en 5 de ellas.

2. El algoritmo lazy.LBR también demostró ser un algoritmo consistente al aparecer en 4 de las 6 categorías.

3. El algoritmo *misc.VFI,* aunque sólo aparece en 3 de las 6 categorías (al igual que muchos otros), lo hace dos veces en el primer lugar y una en el segundo lugar.

Por otro lado, la medida *Precision* demostró ser un buen indicador del desempeño de los clasificadores para la categoría *Cantidad de textos humorísticos clasificados correctamente*, cuando éste ronda el 0.5, no así para cuando su valor es cercano a 1 ó 0. (*véanse Anexos 5, 6 y 7*).

Por su parte *FP_Rate,* también demostró ser un buen indicador del desempeño de los algoritmos para la clase *Cantidad de textos humorísticos clasificados correctamente,* cuando su valor es cercano al 1. (*véanse Anexos 5, 6 y 7*).

De igual manera *TP_Rate* y *Recall,* las cuales representan la proporción de instancias positivas clasificadas correctamente de la categoría ya mencionada. (*véanse Anexos 5, 6 y 7*).

## *6.2 Alcances*

Este proyecto es uno más de los intentos por detectar el muy complejo y variado humor. Y aunque se está todavía muy lejos de llegar a comprenderlo, detectarlo y generarlo de manera infalible, este trabajo representa un pequeño paso que puede servir de inspiración a futuras generaciones. Es por ello que sus alcances todavía son cortos y sus limitaciones muy remarcadas.

Una de la mayores limitantes que presenta este trabajo es que no implementa la antonimia, un rasgo esencial presente en el humor; y esto es porque para el idioma español no se cuenta con un diccionario electrónico completo que pudiese ser utilizado para éste fin.

Otra de las limitantes importantes es que es estático en cuanto al cálculo de la aliteración y de la rima, lo cual reduce en gran medida su eficacia y la veracidad de los resultados.

Para el caso del albur, no implementa el cálculo para cuando existen juegos de palabras ni semejanza homófona.

Una más de las limitaciones de este proyecto de tesis es que no de todos los textos positivos registrados es posible obtener sus atributos, ya que éstos no presentan ninguno de los atributos aquí manejados, como son la aliteración, la rima, el contenido adulto o el albur; sino que están presentes como plantillas u otro tipo de juegos de palabras no contemplados para este proyecto.

Y por último, otra de las limitaciones del proyecto es la cantidad de ejemplos positivos o textos humorísticos, pues éstos están por debajo de los cinco mil ejemplos, en contraste con los diez mil ejemplos negativos o textos no humorísticos; lo que hace que los resultados no puedan ser aún más óptimos.

## *6.3 Aportaciones*

Se ha comprobado que los algoritmos de minería de datos de WEKA se pueden utilizar para la clasificación y, por consiguiente, para la detección de textos

humorísticos también para el idioma español utilizando los rasgos *rima, aliteración* y *contenido adulto,* propuestos por Rada Mihalcea y por Salvatore Attardo [16].

Se ha elaborado una herramienta de procesamiento de textos cortos para el idioma español que permite encontrar algunos de los rasgos del humor como son: *rima, aliteración* y *contenido adulto,* así como un módulo de detección de la expresión humorística mexicana por excelencia, *el albur.*

Se ha encontrado que existen otros algoritmos diferentes a los de Bayes y SVM, utilizados por Rada Mihalcea, que son también muy útiles y quizás más eficientes a la hora de detectar textos humorísticos, los cuales son: *bayes.AODE, lazy,LBR y misc.VFI.*

## *6.4 Trabajo futuro*

El trabajo futuro, sin duda alguna, es extender los alcances de la detección del humor para el idioma español; ya que hasta este momento, para el caso de la rima sólo se calcula para los tres últimos caracteres, haciéndolo un cálculo estático. Y no sólo se pretende hacerlo dinámico, sino que se pretende que contemple también el cálculo de rima asonante.

Está contemplado también mantener actualizados los diccionarios léxicos de los términos utilizados para detectar el contenido adulto y el albur, y para este último, se pretende hacer un módulo de juegos de palabras que permita detectar los albures más complicados de acuerdo con su estructura.

Por último, se pretende continuar con la ampliación de la colección de textos tanto humorísticos como no humorísticos para garantizar resultados más confiables y veraces.

## *6.5 Conclusiones*

Se pudo demostrar la hipótesis planteada al principio que sugiere que es posible utilizar para la detección de textos humorísticos en español los atributos de *rima, aliteración* y *contenido adulto,* que propone Rada para el idioma inglés, utilizando los algoritmos del programa WEKA.

Son importantes los esfuerzos que se hacen para detectar el *albur*; sin embargo, aún dista mucho de poderse detectar gran parte del mismo, dado que muchas veces el albur se presenta como juegos de palabras muy complejos que todavía no está dentro de los alcances de la Lingüística poder detectar eficientemente.

Además se descubrió que para algunas situaciones en conjuntos de datos en el idioma español los algoritmos Naïve Bayes y SVM utilizados por Rada, no son los más eficientes a la hora de hacer una clasificación.

Por último, y con base en los resultados obtenidos en este trabajo de tesis, podemos proponer a los algoritmos ***bayes.AODE, lazy.LBR*** y ***misc.VFI*** como los mejores algoritmos para la detección de humor en textos cortos en español.

# Referencias


[1]. Página oficial de Nueva Acrópolis. Organización internacional de carácter filosófico, cultural y social, con sede en España. www.nueva-acropolis.es/filosofia/humor/buen-humor.htm.

[2]. Jáuregui E. (2007). “El sentido del humor”, Ed. RBA.

[3]. Chazenbalk L. (2006). “El valor del humor en el proceso psicoterapéutico”, Universidad de Palermo.

[4]. Bergson, H. (1900). “La risa”.

[5]. Schopenhauer, A. (1987). “*El mundo como voluntad y representación”*. Introducción de E. Friedrich Sauer. Editorial Porrúa - México. ISBN 968-432-886-9.

[6]. Mihalcea, R., Strapparava, C. (2005). “Making Computers Laugh: Investigations in Automatic Humor Recognition”.

[7]. Página oficial de la Real Academia Española. Diccionario de la lengua española 22ª edición. www.rae.es/rae.html.

[8]. Hernández, V. (2006). “Antologia del Albur”, ISBN: 1-4196-2447-4.

[9]. Página oficial de WordReference.com, Vienna, Virginia, USA. www.wordreference.com.

[10]. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Y Witten, I. H. (2009); "The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1".

[11]. García, M. D. (n.d.) "Técnicas de automatización para el procesado de señales biomédicas basadas en métodos de aprendizaje máquina". Universidad Carlos III de Madrid.

[12]. Vallejos, S. (2006). "Minería de Datos", Universidad Nacional del Nordeste, Argentina.

[13]. Hoffmann, E., Chamie M. (n.d.), "Standard Statistical Classifications: Basic Principles", http://unstats.un.org//unsd//class//family//default.htm.

[14]. Acuña, E. (2008). "Análisis de Regresión", Universidad de Puerto Rico.

[15]. Vélez, I. (October 20, 2003). "Sequential Decisions: Decision Trees (Árboles de Decisión)", Available at http://ssrn.com/abstract=986975.

[16]. Mihalcea, R., Strapparava, C. (October 2005). "Making Computers Laugh: Investigations in Automatic Humor Recognition", Vancouver, Canada.

[17]. Raskin, V. (1984). "Semantic Mechanisms of Humor", ISBN 978-90-277-1821-1.

[18]. De Gruyter, W. (1994). "Lingüistic Theories of Humor", ISBN 3110142554.

[19]. Michelle, J. (2004). "Computational Recognition of Humor in a Focused Domain".

[20]. Mihalcea, R., Attardo, S. (2005). “Making Computers Laugh”.

[21]. Mihalcea, R., Pulman, S. (2007). “Characterizing Humor: An exploration of Features in Humorous Texts”.

[22]. Sjöbergh, J., Araki, K. (2007). “Recognizing Humor Without Recognizing Meaning”.

[23]. Reyes, A., Buscaldi, D., Rosso, P. (n.d). “The Impact of Semantic and Morphosyntactic Ambiguity on Automatic Humor Recognition”.

[24]. Mihalcea, R., Strapparava, C., Pulman, S. (2010). “Computational Models for Incongruity Detection in Humour”.

[25]. Páginas web de las que se recolectaron los textos humorísticos: http://www.loschistes.com, http://www.chistes.com/LosMejores.asp, http://www.estudiantes.info/chistes/lepe/chistes_lepe_1.htm, http://www.buenos-chistes.com/chistes-de-Abogados_1.html, http://historico.portalmix.com/chistes/chi_chistes_tamanyo_L_0_S.shtml, http://lapaginadelgordo.galeon.com/chistes1.html, http://www.todohistorietas.com.ar/chistesgalle2.htm, http://foro.univision.com/t5/Comunidad-de-El-Salvador/MI-RETO-E-MIL-CHISTES-CORTOS/m-p/156995157

[26]. Diccionarios de términos sexuales:

http://www.potenciasexual.com/diccionario_sexual.html,

http://www.desahogate.net/el_original/educacion-sexual/diccionario-de-terminos-de-sexualidad-y-parafilias-120747.html

[27]. Página oficial de donde se recolectaron los dichos: Copyright (c) Valentin Anders www.dechile.net

[28]. Página del concepto de Máquinas de Soporte Vectorial: http://es.wikipedia.org/wiki/Máquinas_de_vectores_de_soporte

[29]. Página del concepto de Naive Bayes: http://en.wikipedia.org/wiki/Naive_Bayes_classifier

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

[31]. Folklore: Electronic Journal of Folklore (Folklore: Electronic Journal of Folklore), issue: 33 / 2006, pages: 2758, on www.ceeol.com.

[32]. Página del concepto de Redes Neuronales: http://es.wikipedia.org/wiki/Red_neuronal_artificial.

[33]. Página del cocenpto de Sistemas Basados en Reglas de Producción: http://aprendizajeyagentes.blogspot.com/2007/05/sistemas-basados-en-reglas-de-produccin.html

[34]. Ruch, W. (n.d.). *"Compurters with a personality? Lessons to be learned from the studies of the psychology of humor".* Queens University Belfast.

[35]. Tipos de rimas: http://www.scribd.com/doc/6420492/Tipos-de-Rimas-Lenguaje-6102008

# Anexos.

## *Anexo 1. Principales características de los algoritmos de WEKA.*

Los algoritmos de WEKA que en este trabajo de tesis se utilizan, están divididos en seis grandes grupos: bayes, functions, lazy, rules, trees y misc.

Los clasificadores *bayes* están basados en teorías estadísticas de aprendizaje.

Los clasificadores *functions* agrupan algoritmos basados en SVM(*Support Vector Machines*), algoritmos de regresión y redes neuronales.

Los clasificadores *lazy* se caracterizan porque el aprendizaje en los algoritmos se realiza al momento en que se hace la predicción, como en el algoritmo k-NN (k-nearest neighbor).

Los clasificadores *rules* agrupan algoritmos de aprendizaje basados en reglas.

Los clasificadores *trees* agrupan algoritmos basados en árboles de decisión.

Por último, los clasificadores *misc* agrupan algoritmos cuyas características no recaen en ninguno de los clasificadores anteriores, pero el programa WEKA tampoco las especifica.

# *Anexo 2. Formato de un archivo .arff*

@relation conjuntodedatos

@attribute $nominal_a${$valornom_{a1}$, $valornom_{a2}$, …, $valornom_{an}$}

@attribute $nominal_b${$valornom_{b1}$, $valornom_{b2}$, …, $valornom_{bm}$}

@attribute *numerico1* real

@attribute *numerico2* real

…

@atribute clasificación{mujer, hombre}

@data

$valornom_{aj}$, $valornom_{bk}$, 20, 15, mujer

$valornom_{ak}$, $valorm_{bj}$, 20, 30, hombre

…

Donde:

*conjuntodedatos*, de la sección @relation, debe ser un nombre descriptivo de la base de datos que se está manejando.

$nominal_x$, de la sección @attribute, define los atributos nominales que se analizarán. Definiendo además un vector con el conjunto de valores nominales posibles para la variable.

*numérico*$_z$, de la sección @attribute, define los atributos numéricos permitidos.

@*data,* es una etiqueta que indica que a continuación se establecen los valores para cada uno de los atributos; éstos separados por comas y representando cada registro en un renglón por separado.

*clasificación,* de la sección @attribute, define las clases a las cuales podrán pertenecer cada uno de los registros. Útil para el proceso de predicción.

# *Anexo 3. Diccionario de Contenido Adulto*

***Advertencia:*** Las palabras aquí presentadas pueden resultar incómodas para algunas personas; se recomienda discreción.

abstinencia
anticonceptivo
argentino
bilbao
bilbaino
bisexual
callate
cama
catalan
castrado
circuncision
climax
clitoris
coito
colmo
condon
consolador
ereccion
esperma
espermatozoide
etiope
etiopia
excitacion
excitado
eyaculacion
gallego
gay
genital
hija
homosexual
impotencia
impotente
jaimito
jode
joder
kamasutra
labios
lepe
lepera
lepero
manolo
maria
marido
masturbacion
masturbado
masturbarte
masturbarse
matrimonio
matriz
menstruacion
menstrual
mujer
negro
negros
ninfomana
norteño
norteños
novia
novio
organo
orgasmo
ovario
paco
parece
parecen
patxi
pene
penetracion
pepe
pezon
politico
precoz
preservativo
procrear
procreacion
protituta
regiomontano
regiomontanos
rubia
semen
sexo
sexual
suegra
telon
testiculo
tontilandes
tontilandia
transexual
travesti
vagina
vasco
venancio
vibrador

# *Anexo 4. Diccionario de Albur*

***Advertencia:*** Las palabras aquí presentadas pueden resultar incómodas para algunas personas; se recomienda discreción.

bujero
fundillo
fundamento
chiquilin
anis
aniceto
chico
ojal
hojalatero
remolino
anillo
chiquito
chiquistriquis
yoyo
anastasio
analgesico
anaranjado
asterisco
cutis
curto
sisirisco
chiquisnais
centro
cajetearla
calabacearla
frijoles
cacahuate
miercoles
calabaza
champurrado
pedernal
pedestal
gaseosa
pedazo
pedal
oloroso
espiritu

ruidoso
ventoso
aroma
ronquido
pedro
follar
pisar
chutar
tirar
ponchar
picar
clochar
cochar
clavar
asentaderas
posaderas
pompas
nalgatorio
asiento
teleras
nachas
ignacias
anchas
aposentaderas
bote
tortas
petacas
coliseo
coliflor
chaqueta
chamarra
gabardina
manuela
jalada
puñeta
jalones
jalisco

manopla
maniobra
paja
pajuela
pito
polla
miembro
pilila
pistola
rifle
sable
minga
chipote
palo
vara
mastil
macana
garrote
pelon
pajaro
cabezon
camaron
salchicha
platano
camote
chorizo
longaniza
langosta
lancha
pilin
estaca
negra
manguera
toda
pelona
versh
verdura

verdolaga
vergüenza
vergonzoso
envergadura
verenice
vergansito
riata
delantero
chile
pescue
flaco
pirinola
leño
pepino
zanahoria
trompeta
prieta
pepito
machete
chilacayote
ciclope
tuerto
blanco
mocos
mecos
leche
yogurt
baba
crema
jocoque
ostiones
cemento
pechos
melones
montes
pechuga
chichis

chicha
senotes
monticulos
mellizas
teclas
teclado
chicharron
huevos
tompeates
tanates
cocos
cojones
campanas
testigos
bolas
canicas
cascarones
pelotas
pelo
huerfanitos
cacahuates
cuates
gemelos
aguacates
tejocotes
blanquillos
papaya
rajada
panocha
pepa
labion
triangulo
chango
simio
peludo
mono
araña

molcajete
tarantula
tesorito
chimuelo
barbona
cucaracha
panochon
peluche
eyacula
peluchin
caga
cagar
cagada
mierda
verga
culo
bragueta
chupar

# Anexo 5. Resultados del CasoTodasInstancias

3675 chistes contra 10,000 dichos. No se deben tomar en cuenta los algoritmos con TP_Rate=1, dado que clasificaron las 13,675 instancias en la clase Humorística.

| Algoritmos | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| rules.Prism | 3675 | 1 | 1 | 0.269 |
| misc.HyperPipes | 3675 | 1 | 1 | 0.269 |
| misc.VFI | 2665 | 0.725 | 0.26 | 0.507 |
| functions.Winnow | 2378 | 0.647 | 0.702 | 0.253 |
| lazy.IB1 | 1721 | 0.468 | 0.203 | 0.458 |
| functions.MultilayerPerceptron | 1612 | 0.438 | 0.069 | 0.701 |
| functions.VotedPerceptron | 1494 | 0.406 | 0.06 | 0.713 |
| functions.RBFNetwork | 1474 | 0.401 | 0.047 | 0.76 |
| bayes.AODE | 1454 | 0.395 | 0.043 | 0.772 |
| bayes.AODEsr | 1454 | 0.395 | 0.043 | 0.772 |
| bayes.BayesNet | 1454 | 0.395 | 0.043 | 0.772 |
| bayes.HNB | 1454 | 0.395 | 0.043 | 0.772 |
| bayes.NaiveBayes | 1454 | 0.395 | 0.043 | 0.772 |
| bayes.NaiveBayesSimple | 1454 | 0.395 | 0.043 | 0.772 |
| bayes.NaiveBayesUpdateable | 1454 | 0.395 | 0.043 | 0.772 |
| bayes.WAODE | 1454 | 0.395 | 0.043 | 0.772 |
| rules.DecisionTable | 1454 | 0.395 | 0.043 | 0.772 |
| lazy.IBk | 1454 | 0.395 | 0.043 | 0.772 |
| lazy.LBR | 1454 | 0.395 | 0.043 | 0.772 |
| trees.Id3 | 1454 | 0.395 | 0.043 | 0.772 |
| trees.J48 | 1454 | 0.395 | 0.043 | 0.772 |
| trees.RandomForest | 1454 | 0.395 | 0.043 | 0.772 |
| functions.Logistic | 1454 | 0.395 | 0.043 | 0.772 |
| functions.SimpleLogistic | 1454 | 0.395 | 0.043 | 0.772 |
| trees.REPTree | 1451 | 0.395 | 0.043 | 0.772 |
| rules.ConjunctiveRule | 1361 | 0.37 | 0.054 | 0.715 |
| rules.OneR | 1361 | 0.37 | 0.054 | 0.715 |
| lazy.LWL | 1361 | 0.37 | 0.054 | 0.715 |
| trees.DecisionStump | 1361 | 0.37 | 0.054 | 0.715 |
| functions.SMO | 1361 | 0.37 | 0.054 | 0.715 |
| lazy.KStar | 865 | 0.235 | 0.015 | 0.853 |
| rules.ZeroR | 0 | 0 | 0 | 0 |

# Anexo 6. Resultados del Caso-3675-Heterogéneo

3675 chistes contra 3675 dichos. No se deben tomar en cuenta los algoritmos con TP_Rate=1, dado que clasificaron las 7350 instancias en la clase Humorística.

| Algoritmo | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| rules.Prism | 3675 | 1 | 1 | 0 |
| misc.HyperPipes | 3675 | 1 | 1 | 0 |
| rules.ZeroR | 3675 | 1 | 1 | 0 |
| functions.Winnow | 3189 | 0.867 | 0.883 | 0.496 |
| misc.VFI | 2651 | 0.721 | 0.248 | 0.744 |
| bayes.AODE | 2646 | 0.72 | 0.247 | 0.745 |
| bayes.AODEsr | 2646 | 0.72 | 0.247 | 0.745 |
| bayes.BayesNet | 2646 | 0.72 | 0.247 | 0.745 |
| bayes.HNB | 2646 | 0.72 | 0.247 | 0.745 |
| bayes.NaiveBayes | 2646 | 0.72 | 0.247 | 0.745 |
| bayes.NaiveBayesSimple | 2646 | 0.72 | 0.247 | 0.745 |
| bayes.NaiveBayesUpdateable | 2646 | 0.72 | 0.247 | 0.745 |
| bayes.WAODE | 2646 | 0.72 | 0.247 | 0.745 |
| lazy.LBR | 2646 | 0.72 | 0.247 | 0.745 |
| lazy.IBk | 2629 | 0.715 | 0.247 | 0.743 |
| trees.Id3 | 2629 | 0.715 | 0.247 | 0.743 |
| functions.VotedPerceptron | 2622 | 0.713 | 0.245 | 0.745 |
| rules.DecisionTable | 2613 | 0.711 | 0.246 | 0.743 |
| trees.RandomForest | 2613 | 0.711 | 0.246 | 0.743 |
| functions.SimpleLogistic | 2612 | 0.71 | 0.245 | 0.744 |
| trees.J48 | 2604 | 0.708 | 0.246 | 0.742 |
| trees.REPTree | 2586 | 0.703 | 0.238 | 0.747 |
| functions.MultilayerPerceptron | 2586 | 0.703 | 0.234 | 0.75 |
| functions.Logistic | 2568 | 0.698 | 0.233 | 0.75 |
| functions.RBFNetwork | 2559 | 0.696 | 0.23 | 0.752 |
| lazy.LWL | 2540 | 0.691 | 0.221 | 0.757 |
| functions.SMO | 2540 | 0.691 | 0.221 | 0.757 |
| lazy.KStar | 2508 | 0.682 | 0.217 | 0.758 |
| lazy.IB1 | 2378 | 0.647 | 0.446 | 0.592 |
| rules.ConjunctiveRule | 1361 | 0.37 | 0.054 | 0.872 |
| rules.OneR | 1361 | 0.37 | 0.054 | 0.872 |
| trees.DecisionStump | 1361 | 0.37 | 0.054 | 0.872 |

# *Anexo 7. Resultados del Caso-2833-Homogéneo*

2,833 chistes contra 2,833 dichos. No se deben tomar en cuenta los algoritmos con TP_Rate=1, dado que clasificaron las 5,666 instancias en la clase Humorística.

| Algoritmos | Clase Humorístico | TP_Rate/Recall | FP_Rate | Precision |
|---|---|---|---|---|
| rules.Prism | 2833 | 1 | 1 | 0.5 |
| rules.ZeroR | 2833 | 1 | 1 | 0.5 |
| misc.HyperPipes | 2833 | 1 | 1 | 0.5 |
| functions.SMO | 2408 | 0.85 | 0.421 | 0.669 |
| functions.Logistic | 2383 | 0.841 | 0.401 | 0.677 |
| functions.VotedPerceptron | 2379 | 0.839 | 0.397 | 0.679 |
| bayes.WAODE | 2355 | 0.831 | 0.396 | 0.678 |
| functions.SimpleLogistic | 2332 | 0.823 | 0.384 | 0.682 |
| bayes.AODE | 2305 | 0.813 | 0.375 | 0.685 |
| bayes.AODEsr | 2305 | 0.813 | 0.375 | 0.685 |
| trees.REPTree | 2301 | 0.812 | 0.367 | 0.689 |
| trees.J48 | 2290 | 0.808 | 0.358 | 0.693 |
| rules.DecisionTable | 2278 | 0.804 | 0.364 | 0.689 |
| bayes.BayesNet | 2252 | 0.795 | 0.372 | 0.681 |
| bayes.NaiveBayes | 2252 | 0.795 | 0.372 | 0.681 |
| bayes.NaiveBayesSimple | 2252 | 0.795 | 0.372 | 0.681 |
| bayes.NaiveBayesUpdateable | 2252 | 0.795 | 0.372 | 0.681 |
| misc.VFI | 2252 | 0.795 | 0.372 | 0.681 |
| lazy.IBk | 2245 | 0.792 | 0.347 | 0.695 |
| trees.Id3 | 2245 | 0.792 | 0.347 | 0.695 |
| trees.RandomForest | 2239 | 0.79 | 0.345 | 0.696 |
| lazy.LBR | 2206 | 0.778 | 0.331 | 0.702 |
| lazy.KStar | 2199 | 0.776 | 0.326 | 0.704 |
| rules.OneR | 2086 | 0.736 | 0.362 | 0.671 |
| functions.RBFNetwork | 2055 | 0.725 | 0.318 | 0.695 |
| functions.MultilayerPerceptron | 2013 | 0.71 | 0.285 | 0.714 |
| lazy.IB1 | 1906 | 0.673 | 0.406 | 0.624 |
| lazy.LWL | 1780 | 0.628 | 0.277 | 0.694 |
| functions.Winnow | 1773 | 0.626 | 0.55 | 0.532 |
| bayes.HNB | 1560 | 0.55 | 0.143 | 0.793 |
| rules.ConjunctiveRule | 1361 | 0.48 | 0.119 | 0.802 |
| trees.DecisionStump | 1361 | 0.48 | 0.119 | 0.802 |

# Anexo 8. Resultados de los algoritmos de WEKA

## BAYES.AODE

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2646 1031 |    a = Humorístico
  908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4077 71.9428 %
Incorrectly Classified Instances 1590 28.0572 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.813** | 0.375 | 0.685 | 0.813 | Humorístico |
| **0.625** | 0.187 | 0.77 | 0.625 | NoHumorístico |

```
   a    b   <-- classified as
 2305  529 |    a = Humorístico
 1061 1772 |    b = NoHumorístico
```

## *BAYES.AODESR*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2646 1031 |    a = Humorístico
  908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4077 71.9428 %
Incorrectly Classified Instances 1590 28.0572 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.813** | 0.375 | 0.685 | 0.813 | Humorístico |
| **0.625** | 0.187 | 0.77 | 0.625 | NoHumorístico |

```
   a    b   <-- classified as
 2305  529 |    a = Humorístico
 1061 1772 |    b = NoHumorístico
```

## *BAYES.BAYESNET*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
1454 2223 |    a = Humorístico
 429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
2646 1031 |    a = Humorístico
 908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4030 71.1135 %
Incorrectly Classified Instances 1637 28.8865 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.795** | 0.372 | 0.681 | 0.795 | Humorístico |
| **0.628** | 0.205 | 0.753 | 0.628 | NoHumorístico |

```
   a    b   <-- classified as
2252  582 |    a = Humorístico
1055 1778 |    b = NoHumorístico
```

## *BAYES.HNB*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2646 1031 |    a = Humorístico
  908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 3987 70.3547 %
Incorrectly Classified Instances 1680 29.6453 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.55** | 0.143 | 0.793 | 0.55 | Humorístico |
| **0.857** | 0.45 | 0.656 | 0.857 | NoHumorístico |

```
   a    b   <-- classified as
 1560 1274 |    a = Humorístico
  406 2427 |    b = NoHumorístico
```

## *BAYES.NAIVEBAYES*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72Humorístico | |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2646 1031 |    a = Humorístico
  908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4030 71.1135 %
Incorrectly Classified Instances 1637 28.8865 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.795** | 0.372 | 0.681 | 0.795 | Humorístico |
| **0.628** | 0.205 | 0.753 | 0.628 | NoHumorístico |

```
   a    b   <-- classified as
 2252  582 |    a = Humorístico
 1055 1778 |    b = NoHumorístico
```

## *BAYES.NAIVEBAYESSIMPLE*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

a b <-- classified as
1454 2223 | a = Humorístico
429 9571 | b = NoHumorístico

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

a b <-- classified as
2646 1031 | a = Humorístico
908 2768 | b = NoHumorístico

### Caso-2833-Homogéneo

Correctly Classified Instances 4030 71.1135 %
Incorrectly Classified Instances 1637 28.8865 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.795** | 0.372 | 0.681 | 0.795 | Humorístico |
| **0.628** | 0.205 | 0.753 | 0.628 | NoHumorístico |

a b <-- classified as
2252 582 | a = Humorístico
1055 1778 | b = NoHumorístico

## *BAYES.NAIVEBAYESUPDATEABLE*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
1454 2223 |    a = Humorístico
 429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
2646 1031 |    a = Humorístico
 908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4030 71.1135 %
Incorrectly Classified Instances 1637 28.8865 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.795** | 0.372 | 0.681 | 0.795 | Humorístico |
| **0.628** | 0.205 | 0.753 | 0.628 | NoHumorístico |

```
   a    b   <-- classified as
2252  582 |    a = Humorístico
1055 1778 |    b = NoHumorístico
```

## *BAYES.WAODE*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2646 1031 |    a = Humorístico
  908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4067 71.7664 %
Incorrectly Classified Instances 1600 28.2336 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.831** | 0.396 | 0.678 | 0.831 | Humorístico |
| **0.604** | 0.169 | 0.781 | 0.604 | NoHumorístico |

```
   a    b   <-- classified as
 2355  479 |    a = Humorístico
 1121 1712 |    b = NoHumorístico
```

## *RULES.CONJUNCTIVERULE*

### CasoTodasInstancias

Correctly Classified Instances 10818 79.0963 %
Incorrectly Classified Instances 2859 20.9037 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.715 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.803 | 0.946 | NoHumorístico |

```
   a    b   <-- classified as
1361 2316 |    a = Humorístico
 543 9457 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 4837 65.7827 %
Incorrectly Classified Instances 2516 34.2173 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.872 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.6 | 0.946 | NoHumorístico |

```
   a    b   <-- classified as
1361 2316 |    a = Humorístico
 200 3476 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 3857 68.0607 %
Incorrectly Classified Instances 1810 31.9393 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.48** | 0.119 | 0.802 | 0.48 | Humorístico |
| **0.881** | 0.52 | 0.629 | 0.881 | NoHumorístico |

```
   a    b   <-- classified as
1361 1473 |    a = Humorístico
 337 2496 |    b = NoHumorístico
```

## *RULES.DECISIONTABLE*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5384 73.2218 %
Incorrectly Classified Instances 1969 26.7782 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.711** | 0.246 | 0.743 | 0.711 | Humorístico |
| **0.754** | 0.289 | 0.723 | 0.754 | NoHumorístico |

```
   a    b   <-- classified as
 2613 1064 |    a = Humorístico
  905 2771 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4081 72.0134 %
Incorrectly Classified Instances 1586 27.9866 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.804** | 0.364 | 0.689 | 0.804 | Humorístico |
| **0.636** | 0.196 | 0.764 | 0.636 | NoHumorístico |

```
   a    b   <-- classified as
 2278  556 |    a = Humorístico
 1030 1803 |    b = NoHumorístico
```

## *RULES.ONER*

### CasoTodasInstancias

Correctly Classified Instances 10818 79.0963 %
Incorrectly Classified Instances 2859 20.9037 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.715 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.803 | 0.946 | NoHumorístico |

```
   a    b   <-- classified as
1361 2316 |    a = Humorístico
 543 9457 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 4837 65.7827 %
Incorrectly Classified Instances 2516 34.2173 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.872 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.6 | 0.946 | NoHumorístico |

```
   a    b   <-- classified as
1361 2316 |    a = Humorístico
 200 3476 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 3894 68.7136 %
Incorrectly Classified Instances 1773 31.2864 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.736** | 0.362 | 0.671 | 0.736 | Humorístico |
| **0.638** | 0.264 | 0.707 | 0.638 | NoHumorístico |

```
   a    b   <-- classified as
2086  748 |    a = Humorístico
1025 1808 |    b = NoHumorístico
```

## *RULES.PRISM*

### CasoTodasInstancias

Correctly Classified Instances 3677 26.8846 %
Incorrectly Classified Instances 10000 73.1154 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.269 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
   a    b   <-- classified as
 3677    0 |    a = Humorístico
10000    0 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 3677 50.0068 %
Incorrectly Classified Instances 3676 49.9932 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.5 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
   a    b   <-- classified as
3677    0 |    a = Humorístico
3676    0 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 2834 50.0088 %
Incorrectly Classified Instances 2833 49.9912 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.5 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
   a    b   <-- classified as
2834    0 |    a = Humorístico
2833    0 |    b = NoHumorístico
```

## *RULES.ZEROR*

### CasoTodasInstancias

Correctly Classified Instances 10000 73.1154 %
Incorrectly Classified Instances 3677 26.8846 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0** | 0 | 0 | 0 | Humorístico |
| **1** | 1 | 0.731 | 1 | NoHumorístico |

```
a    b   <-- classified as
0  3677 |    a = Humorístico
0 10000 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 3677 50.0068 %
Incorrectly Classified Instances 3676 49.9932 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.5 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
  a    b   <-- classified as
3677   0 |    a = Humorístico
3676   0 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 2834 50.0088 %
Incorrectly Classified Instances 2833 49.9912 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.5 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
  a    b   <-- classified as
2834   0 |    a = Humorístico
2833   0 |    b = NoHumorístico
```

## *LAZY.IB1*

### CasoTodasInstancias

Correctly Classified Instances 9688 70.8342 %
Incorrectly Classified Instances 3989 29.1658 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.468** | 0.203 | 0.458 | 0.468 | Humorístico |
| **0.797** | 0.532 | 0.803 | 0.797 | NoHumorístico |

```
   a    b   <-- classified as
 1721 1956 |    a = Humorístico
 2033 7967 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 4414 60.0299 %
Incorrectly Classified Instances 2939 39.9701 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.647** | 0.446 | 0.592 | 0.647 | Humorístico |
| **0.554** | 0.353 | 0.61 | 0.554 | NoHumorístico |

```
   a    b   <-- classified as
 2378 1299 |    a = Humorístico
 1640 2036 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 3589 63.3316 %
Incorrectly Classified Instances 2078 36.6684 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.673** | 0.406 | 0.624 | 0.673 | Humorístico |
| **0.594** | 0.327 | 0.645 | 0.594 | NoHumorístico |

```
   a    b   <-- classified as
 1906  928 |    a = Humorístico
 1150 1683 |    b = NoHumorístico
```

## *LAZY.IBK*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5396 73.385 %
Incorrectly Classified Instances 1957 26.615 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.715** | 0.247 | 0.743 | 0.715 | Humorístico |
| **0.753** | 0.285 | 0.725 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2629 1048 |    a = Humorístico
  909 2767 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4095 72.2605 %
Incorrectly Classified Instances 1572 27.7395 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.792** | 0.347 | 0.695 | 0.792 | Humorístico |
| **0.653** | 0.208 | 0.759 | 0.653 | NoHumorístico |

```
   a    b   <-- classified as
 2245  589 |    a = Humorístico
  983 1850 |    b = NoHumorístico
```

## *LAZY.KSTAR*

### CasoTodasInstancias

Correctly Classified Instances 10716 78.3505 %
Incorrectly Classified Instances 2961 21.6495 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.235** | 0.015 | 0.853 | 0.235 | Humorístico |
| **0.985** | 0.765 | 0.778 | 0.985 | NoHumorístico |

```
   a    b   <-- classified as
 865 2812 |    a = Humorístico
 149 9851 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5385 73.2354 %
Incorrectly Classified Instances 1968 26.7646 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.682** | 0.217 | 0.758 | 0.682 | Humorístico |
| **0.783** | 0.318 | 0.711 | 0.783 | NoHumorístico |

```
   a    b   <-- classified as
 2508 1169 |    a = Humorístico
  799 2877 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4108 72.4899 %
Incorrectly Classified Instances 1559 27.5101 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.776** | 0.326 | 0.704 | 0.776 | Humorístico |
| **0.674** | 0.224 | 0.75 | 0.674 | NoHumorístico |

```
   a    b   <-- classified as
 2199  635 |    a = Humorístico
  924 1909 |    b = NoHumorístico
```

## *LAZY.LBR*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5414 73.6298 %
Incorrectly Classified Instances 1939 26.3702 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.72** | 0.247 | 0.745 | 0.72 | Humorístico |
| **0.753** | 0.28 | 0.729 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2646 1031 |    a = Humorístico
  908 2768 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4101 72.3663 %
Incorrectly Classified Instances 1566 27.6337 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.778** | 0.331 | 0.702 | 0.778 | Humorístico |
| **0.669** | 0.222 | 0.751 | 0.669 | NoHumorístico |

```
   a    b   <-- classified as
 2206  628 |    a = Humorístico
  938 1895 |    b = NoHumorístico
```

## *LAZY.LWL*

### CasoTodasInstancias

Correctly Classified Instances 10818 79.0963 %
Incorrectly Classified Instances 2859 20.9037 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.715 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.803 | 0.946 | NoHumorístico |

a b <-- classified as
1361 2316 | a = Humorístico
543 9457 | b = NoHumorístico

### Caso-3675-Heterogéneo

Correctly Classified Instances 5402 73.4666 %
Incorrectly Classified Instances 1951 26.5334 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.691** | 0.221 | 0.757 | 0.691 | Humorístico |
| **0.779** | 0.309 | 0.716 | 0.779 | NoHumorístico |

a b <-- classified as
2540 1137 | a = Humorístico
814 2862 | b = NoHumorístico

### Caso-2833-Homogéneo

Correctly Classified Instances 3829 67.5666 %
Incorrectly Classified Instances 1838 32.4334 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.628** | 0.277 | 0.694 | 0.628 | Humorístico |
| **0.723** | 0.372 | 0.66 | 0.723 | NoHumorístico |

a b <-- classified as
1780 1054 | a = Humorístico
784 2049 | b = NoHumorístico

## *TREES.DECISIONSTUMP*

### CasoTodasInstancias

Correctly Classified Instances 10818 79.0963 %
Incorrectly Classified Instances 2859 20.9037 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.715 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.803 | 0.946 | NoHumorístico |

```
   a    b   <-- classified as
1361 2316 |    a = Humorístico
 543 9457 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 4837 65.7827 %
Incorrectly Classified Instances 2516 34.2173 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.872 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.6 | 0.946 | NoHumorístico |

```
   a    b   <-- classified as
1361 2316 |    a = Humorístico
 200 3476 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 3857 68.0607 %
Incorrectly Classified Instances 1810 31.9393 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.48** | 0.119 | 0.802 | 0.48 | Humorístico |
| **0.881** | 0.52 | 0.629 | 0.881 | NoHumorístico |

```
   a    b   <-- classified as
1361 1473 |    a = Humorístico
 337 2496 |    b = NoHumorístico
```

## *TREES.ID3*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5396 73.385 %
Incorrectly Classified Instances 1957 26.615 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.715** | 0.247 | 0.743 | 0.715 | Humorístico |
| **0.753** | 0.285 | 0.725 | 0.753 | NoHumorístico |

```
   a    b   <-- classified as
 2629 1048 |    a = Humorístico
  909 2767 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4095 72.2605 %
Incorrectly Classified Instances 1572 27.7395 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.792** | 0.347 | 0.695 | 0.792 | Humorístico |
| **0.653** | 0.208 | 0.759 | 0.653 | NoHumorístico |

```
   a    b   <-- classified as
 2245  589 |    a = Humorístico
  983 1850 |    b = NoHumorístico
```

## *TREES.J48*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
1454 2223 |    a = Humorístico
 429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5375 73.0994 %
Incorrectly Classified Instances 1978 26.9006 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.708** | 0.246 | 0.742 | 0.708 | Humorístico |
| **0.754** | 0.292 | 0.721 | 0.754 | NoHumorístico |

```
   a    b   <-- classified as
2604 1073 |    a = Humorístico
 905 2771 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4109 72.5075 %
Incorrectly Classified Instances 1558 27.4925 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.808** | 0.358 | 0.693 | 0.808 | Humorístico |
| **0.642** | 0.192 | 0.77 | 0.642 | NoHumorístico |

```
   a    b   <-- classified as
2290  544 |    a = Humorístico
1014 1819 |    b = NoHumorístico
```

## *TREES.RANDOMFOREST*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5384 73.2218 %
Incorrectly Classified Instances 1969 26.7782 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.711** | 0.246 | 0.743 | 0.711 | Humorístico |
| **0.754** | 0.289 | 0.723 | 0.754 | NoHumorístico |

```
   a    b   <-- classified as
 2613 1064 |    a = Humorístico
  905 2771 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4096 72.2781 %
Incorrectly Classified Instances 1571 27.7219 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.79** | 0.345 | 0.696 | 0.79 | Humorístico |
| **0.655** | 0.21 | 0.757 | 0.655 | NoHumorístico |

```
   a    b   <-- classified as
 2239  595 |    a = Humorístico
  976 1857 |    b = NoHumorístico
```

## *TREES.REPTREE*

### CasoTodasInstancias

Correctly Classified Instances 11022 80.5878 %
Incorrectly Classified Instances 2655 19.4122 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.811 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
1451 2226 |    a = Humorístico
 429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5387 73.2626 %
Incorrectly Classified Instances 1966 26.7374 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.703** | 0.238 | 0.747 | 0.703 | Humorístico |
| **0.762** | 0.297 | 0.72 | 0.762 | NoHumorístico |

```
   a    b   <-- classified as
2586 1091 |    a = Humorístico
 875 2801 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4093 72.2252 %
Incorrectly Classified Instances 1574 27.7748 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.812** | 0.367 | 0.689 | 0.812 | Humorístico |
| **0.633** | 0.188 | 0.771 | 0.633 | NoHumorístico |

```
   a    b   <-- classified as
2301  533 |    a = Humorístico
1041 1792 |    b = NoHumorístico
```

## *FUNCTIONS.LOGISTIC*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5386 73.249 %
Incorrectly Classified Instances 1967 26.751 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.698** | 0.233 | 0.75 | 0.698 | Humorístico |
| **0.767** | 0.302 | 0.718 | 0.767 | NoHumorístico |

```
   a    b   <-- classified as
 2568 1109 |    a = Humorístico
  858 2818 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4081 72.0134 %
Incorrectly Classified Instances 1586 27.9866 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.841** | 0.401 | 0.677 | 0.841 | Humorístico |
| **0.599** | 0.159 | 0.79 | 0.599 | NoHumorístico |

```
   a    b   <-- classified as
 2383  451 |    a = Humorístico
 1135 1698 |    b = NoHumorístico
```

## *FUNCTIONS.MULTILAYERPERCEPTRON*

### CasoTodasInstancias

Correctly Classified Instances 10924 79.8713 %
Incorrectly Classified Instances 2753 20.1287 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.438** | 0.069 | 0.701 | 0.438 | Humorístico |
| **0.931** | 0.562 | 0.818 | 0.931 | NoHumorístico |

```
   a    b   <-- classified as
1612 2065 |    a = Humorístico
 688 9312 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5401 73.453 %
Incorrectly Classified Instances 1952 26.547 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.703** | 0.234 | 0.75 | 0.703 | Humorístico |
| **0.766** | 0.297 | 0.721 | 0.766 | NoHumorístico |

```
   a    b   <-- classified as
2586 1091 |    a = Humorístico
 861 2815 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4040 71.2899 %
Incorrectly Classified Instances 1627 28.7101 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.71** | 0.285 | 0.714 | 0.71 | Humorístico |
| **0.715** | 0.29 | 0.712 | 0.715 | NoHumorístico |

```
   a    b   <-- classified as
2013  821 |    a = Humorístico
 806 2027 |    b = NoHumorístico
```

## *FUNCTIONS.RBFNETWORK*

### CasoTodasInstancias

Correctly Classified Instances 11009 80.4928 %
Incorrectly Classified Instances 2668 19.5072 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.401** | 0.047 | 0.76 | 0.401 | Humorístico |
| **0.954** | 0.599 | 0.812 | 0.954 | NoHumorístico |

```
   a    b   <-- classified as
 1474 2203 |    a = Humorístico
  465 9535 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5390 73.3034 %
Incorrectly Classified Instances 1963 26.6966 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.696** | 0.23 | 0.752 | 0.696 | Humorístico |
| **0.77** | 0.304 | 0.717 | 0.77 | NoHumorístico |

```
   a    b   <-- classified as
 2559 1118 |    a = Humorístico
  845 2831 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 3986 70.337 %
Incorrectly Classified Instances 1681 29.663 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.725** | 0.318 | 0.695 | 0.725 | Humorístico |
| **0.682** | 0.275 | 0.713 | 0.682 | NoHumorístico |

```
   a    b   <-- classified as
 2055  779 |    a = Humorístico
  902 1931 |    b = NoHumorístico
```

## *FUNCTIONS.SIMPLELOGISTIC*

### CasoTodasInstancias

Correctly Classified Instances 11025 80.6098 %
Incorrectly Classified Instances 2652 19.3902 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.395** | 0.043 | 0.772 | 0.395 | Humorístico |
| **0.957** | 0.605 | 0.812 | 0.957 | NoHumorístico |

```
   a    b   <-- classified as
 1454 2223 |    a = Humorístico
  429 9571 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5389 73.2898 %
Incorrectly Classified Instances 1964 26.7102 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.71** | 0.245 | 0.744 | 0.71 | Humorístico |
| **0.755** | 0.29 | 0.723 | 0.755 | NoHumorístico |

```
   a    b   <-- classified as
 2612 1065 |    a = Humorístico
  899 2777 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4078 71.9605 %
Incorrectly Classified Instances 1589 28.0395 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.823** | 0.384 | 0.682 | 0.823 | Humorístico |
| **0.616** | 0.177 | 0.777 | 0.616 | NoHumorístico |

```
   a    b   <-- classified as
 2332  502 |    a = Humorístico
 1087 1746 |    b = NoHumorístico
```

## *FUNCTIONS.SMO*

### CasoTodasInstancias

Correctly Classified Instances 10818 79.0963 %
Incorrectly Classified Instances 2859 20.9037 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.37** | 0.054 | 0.715 | 0.37 | Humorístico |
| **0.946** | 0.63 | 0.803 | 0.946 | NoHumorístico |

```
   a    b   <-- classified as
 1361 2316 |    a = Humorístico
  543 9457 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5402 73.4666 %
Incorrectly Classified Instances 1951 26.5334 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.691** | 0.221 | 0.757 | 0.691 | Humorístico |
| **0.779** | 0.309 | 0.716 | 0.779 | NoHumorístico |

```
   a    b   <-- classified as
 2540 1137 |    a = Humorístico
  814 2862 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4049 71.4487 %
Incorrectly Classified Instances 1618 28.5513 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.85** | 0.421 | 0.669 | 0.85 | Humorístico |
| **0.579** | 0.15 | 0.794 | 0.579 | NoHumorístico |

```
   a    b   <-- classified as
 2408  426 |    a = Humorístico
 1192 1641 |    b = NoHumorístico
```

## *FUNCTIONS.VOTEDPERCEPTRON*

### CasoTodasInstancias

Correctly Classified Instances 10893 79.6447 %
Incorrectly Classified Instances 2784 20.3553 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.406** | 0.06 | 0.713 | 0.406 | Humorístico |
| **0.94** | 0.594 | 0.812 | 0.94 | NoHumorístico |

a b <-- classified as
1494 2183 | a = Humorístico
601 9399 | b = NoHumorístico

### Caso-3675-Heterogéneo

Correctly Classified Instances 5399 73.4258 %
Incorrectly Classified Instances 1954 26.5742 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.713** | 0.245 | 0.745 | 0.713 | Humorístico |
| **0.755** | 0.287 | 0.725 | 0.755 | NoHumorístico |

a b <-- classified as
2622 1055 | a = Humorístico
899 2777 | b = NoHumorístico

### Caso-2833-Homogéneo

Correctly Classified Instances 4086 72.1016 %
Incorrectly Classified Instances 1581 27.8984 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.839** | 0.397 | 0.679 | 0.839 | Humorístico |
| **0.603** | 0.161 | 0.79 | 0.603 | NoHumorístico |

a b <-- classified as
2379 455 | a = Humorístico
1126 1707 | b = NoHumorístico

## *FUNCTIONS.WINNOW*

### CasoTodasInstancias

Correctly Classified Instances 5356 39.1606 %
Incorrectly Classified Instances 8321 60.8394 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.647** | 0.702 | 0.253 | 0.647 | Humorístico |
| **0.298** | 0.353 | 0.696 | 0.298 | NoHumorístico |

```
   a    b   <-- classified as
 2378 1299 |    a = Humorístico
 7022 2978 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 3619 49.218 %
Incorrectly Classified Instances 3734 50.782 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.867** | 0.883 | 0.496 | 0.867 | Humorístico |
| **0.117** | 0.133 | 0.468 | 0.117 | NoHumorístico |

```
   a    b   <-- classified as
 3189  488 |    a = Humorístico
 3246  430 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 3047 53.7674 %
Incorrectly Classified Instances 2620 46.2326 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.626** | 0.55 | 0.532 | 0.626 | Humorístico |
| **0.45** | 0.374 | 0.546 | 0.45 | NoHumorístico |

```
   a    b   <-- classified as
 1773 1061 |    a = Humorístico
 1559 1274 |    b = NoHumorístico
```

## *MISC.HYPERPIPES*

### CasoTodasInstancias

Correctly Classified Instances 3677 26.8846 %
Incorrectly Classified Instances 10000 73.1154 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.269 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
   a    b   <-- classified as
 3677    0 |    a = Humorístico
10000    0 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 3677 50.0068 %
Incorrectly Classified Instances 3676 49.9932 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.5 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
   a    b   <-- classified as
3677    0 |    a = Humorístico
3676    0 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 2834 50.0088 %
Incorrectly Classified Instances 2833 49.9912 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **1** | 1 | 0.5 | 1 | Humorístico |
| **0** | 0 | 0 | 0 | NoHumorístico |

```
   a    b   <-- classified as
2834    0 |    a = Humorístico
2833    0 |    b = NoHumorístico
```

## *MISC.VFI*

### CasoTodasInstancias

Correctly Classified Instances 10069 73.6199 %
Incorrectly Classified Instances 3608 26.3801 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.725** | 0.26 | 0.507 | 0.725 | Humorístico |
| **0.74** | 0.275 | 0.88 | 0.74 | NoHumorístico |

```
   a    b   <-- classified as
 2665 1012 |    a = Humorístico
 2596 7404 |    b = NoHumorístico
```

### Caso-3675-Heterogéneo

Correctly Classified Instances 5415 73.6434 %
Incorrectly Classified Instances 1938 26.3566 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.721** | 0.248 | 0.744 | 0.721 | Humorístico |
| **0.752** | 0.279 | 0.729 | 0.752 | NoHumorístico |

```
   a    b   <-- classified as
 2651 1026 |    a = Humorístico
  912 2764 |    b = NoHumorístico
```

### Caso-2833-Homogéneo

Correctly Classified Instances 4030 71.1135 %
Incorrectly Classified Instances 1637 28.8865 %

| TP_Rate | FP_Rate | Precision | Recall | Class |
|---|---|---|---|---|
| **0.795** | 0.372 | 0.681 | 0.795 | Humorístico |
| **0.628** | 0.205 | 0.753 | 0.628 | NoHumorístico |

```
   a    b   <-- classified as
 2252  582 |    a = Humorístico
 1055 1778 |    b = NoHumorístico
```