

Detección de humor en textos en español

PROYECTO DE GRADO

Santiago CASTRO
Matías CUBERO

Tutores:

Dr. Guillermo MONCECCHI
MSc. Diego GARAT



Universidad de la República
Facultad de Ingeniería

Marzo de 2015



«Un día sin reír es un día perdido.»

Charles Chaplin

Resumen

Aun cuando el humor ha sido estudiado desde el punto de vista psicológico, cognitivo, y lingüístico, su estudio desde un punto de vista computacional es todavía un área a explorar en el marco de la lingüística computacional. Algunos trabajos previos existen, pero se está aún lejos de concretar una caracterización del humor que permita su reconocimiento y generación automática. En este proyecto se elabora un estado del arte del reconocimiento de humor, se construye un corpus de tweets que se etiqueta en base a votaciones de personas y se realiza un clasificador estadístico y supervisado de humor en tweets en español utilizando Support Vector Machine, k Nearest Neighbors, Árboles de decisión y dos tipos de clasificadores Naïve Bayes, llegando a una precisión de 83,61 % y un recall de 68,85 %.

Palabras clave: Humor, Humor Computacional, Reconocimiento de Humor, Aprendizaje Automático, Procesamiento del Lenguaje Natural.

Agradecimientos

A nuestras familias y amigos por el apoyo brindado durante nuestra carrera y en la elaboración de este proyecto. A todo el MuDi, por hacer siempre llevadero el proyecto y por los constantes apoyos multidisciplinarios. A nuestros trabajos, que nos bancaron en tiempos difíciles. También a todas las personas que entraron a la página o se descargaron la aplicación para votar tweets. Además agradecemos a nuestros tutores, Guillermo y Diego, por acompañarnos en la locura de intentar abarcar esta difícil temática que es encontrar el humor, y a la Agencia Nacional de Investigación e Innovación por el apoyo económico brindado. Santiago quiere agradecer también a Cecilia por el apoyo y la comprensión durante este proceso.

Índice general

1	Introducción	1
1.1	Objetivos y alcance	2
2	Estado del arte	3
2.1	Humor	3
2.1.1	Teoría de la superioridad	4
2.1.2	Teoría del alivio	4
2.1.3	Teoría de la resolución de incongruencias	4
2.1.4	Teoría de la violación	5
2.1.5	Teorías sociológicas	6
2.1.6	Teoría de los guiones semánticos del humor	6
2.1.7	Teoría general del humor verbal	7
2.2	Detección de humor	8
2.2.1	Aliteración	8
2.2.2	Ambigüedad	8
2.2.3	Antonimia	9
2.2.4	Centrado en personas	10
2.2.5	Jerga Sexual	10
2.2.6	Negatividad	10
2.2.7	Palabras clave	10
2.2.8	Perplejidad del modelo de lenguaje	11
2.3	Evaluación de los trabajos	11
3	Corpus	13
3.1	Extracción de tweets	13
3.2	Anotación del corpus	15
3.2.1	Proceso de Anotación	15
3.2.2	Criterios de anotación	17
3.2.3	Resultado de la votación	17
3.2.4	Humor según la votación	19
3.2.5	Promedio de estrellas	22
3.2.6	Mejores chistes	22
3.2.7	Concordancia entre los anotadores	23
4	Aprendizaje Automático	26
4.1	Métodos de Aprendizaje Automático	26
4.1.1	Support Vector Machine	26
4.1.2	Árbol de decisión	28

4.1.3	Naïve Bayes	29
4.1.4	k Nearest Neighbors	30
4.2	Medidas de evaluación	30
4.3	Escalado de características	34
5	Clasificador	35
5.1	Conjunto de entrenamiento/evaluación	35
5.1.1	Subcorpus	35
5.2	Línea base	36
5.3	Características	36
5.3.1	Presencia de animales	37
5.3.2	Jerga sexual	37
5.3.3	Primera persona	37
5.3.4	Segunda persona	38
5.3.5	Distancia temática	38
5.3.6	Diálogo	38
5.3.7	Preguntas-respuestas	38
5.3.8	Palabras frecuentes	38
5.3.9	Links	39
5.3.10	Antónimos	39
5.3.11	Hashtags	39
5.3.12	Exclamación	39
5.3.13	Palabras mayúsculas	39
5.3.14	Negación	39
5.3.15	Palabras fuera del vocabulario	39
5.3.16	Palabras no españolas	40
5.4	Resultados iniciales	40
5.5	Selección de características	41
5.5.1	Extremely Randomized Trees	41
5.5.2	Eliminación recursiva de atributos	41
5.6	Resultados finales	41
5.7	Evaluación en el conjunto de entrenamiento	44
5.7.1	Inconsistencias en el corpus	44
5.7.2	SVM	46
5.8	Restricción al subcorpus de cuentas de humor	46
5.8.1	Métricas ponderadas según calificación	47
5.9	Tweets censurados	48
5.10	Diferentes subconjuntos no humorísticos	48
5.11	<i>Tweets</i> dudosos	49
5.12	Independencia y combinación de características	49
6	Conclusiones	50
6.1	Trabajo futuro	51
	Glosario	52
	Apéndices	53
A	Arquitectura del clasificador	54

B Ajuste de parámetros	57
C Uso del programa clasificador	60
Bibliografía	64

Capítulo 1

Introducción

Aun cuando el humor ha sido estudiado desde el punto de vista psicológico y cognitivo (International Journal of Humor Research 1988), o lingüístico (Raskin 1985), su estudio desde un punto de vista computacional es todavía un área a explorar en el marco de la lingüística computacional. El humano, como ser social, convive a diario con el humor y por tanto se convierte en un tema a estudiar de fundamental interés para las Ciencias de la Computación. Algunos trabajos previos existen (Mulder y Nijholt 2002), pero se está aún lejos de concretar una caracterización del humor que permita su reconocimiento y generación automática. Este proyecto, en el marco del Proyecto de grado de la Facultad de Ingeniería — Universidad de la República — intenta buscar y eventualmente proponer métodos y mecanismos para el reconocimiento automático de expresiones humorísticas, en particular para tweets en el idioma español.

La risa caracteriza al ser humano como especie. El humor, que es un potencial causante de la risa, es un componente esencial en la comunicación humana. Permite que la gente no se sienta infeliz, a la vez que produce un ambiente más distendido. Es de especial interés que la interfaz con la computadora sea también lo más amigable posible, en donde el humor puede ser una pieza fundamental, teniendo en cuenta que las computadoras juegan un rol social cada vez más activo (incluyendo celulares, tablets y dispositivos inteligentes en general).

Identificar humor en un texto es un paso intermedio en la resolución de cuestiones más complicadas. Resultaría atractivo poder llegar a generar chistes, o más en general humor, en base a conocer qué atributos enriquecen en mejor medida al texto. Una etapa previa a la generación de humor puede ser entonces el reconocimiento. El hecho de aprender a generar ejemplos a partir de saber reconocerlos es una idea clásica; una vez que se sabe cómo identificarlos se tiene una idea de qué características intentar hacer cumplir a los ejemplos que uno quiere construir. Por otro lado, como uso final, este trabajo puede ser usado para encontrar chistes en Twitter, para buscar los tweets graciosos de algún tema del momento, o las respuestas graciosas a comentarios en Twitter.

Para mostrar la dificultad del problema de reconocimiento de humor, se muestra el chiste (1), que es del año 2012, y es un ejemplo de tipo de chiste que se pretende poder detectar. Al igual que muchas tareas del área Procesamiento del Lenguaje Natural, es necesario contar con cierto conocimiento sobre el mundo y la realidad. Primero, aunque parezca trivial, para interpretar el lenguaje y al chiste hay que saber que el 28 de diciembre es el Día de los inocentes (al menos en Uruguay), y también saber la teoría que afirma el fin del mundo, basándose en el calendario maya, afirmando que termina en el año 2012 de nuestro calendario. También que los mayas fueron personas y por lo tanto pueden hablar y celebrar un día. El humor está dentro de un contexto que es implícito al texto. A su vez hay que entender que el relator está diciendo que la teoría de

los mayas era una broma para el día de los inocentes. Una computadora para entender este texto debería ser capaz de darse cuenta que esto no es serio: los mayas no pueden decir nada porque no existe más. Se tiene que dar cuenta que esto no puede pasar, no es una deducción válida.

- (1) El 28 de diciembre del 2012 van a salir los mayas a decir: “¡Feliz día de los inocentes!”

1.1. Objetivos y alcance

El objetivo de este proyecto es construir un clasificador de humor en textos en español utilizando métodos de aprendizaje automático, en particular técnicas supervisadas. Como la problemática es muy variada se decide restringir el problema a solamente tweets. Para poder lograr este cometido es necesario contar con un corpus de tweets clasificados como humorísticos o no, por tanto un objetivo secundario de este trabajo es construir un corpus de tweets en español para poder trabajar.

Adicionalmente se desea saber si el aprendizaje supervisado estadístico funciona para detectar humor, ya que podría ser el caso que las técnicas utilizadas no den resultados satisfactorios. La idea es usar esta forma de aprendizaje sin involucrarse profundamente a intentar interpretar un texto, ya que como se muestra es una tarea difícil.

Una meta indirecta es hacer un relevamiento del estado del arte para entender cuál es el estado de resolución de esta tarea.

Capítulo 2

Estado del arte

El presente capítulo intenta abordar el estado del arte de la Detección de humor. De esta forma se analiza lo ya existente y se fija un punto de partida para la tarea a realizar, siendo de gran valor para las etapas siguientes del proyecto. Primero se analiza el humor como concepto, introduciendo principalmente definiciones, teorías y puntos de vista, obteniendo una visión global sobre el tema a tratar. Luego, se describen ciertas características intensamente relacionadas con el humor, que fueron sugeridas por distintos autores para ser utilizadas en la detección de humor. Por último, se analizan dos trabajos similares al proyecto a realizar.

2.1. Humor

Según el Diccionario de la Real Academia Española (2001) (DRAE), *humorismo* es un “modo de presentar, enjuiciar o comentar la realidad, resaltando el lado cómico, risueño o ridículo de las cosas”. Este trabajo se basa únicamente en aquel humor que pueda ser expresado por escrito. Por *chiste* se entiende por “dicho u ocurrencia aguda y graciosa”, aunque en este documento se usa esta palabra y la palabra *humor* de manera intercambiada, más allá que no signifiquen lo mismo. Cabe destacar también que *humor* puede causar *risa* pero no siempre.

Suelen confundirse los términos *ironía*, *sátira*, *sarcasmo* y *humor*. La *ironía*, a diferencia de los otros conceptos, según la DRAE es una “figura retórica que consiste en dar a entender lo contrario de lo que se dice”. *Sátira* en cambio es una “composición en verso o prosa cuyo objeto es censurar o ridiculizar a alguien o algo”. A esta última se la puede ver como forma de humor, aunque tiene un fin crítico, a diferencia del humor que tiene un fin cómico (causar risa). El *sarcasmo* es una “burla sangrienta, ironía mordaz y cruel con que se ofende o maltrata a alguien o algo”, teniendo un objetivo de burla. Muchas veces se mezcla al *ingenio* también en la confusión ya que varias figuras humorísticas que se escuchan frecuentemente vienen acompañadas de él, pero son obviamente conceptos distintos.

Existen varias teorías que explican qué es el humor. Se destaca un reporte del estado del arte de Mulder y Nijholt (2002) respecto al humor en sí y respecto al Humor Computacional en el que se engloban algunas de ellas. En las siguientes subsecciones se presentan la Teoría de la superioridad, la Teoría del Alivio, la Teoría de resolución de incongruencias y la Teoría de la violación que intentan explicar el humor desde un punto de vista general. Luego se presentan teorías del humor enfocado desde el punto de vista sociológico, es decir, como afecta el humor en los grupos sociales. Por último se presentan la Teoría de los guiones semánticos del humor y la Teoría general del humor verbal que enfocan el humor desde el punto de vista lingüístico.

2.1.1. Teoría de la superioridad

Gruner (2000) desarrolla una teoría que afirma que el humor está relacionado con la superioridad, que alguien gana sobre otro dentro de un chiste. Al humor se lo ve como al deporte; es más agradable cuando se crea una tensión sobre el resultado, sobre cómo concluirá, y cuando a la vez se “gana” de repente y por sorpresa. Es decir, es más disfrutable ganar un partido parejo que una victoria demasiado aplastante desde el principio. Lo mismo ocurre con el humor; en un chiste uno genera una expectativa sobre cómo concluirá y de forma repentina aparece el “ganador” y la potencial risa.

Cuando se encuentra humor en algo, hay risa sobre las desgracias o defectos de otro (del perdedor), hay un sentimiento de repentina superioridad, ya que momentáneamente no se está pasando por la situación que le está ocurriendo al otro. En este sentido, sentirse superior es “sentirse bien”, es conseguir lo que uno quiere, es ganar. Un ejemplo de chiste visto fácilmente según esta teoría es (2). El mismo tilda de perdedores a los políticos frente a esta situación, en contraste con los que no lo son, que se sienten ganadores.

- (2) — ¿En qué se parece Superman a un político honesto?
— En que ninguno de los dos existe.

Risa es igual a ganar según el autor, pero subraya que no es siempre ganar en el sentido de vencer a alguien, sino más en general de conseguir lo que uno quiere, por ejemplo probar estar en lo cierto, o llegar a una situación que uno quería estar. Bajo este esquema uno podría intentar reconocer humor buscando dentro de un conjunto de oraciones por alguien que esté siendo el objetivo de una crítica, y buscar que sea el “perdedor”, siendo éste un requisito necesario para el humor bajo esta forma de pensar.

2.1.2. Teoría del alivio

En este caso se tiene una naturaleza más psicológica. Está relacionada con la propuesta de Freud y Strachey (1905) que la risa libera la tensión y la “energía síquica”. La misma se acumularía cuando ocurren sentimientos represivos sobre áreas tabú como la muerte o el sexo. Se muestra el chiste (3), que hace referencia a temas que generan tensión. Rutter (1997) afirma que esta es más una teoría sobre la risa que sobre el humor. Establece que nos reímos para liberarnos de temas difíciles para nosotros, pero no dice cómo el humor que trata de esos temas nos hace reír, qué es lo que cumple ese humor.

- (3) Cómo sería de mala aquella suegra que cuando murió, le pusieron este epitafio:
“Aquí descansa doña Juana Baltasar García. En casa descansamos todos”.

Una versión de por qué uno se ríe en base a esta teoría es porque se experimenta una sensación placentera cuando el humor reemplaza sentimientos negativos como dolor o tristeza.

Es interesante destacar que hay ciertos temas que, según esta teoría, son más graciosos ya que generan más tensión sobre las personas. Resulta cautivador fijarse si un texto habla por ejemplo de sexo, muerte o superación personal para encontrar humor.

Existen otras teorías derivadas de la Teoría del alivio, que por más información se puede consultar el Documento del estado del arte de este proyecto.

2.1.3. Teoría de la resolución de incongruencias

La Teoría de la resolución de incongruencias (conocida en inglés como *Incongruity-Resolution Theory*) se basa en que la causa de la risa está en la percepción repentina de que hay incongruencia entre un concepto y los objetos reales que se tenían en mente. Según Rutter (1997), la

incongruencia ocurre cuando dos objetos son presentados bajo un concepto, que parece aplicar a ambos y los hace similares, pero con el progreso del relato humorístico se vuelve aparente que el concepto sólo aplica a uno y la diferencia se hace visible. Puede ser a su vez que no es la incongruencia la que da la gracia, sino que es la resolución coherente de la aparente incoherencia que la da. Es decir, que cuando aparece una incoherencia en un razonamiento, lo gracioso sale de entender que era otro el camino del razonamiento. Se puede ver un ejemplo de aplicación de la teoría en el chiste (4).

- (4) — Mi amor llevamos 30 años juntos, ¿por qué no nos casamos?
— Mejor no, ¿quién se va a querer casar con nosotros?

En este caso, en el comienzo se piensa que “casamos” hace referencia a la pareja entre sí, pero termina haciendo referencia implícita a ellos casándose con otros. Es decir, la segunda oración no es congruente con el comienzo y se tiene que hacer un nuevo razonamiento sobre cómo realmente es.

Al igual que la Teoría de la superioridad, no explica por qué un chiste no es gracioso más de una vez, ni por qué no todas las incongruencias son graciosas. Una posible explicación es que necesitamos distinguir lo serio de los razonamientos equivocados, para saber cómo tratar dicha información, y el humor es una reacción del segundo caso, basándose en Suslov (1992). No se profundizará esta visión ni ninguna otra que fundamente los motivos de la risa ya que no se encuentran dentro del alcance de este proyecto.

La teoría, según Ritchie (1999), no se desarrolla en detalle y deja parcialmente abiertos los conceptos de incongruencia y de resolución. Se puede dividir en otras teorías que por más información se puede consultar el Documento del estado del arte de este proyecto.

2.1.4. Teoría de la violación

Veatch (1998) propone que la condición necesaria y suficiente para la percepción de humor es que se cumplan los siguientes tres puntos:

- **V:** que se viole cierto compromiso de cómo las cosas deben ser según quien percibe.
- **N:** el perceptor tiene un sentimiento dominante de que la situación es normal.
- **Simultaneidad:** V y N están presentes en la mente del perceptor al mismo tiempo.

Es decir la situación parece normal (N) pero al mismo tiempo parece que hay algo mal (V). Por ejemplo, se tiene el chiste (5). La situación parece normal, está mencionando que cuida su celular, que lo tiene en la caja la cual seguramente vino cuando lo compró. Pero luego se viola el pensamiento de que lo compró al decir que está en la tienda, simultáneamente, ya que habla de lo mismo que antes, que era normal hasta el momento.

- (5) Mi iPhone 6 Plus lo cuido tanto que todavía está en su caja... en la tienda... en el shopping...

Si se cumple V y no se cumple N, lo que está pasando no es normal entonces, pero sí hay una violación. El perceptor en este caso está muy comprometido con la situación y lo ve como algo ofensivo. Esto pasa por ejemplo con los chistes racistas, cuando el que lo escucha pertenece a la etnia objetivo del chiste.

2.1.5. Teorías sociológicas

El humor además puede ser visto desde el punto de vista de la Sociología, es decir los comportamientos que aparecen en el humor según las relaciones humanas. Lo que dicen estas teorías no es lo que pasa dentro del ser humano frente al humor, sino entre los seres humanos. El trabajo de Rutter (1997) separa en las siguientes 3 subramas.

Teorías de mantenimiento

Estas teorías afirman que el humor mantiene grupos sociales: se fortalece el vínculo entre quien cuenta el chiste y quienes están en su grupo. Por ejemplo si se relata un chiste, los que se ríen pertenecen al mismo grupo de quien lo relata — porque lo entendieron — mientras que los otros están fuera. O también aquellas personas a las que son alcanzadas por un chiste son separadas en un grupo y al resto en otro, como los chistes racistas por ejemplo. Fortalecen vínculos entre quien cuenta el chiste y quienes están en el mismo grupo que él.

Teorías de negociación

Miran el rol del humor como medios de interacción o pasatiempos. El perceptor define si es gracioso según su contexto social y cultural. Él está de acuerdo con lo que dice el chiste riéndose, o en desacuerdo si no lo hace. Por ejemplo, alguien podría haber estado mirando con otra persona el partido Brasil-Alemania de la edición 2014 de la Copa Mundial de Fútbol, donde el locatario Brasil pierde con Alemania siete a uno en semifinales, dejando en la etapa anterior afuera a Colombia, y decirle el chiste (6).

(6) ¿Para eso eliminaron a Colombia?

La otra persona podría reírse, lo cual significaría que está de acuerdo con él, que está perdiendo Brasil de una mala manera. O podría no reírse por algún motivo. Puede no estar de acuerdo por ejemplo porque fueron partidos jugados de manera totalmente distinta, o se lo podría tomar muy en serio el chiste y enojarse. O tal vez la persona, porque ya estaba enojada desde antes con él, simplemente elige no estar de acuerdo y no reírse, por decir motivos.

Teorías de contextos

Estas teorías afirman que el humorista hace un salto desde el contexto serio al humorístico, puede criticar sin miedo y hasta puede hablar de temas tabú. Los chistes se alejan de lo que es un discurso normal. Al cambiar a un contexto humorístico introduciendo un tema de tratamiento delicado, uno siempre puede decir “era sólo un chiste”.

2.1.6. Teoría de los guiones semánticos del humor

Esta teoría (SSTH — Semantic Script Theory of Humor, en inglés) asume que siempre un chiste está relacionado con dos guiones opuestos entre sí en algún sentido. Dice que hasta antes del remate el chiste no es ambiguo, y que éste es el que dispara la ambigüedad. Postula también que hay tres niveles de abstracción de oposición de guiones:

- **El más alto:** real vs irreal (algo que ocurrió en el chiste con algo imaginario).
- **El del medio:** actual vs no actual, normal vs anormal o posible vs imposible (que explica la diferencia entre lo ocurrido y lo imaginado).

- **El más bajo:** oposiciones temáticas como bueno vs malo, muerte vs vida, sexual vs no sexual, etc (especifica aún más la diferencia).

El remate puede traer ambigüedad. Si trae contradicción semántica (oposición), se trata de *ironía*.

2.1.7. Teoría general del humor verbal

Esta teoría (GTVH — The General Theory of Verbal Humor, en inglés) es formada a partir de la teoría presentada anteriormente, entre otras. Consiste de parámetros llamados Recursos de conocimiento (KR — Knowledge Resources):

- **Oposición de guiones:** definida igual que en SSTH.
- **Mecanismo lógico:** la forma en la cual juntar dos guiones diferentes en un chiste.
- **Situación:** el contexto, ya sea los involucrados, objetos, actividades, etc.
- **Víctima:** la víctima del chiste, si la hay.
- **Estrategia narrativa:** organización narrativa (cómo se cuenta) el chiste.
- **Lenguaje:** conjunto de componentes lingüísticos elegidos para formar el texto del chiste.

El orden en que fueron presentados fue estudiado por Ruch, Attardo y Raskin (1993) y se basa en que hay un incremento lineal de similitud entre chistes según esta jerarquía. Primero, cuanto más parámetros tengan en común dos chistes, más parecidos son. Además, que dos chistes tengan iguales sus parámetros menos Lenguaje, por decir un ejemplo, los hace mucho más parecidos entre sí que dos chistes que tengan iguales sus parámetros menos Oposición de guiones. Se puede entender mejor con los siguientes ejemplos:

- (7) — ¿Cuántos gallegos se precisan para cambiar una lamparita?
— Cinco. Uno para agarrarla y cuatro para girar la mesa.
- (8) — ¿Cuántos gallegos se precisan para lavar un auto?
— Dos. Uno para agarrar la esponja y otro para mover el auto adelante y hacia atrás.

Ambos chistes tienen la misma Oposición de guiones (tonto/inteligente), el mismo Mecanismo lógico (se intercambia qué cosa se mueve), la misma Víctima (los gallegos), la misma Estrategia narrativa (en forma de pregunta), mismo Lenguaje pero difieren en la Situación: en el chiste (7) está la lamparita y en el chiste (8) está el lavado del auto. En conclusión, y como se puede observar, estos chistes son muy similares entre sí. Más aún, reforzando la jerarquía, podemos notar que si comparamos el chiste (7) consigo mismo pero contado con ingenieros en lugar de gallegos, tienen entre sí un nivel de similitud que es más alto que el de los dos chistes anteriores.

Este orden a su vez indica determinación: un recurso de conocimiento (KR) determina a los de abajo. Es decir, un KR limita las opciones de los KR bajo el mismo, según el orden jerárquico presentado. Por ejemplo, Oposición de guiones afecta la elección del Mecanismo lógico.

Se pone en duda si el Mecanismo lógico es un KR o no, ya que si se comparan dos chistes cualesquiera, la similitud de cada KR es creciente siguiendo el orden jerárquico, exceptuando el anterior. Éste a su vez se puede ver como la parte de resolución de la incongruencia según las teorías que afirman su existencia. Asimismo, los chistes “sin sentido” (que fueron mencionados anteriormente) no tienen Mecanismo lógico, lo cual refuerza la opinión de que se trata de un parámetro opcional.

2.2. Detección de humor

Presentadas las teorías que dan un marco global, se pasa a un marco meramente computacional de la detección del humor. En esta sección se analizan trabajos realizados por otros autores en tareas similares a la tarea a resolver. No se han encontrado trabajos que tratan la detección de humor en textos en español. Existen trabajos para el idioma inglés que intentan detectar humor en textos de usualmente un promedio de quince palabras, denominados *one-liners*. En dichos trabajos se intentan utilizar características específicas del humor y en todas se utilizan métodos de clasificación estadísticos.

En general, se analizan dos trabajos donde se estudian características para reconocer el humor en textos en inglés (*one-liners*) y luego se aplican para clasificar frases de humor. El primer trabajo es «Making Computers Laugh: Investigations in Automatic Humor Recognition» de Mihalcea y Strapparava (2005) y el segundo trabajo es «Recognizing Humor Without Recognizing Meaning» de Sjöbergh y Araki (2007), que está fuertemente basado en el anterior. También existen diferentes trabajos donde se estudian ciertas características específicas del humor como la ambigüedad, las palabras fuera del vocabulario, etc., sin el objetivo de construir un clasificador como «An Analysis of the Impact of Ambiguity on Automatic Humour Recognition» Reyes, Buscaldi y Rosso 2009 y «Características y rasgos afectivos del humor: un estudio de reconocimiento automático del humor en textos escolares en catalán» Reyes, Rosso et al. 2009, entre otros. A continuación se describen las características específicas para la detección de humor utilizadas en los trabajos analizados: Aliteración, Ambigüedad, Antonimia, Centrado en personas, Jerga sexual, Negatividad, Palabras clave, y Perplejidad del modelo de lenguaje.

2.2.1. Aliteración

La aliteración es la “repetición notoria del mismo o de los mismos fonemas, sobre todo consonánticos, en una frase”, basándose en la Real Academia Española (2001). Según Mihalcea y Strapparava (2005) las propiedades estructurales y fonéticas de los chistes son tan importantes como el contenido; muchas veces esta característica importa más que el contenido en sí. Los chistes de una línea son graciosos por aliteración, rimas y repetición de palabras aunque no estén pensados para leerlos en voz alta. Un ejemplo de esta característica está dado en el chiste (9). Es importante notar que similarmente la retórica juega un importante rol en los chistes, pero también en otros campos como por ejemplo en los titulares de una noticia.

- (9) Cuando pasé por tu puerta me tiraste una flor, la próxima vez que pase, ¡sin maceta por favor!

Esta característica parece no vincularse con ninguna de las teorías estudiadas. Sin embargo, la aliteración es un condimento que da gusto no sólo al humor, sino a los textos literarios, a las narraciones, canciones, etc.

2.2.2. Ambigüedad

Algunas teorías, como por ejemplo la Desambiguación sorpresiva que deriva de la Teoría de la resolución de incongruencias (ver más en el Documento del estado del arte), afirman enfáticamente que la ambigüedad juega un papel fundamental en el humor. La forma más fácil de calcular la ambigüedad es mirando la cantidad de significados de cada palabra en una oración, siguiendo el trabajo de Sjöbergh y Araki (2007). En el mismo también se presenta otra forma de detección de ambigüedad, que viene dada por la cantidad de posibles árboles de parsing encontrados para una oración, aunque una oración bajo un mismo árbol sintáctico puede ser semánticamente ambigua igualmente, como ocurre en (10).

(10) La perra de mi vecina me ladró.

Existe una medida de la complejidad de una oración, que representa el promedio de algunas de las dependencias sintácticas de la misma, sugerida por Basili y Zanzotto (2002). La fórmula es

$$\frac{N + V}{C}$$

N representa la cantidad de complementos nominales. Ellos son las relaciones sintácticas entre sustantivos y grupos preposicionales (N-GP), como en “té de tilo”. V representa el número de complementos verbales, es decir, las relaciones entre verbos y grupo preposicionales (V-GP), como en “ir al parque”. Por último, C representa el número de cláusulas. Las cláusulas son oraciones atómicas. Por ejemplo, la oración (11) tiene 2 cláusulas.

(11) Voy a lo de Juan y vuelvo más tarde.

Según Reyes, Buscaldi y Rosso (2009) esta medida puede ser utilizada para reconocer la ambigüedad en chistes, dado que, siguiendo su estudio, los chistes tienen una estructura sintáctica más compleja teniendo así un mayor grado de ambigüedad.

Trivialmente estas estructuras suelen ser muy ambiguas. Considerar el ejemplo (12): ¿quién tiene la escopeta? Mirando el contenido, es obvio. Sin embargo, sintácticamente tanto el cazador como la liebre pueden tener la escopeta.

(12) Cacé a la liebre con la escopeta.

2.2.3. Antonimia

La antonimia es la relación de oposición entre los significados de dos palabras (útil — inútil). El humor muchas veces proviene de la incongruencia, oposición u otras formas de contradicción, según Sjöbergh y Araki (2007). Esto es comprobado por la Teoría de los guiones semánticos del humor (Subsección 2.1.6) en su esencia, como así también por la teoría de Minsky (ver Documento del estado del arte) que argumenta que el humor puede ser causado por lógica sin sentido. Ésta se puede dar si se presentan antónimos simultáneamente en una oración. Como ejemplo de chiste que utiliza antonimia está el ejemplo (13). Esta característica se puede medir buscando, para cada palabra en la oración, los antónimos. Si está presente un antónimo en la oración para una palabra, se suma uno obteniendo así el total de antónimos presentes de la palabra. En el trabajo de Mihálcea y Strapparava (2005) se utiliza, para lograr lo anteriormente descrito, la relación antonimia brindada por la herramienta WordNet (que es un corpus de referencia léxica) entre sustantivos, verbos, adjetivos y adverbios. Como características interesantes para decir que hay antonimia en un texto se pueden mirar tanto la cantidad de antónimos de las palabras que componen la oración o también la palabra que tiene más antónimos presentes.

(13) — ¿Qué le dice Tarzán a un ratón?
— ¡Tan pequeño y con bigote!
— ¿Y qué le dice el ratón a Tarzán?.
— ¡Tan grandote y con pañal!

2.2.4. Centrado en personas

Los textos de humor parecen referenciar constantemente a escenarios relacionados con las personas, según Mihalcea y Pulman (2007). Se comprobó en el trabajo de Mihalcea y Strapparava (2005) que referencias humanas como “you”, “I”, “person”, “woman”, “man” y “my” son altamente utilizadas en chistes. También en otro estudio realizado por Mihalcea y Strapparava (2010) se comprueba que la palabra “you” se encuentra en un 25 % de los chistes y la palabra “I” se encuentra un 15 %, en un corpus de chistes de one-liners.

2.2.5. Jerga Sexual

La Teoría del alivio (Subsección 2.1.2) argumenta que el humor libera la tensión y energía síquica que se acumula cuando ocurren sentimientos represivos sobre áreas tabú como el sexo. Es por esto que palabras con un tinte sexual podrían constituir una característica específica del humor. Según Mihalcea y Strapparava (2005) el humor basado en jerga adulta es altamente popular, un ejemplo de esto se da en el chiste (14). Para poder medir esta característica se debe tener un conjunto de palabras de orientación sexual. Para esto los autores formaron un léxico extrayendo, de los dominios de la herramienta WordNet, todas las oraciones etiquetadas con el dominio sexualidad. Luego se procesa la lista quitando todas las palabras que tienen más de cuatro significados distintos (*polisemia* ≥ 4).

- (14) — Che, ¿qué significa “let’s fuck”?
— Hagamos el amor.
— Bueno, haremos el amor. Pero después me dices qué significa.

2.2.6. Negatividad

Según Reyes, Rosso et al. (2009) se destaca que el humor tiene una tendencia hacia las connotaciones negativas. Ellos intentan evaluar si existen verbos o adjetivos que denotan un carácter negativo. Como ejemplo podemos encontrar el chiste (15). Según Mihalcea y Pulman (2007) la característica negatividad también es una característica específica del humor y la subdividen en: palabras de negación como “not” o su abreviación “n’t”, orientación negativa con palabras que tienen polaridad negativa como “bad”, “illegal” y “wrong”, y todos los eventos o entidades que están asociadas con un momento de debilidad humana como por ejemplo los sustantivos “ignorance”, “stupidity”, “trouble”, “beer”, “alcohol” o verbos como “quit”, “steal”, “lie” o “drink”. Esto último se relaciona con la Teoría de la superioridad (Subsección 2.1.1) anteriormente explicada, porque presumen la existencia de un perdedor (como “ignorance” y “stupidity”). También se puede expresar en función de la Teoría del alivio (Subsección 2.1.2), ya que estos temas, como “alcohol”, “illegal” y “lie”, generan tensión.

- (15) — Dr., ¿cómo hago para vivir 100 años?
— Nada de sexo, alcohol ni vicios.
— ¿Y funciona?
— No sé, pero seguro que se le va a hacer largo.

2.2.7. Palabras clave

Según Sjöbergh y Araki (2007) existen palabras que son más referenciadas en textos de humor que en textos de no humor, y palabras que son más aludidas en textos de no humor que en textos de humor. Un ejemplo son los animales, que usualmente son mencionados en chistes, como en

el chiste (16), así como los profesionales como abogados (ver chiste (17)) y programadores, según Mihalcea y Pulman (2007). Este razonamiento se puede, no sólo realizar sobre palabras, sino que también se puede realizar con bigramas y trigramas, y más genéricamente, con modelo de lenguaje. Sjöbergh y Araki (2007) construyen dos listas de palabras, bigramas y trigramas de humor y no humor respectivamente a partir del corpus de entrenamiento, asignando pesos a cada palabra, bigrama o trigrama. Todas las palabras, bigramas o trigramas que no estén en las listas tienen peso nulo; son neutras.

- (16) — ¿Cómo se me mete una jirafa en una heladera?
— Abrís la heladera, metés la jirafa y cerrás la heladera.
- (17) — ¿En qué se diferencia un abogado a un cuervo?
— En que uno es rapaz, ladrón y traicionero, y si puede te saca los ojos, y el otro es un inocente pajarito negro.

2.2.8. Perplejidad del modelo de lenguaje

En el trabajo realizado por Reyes, Rosso et al. (2009) se construye un modelo del lenguaje a partir de narraciones. Luego al graficar la perplejidad de un conjunto de prueba de narraciones y otro de textos de humor se puede ver que la perplejidad del conjunto de textos de humor es mayor a la de las narraciones. La perplejidad es una medida que permite evaluar un modelo de lenguaje: cuanto más bajo sea este factor indica que más se adapta al modelo de lenguaje. Esto indica, entonces, que en los textos de humor se dicen cosas “menos predecibles” comparado con un texto representativo, como las narraciones. Por lo tanto, es menos predecible saber qué palabra sigue a otra cuando aparece en los chistes que cuando aparece en otros textos. Esto también es comprobado por Reyes, Buscaldi y Rosso (2009). La perplejidad puede medir la sorpresa de un chiste, lo que es analizado por la mayoría de las teorías de la sección anterior, como las Teorías de la resolución de incongruencias, SSTH y GTVH, entre otras. También según los trabajos de Reyes, Rosso et al. (2009) y de Reyes, Buscaldi y Rosso (2009) en los textos de humor es más común encontrar palabras que están fuera del vocabulario (OOV — Out Of Vocabulary) que en textos de no humor.

2.3. Evaluación de los trabajos

En esta sección se analizan y evalúan dos trabajos similares a la tarea a realizar con textos en inglés; clasificar textos en humorísticos y no humorísticos. En el primer trabajo, realizado por Mihalcea y Strapparava (2005) «Making Computers Laugh: Investigations in Automatic Humor Recognition», se utilizan como características específicas del humor la Aliteración, Antonimia y Jerga adulta. Se puede observar que la Aliteración es la característica que parece tener más precisión en la detección de humor. En la propuesta de Sjöbergh y Araki (2007) «Recognizing Humor Without Recognizing Meaning» utilizan como características específicas del humor las Palabras clave, Ambigüedad, Aliteración, Antonimia y Palabras centradas en las personas, y se llega a que las Palabras clave es la característica que parece tener más precisión con la detección de humor.

En ambos trabajos mencionados anteriormente resaltan que se obtienen resultados similares con los clasificadores Naïve Bayes y Support Vector Machine (explicados en el Capítulo 4), siendo no preferible uno en vez del otro. También comprueban que una mayor cantidad de datos no mejora la precisión al usar los clasificadores luego que la cantidad de datos supera un umbral.

La gran mayoría de los trabajos encontrados utilizan como datos negativos (de no humor) titulares de noticias, proverbios y el British National Corpus (BNC). Tanto Mihalcea y Strapparava (2005) como Sjöbergh y Araki (2007) tenían la intuición de que las frases graciosas son similares a textos creativos, como por ejemplo proverbios, y por lo tanto más difícil de clasificar contra éstos. Sin embargo, se observa que la tarea más difícil es distinguir entre textos de humor y texto regular (BNC).

Combinando las características específicas del humor y los clasificadores basados en contenido del trabajo Mihalcea y Strapparava (2005) se tiene una precisión de 96,95% con el corpus de los titulares de diario, 79,15% con el corpus BNC y 84,82% con el corpus de proverbios. Mientras que Sjöbergh y Araki (2007) tienen una precisión global del 85,4%. Vale aclarar que entre los dos trabajos anteriormente mencionados no se utilizan los mismos conjuntos de chistes y de oraciones dentro de los corpus de no humor por lo tanto los resultados no son comparables.

Los autores en general son positivos en cuanto a los resultados, teniendo en cuenta la precisión obtenida. Hay que considerar que se tratan de versiones simplificadas y muy controladas de la realidad, como se ha mencionado. Clasificar por características consideradas esenciales y por algunas características accidentales en el humor parece haber sido buena idea. Se pueden mejorar los resultados encontrando características que revelen nuevos puntos de vista, o mejorando la forma de buscar ciertos rasgos difíciles de medir como la ambigüedad.

Capítulo 3

Corpus

Relevado el estado del arte se procede a construir un corpus de textos que son considerados humorísticos y no humorísticos. De esta forma se cuenta con datos de ejemplos que se analizan para poder llegar a caracterizar el humor. Como se menciona en el Estado del arte los estudios realizados hasta la actualidad son respecto a oraciones de pocas palabras. Por este motivo se decide utilizar la red social Twitter para extraer texto, ya que restringe a un máximo de 140 caracteres. Luego de extraídos los textos, de ahora en más denominados tweets, se realiza un proceso de anotación, donde a cada tweet se le agrega una etiqueta simbolizando si es humorístico o no. Terminada esta fase, se obtiene el corpus de textos, en este caso tweets, que sirve como principal insumo para la tarea de clasificación.

3.1. Extracción de tweets

La extracción del corpus se basa en gran parte en los trabajos realizados por Mihalcea y Strapparava (2005) y Sjöbergh y Araki (2007). En estos trabajos se utiliza un corpus con frases de humor denominadas *one-liners* (oraciones con humor de aproximadamente 15 palabras), títulos de noticias, fragmentos del *British National Corpus* (BNC) y proverbios.

Para recabar tweets humorísticos se seleccionan cuentas de humor de Twitter realizando una búsqueda por la palabra clave “chistes”. Del resultado se seleccionan diez cuentas que cumplan que en los primeros veinte tweets el 80% sean efectivamente humor. Se llega a 16.488 tweets, habiendo quitado los repetidos y aquellos tweets que tienen contenido vacío.

Se utiliza, de forma similar a los trabajos mencionados anteriormente, una categorización para obtener tweets no humorísticos: cuentas de noticias, cuentas que tengan reflexiones filosóficas (proverbios) y como no aplica el uso del corpus BNC se opta por la elección de cuentas de curiosidades, reuniendo nueve cuentas, tres por cada categoría. En su totalidad conforman 22.875 tweets.

Para llevar a cabo la extracción se utiliza una rutina que extrae todos los tweets de las cuentas seleccionadas, y los guarda en una base de datos relacional, utilizando la API (*Application Programming Interface*) provista por la red social *Twitter*. En la Tabla 3.1 se presentan las cuentas de humor seleccionadas y en la Tabla 3.2 se presentan las no humorísticas.

Cuenta	Nombre	Seguidores	Descripción
@EresChiste	Puros Chistes!	1.901.999	¿Aburrido? Has venido al mejor lugar, sigueme y diviértete con el mejor Humor y los mejores Chistes! Contacto: publicidad@publicidad140.com
@LosChistes	LosChistes.com	654.005	LosChistes.com , tu fuente de risa instantánea y buen humor. Síguenos en Twitter con los mejores chistes cortos y novedades de nuestro sitio
@Riete	RÍETE	935.062	Nuestra misión es hacerte reír... Nuestra Visión es un Mundo lleno de alegría. (Chistes, Frases, Humor, Piropos...) @CarlosLesma PUBLICIDAD: riete@live.com
@ChisteTipico	Chistes♣	2.512.589	Reír es la cura de todo síguenos y pasa un buen rato. (♡) Animate a ser leído CONTACTO - PUBLICIDAD - ChisteTipico@gmail.com
@A_reirse	A reirse	2.061.151	Mi misión es hacerte reír :) Cuenta de humor y entretenimiento en la que podrás divertirti leyendo mis tweets (chistes, frases y mucho más)
@ChistesPicantes	ChistesPicantes	98.347	Sólo los mejores Chistes Picantes (rojos, verdes, fuertes, etc) Si prefieres chistes más variados, te recomendamos seguir a @loschistes
@ChisteVeloz	CHISTES!	339.244	¿Aburrido? Has venido al mejor lugar, sigueme y diviértete con el mejor Humor y los mejores Chistes! Contacto: publicidad@publicidad140.com
@ChistesAlDia	Chistes Al Dia	146.182	Los mejores chistes de twitter. Ríete un rato y comparte tus risas con los mejores chistes.
@ChisteUniversal	CHISTES!	1.259.844	El que se enoja siempre pierde, aquí es un lugar universal para MORIR de RISA. Contacto: publicidad@publicidad140.com
@TWICHISTE	Humor y Chistes	2.368.491	Twichiste. Entretenimiento, Humor y los mejores Chistes! SÍGUEME envía tu Chiste para Retweet Contacto: http://www.twichiste.com/p/contact.html o en twichiste-publi@gmail.com

Tabla 3.1: Cuentas de humor seleccionadas. Último acceso a la información: 22/08/2014

Cuenta	Nombre	Seguidores	Descripción	Categoría
@iDescubrelo	CURIOSIDADES	507.878	Aquí descubrirás #Curiosidades #Frases #Noticias /De Todo Un Poco/ #PUBLICIDAD Contacto: publicidad@publicidad140.com	Curiosidades
@sabiastuque_	Sabías? Curiosidades	1.671.318	¿Sabías Que? Curiosidades. Cuenta sobre todo tipo de imagenes curiosas en el mundo actual. ¿Eres curioso? Entonces esta es tu cuenta #Sabias-que #curiosidades	Curiosidades
@sabiasundato_o	SABÍAS UN DATO™	1.634.470	Eres curioso? aquí podrás descubrir las cosas más curiosas de la internet ... PUBLICIDAD: info@sabiasundato.com	Curiosidades
@GranReflexion	Reflexiones	3.664.332	Son Reflexiones de la Vida: Frases y Pensamientos para reflexionar e inspirarte, Consejos para pensar y alcanzar la felicidad.	Proverbios
@ifilosofia	Miguel	2.216.617	Matemático. Si te gustan los pensamientos para reflexionar, las noticias de ciencia y los acertijos de ingenio, quédate.	Proverbios
@TwitsReflexion	Relfexiones☞	132.925	Cuenta Oficial de Reflexiones.	Proverbios
@CNNEE	CNN en Español	8.625.822	CNN en Español es tu principal fuente de noticias. Cubrimos lo que pasa en América Latina y el resto del mundo. Vive la noticia	Noticias
@Subrayado	Subrayado	142.172	Cuenta oficial del informativo Subrayado, de Canal 10. Uruguay.	Noticias
@telenocheonline	TelenocheOnline	30.575		Noticias

Tabla 3.2: Cuentas no humorísticas elegidas. Último acceso a la información: 22/08/2014

3.2. Anotación del corpus

Luego de extraídos los tweets se debe distinguir si son humorísticos o no. Una posible forma es denotar a todos los tweets extraídos de cuentas de humor como humorísticos y los extraídos de las restantes cuentas como no humorísticos. Sin embargo, no todos los tweets de cada cuenta se caracterizan por ser de la misma categoría (humor — no humor), por lo que se decide que deben ser anotados manualmente. En los tweets extraídos de cuentas no humorísticas no se han encontrado ejemplos que sean incorrectos, es decir, que sí sean de humor. Sin embargo, en los tweets extraídos de cuentas humorísticas sí. Dado que la cantidad de tweets extraídos es excesiva para ser procesados manualmente por una reducida cantidad de personas, se procede a crear una aplicación tanto web como de celulares donde cualquier persona a través de Internet puede clasificar un tweet.

3.2.1. Proceso de Anotación

En consecuencia a la decisión tomada anteriormente se crea una página web¹ y una aplicación para el sistema operativo Android² con el objetivo de que cualquier usuario pueda clasificar los tweets como humorísticos, como no humorísticos o saltárselo en caso de duda.

Los tweets que se presentan al anotador son al azar, pretendiendo que no se repita un tweet a la persona para evitar que lo clasifique varias veces. En la Figura 3.1 se ve una captura de pantalla de la página web, en donde se muestra uno de los tweets extraídos como humorísticos, y en la Figura 3.2 se puede ver una captura de la aplicación Android.



Figura 3.1: Página que usan los usuarios para clasificar los tweets

¹<http://www.clasificahumor.com>

²<https://play.google.com/store/apps/details?id=com.clasificahumor.android>

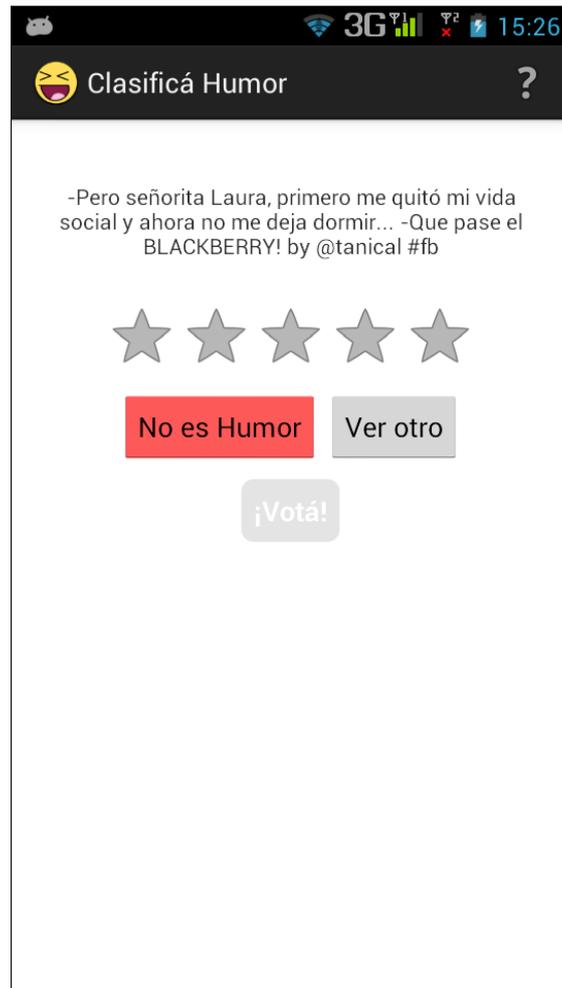


Figura 3.2: Aplicación Android que usan los usuarios para clasificar los tweets

Se busca que la aplicación utilizada sea funcional, simple y eficiente, trasladando todo el procesamiento posible al lado del cliente y el mínimo al lado del servidor. De esta forma la aplicación puede escalar en cantidad de clientes, permitiendo adicionalmente que una vez anotado un tweet pasar al siguiente sea instantáneo.

Asimismo, se intenta que la página y la aplicación no ofrezcan contenido explícito y puedan ser usadas por personas de cualquier edad, por lo que se remueven aquellos tweets que contengan palabras sexuales explícitas o insultos que se encuentren dentro de una lista de 32 palabras conformada para este fin. Se eliminan 303 tweets, quedando 16.185 tweets de humor en total.

3.2.2. Criterios de anotación

Se debe analizar si el anotador debe indicar si es humor o no, o si debe asignarle un puntaje en el caso que sea de humor el tweet. Si el anotador siguiera la primera opción, se le quitaría dificultad a la hora de anotar, pudiendo eventualmente llegar a clasificar más cantidad de tweets debido a la simpleza de elegir solamente entre dos opciones. Sin embargo, se tendría menos información sobre la certeza de que el tweet sea humor, lo cual es considerado necesario ya que existen ciertos tweets que en general se duda si son humor o no, como por ejemplo el tweet (18). Tampoco es adecuado pedir mucha información al anotador ya que haría que rápidamente se agote por tener que pensar constantemente qué calificación poner, y al ser voluntario esto generaría que deje de clasificar. Finalmente se decide llegar a una solución de compromiso que consiste en una escala de humor de uno a cinco, en caso de que un tweet sea considerado humorístico, y una opción para indicar no humor. De esta forma, debido a que la definición de humor es subjetiva, denotaremos como humor a un tweet en función de lo que los anotadores dictaminen.

- (18) En un examen. Ver “Nombre”. Escribir tu nombre y apellido. Ver en la siguiente línea: “Apellido”. Sentirte un tremendo idiota.

Es importante notar que dadas las etiquetas elegidas, se podrán derivar otras. Por ejemplo, la etiqueta *humor*, que se corresponde a la suma de la cantidad de anotaciones de 1, 2, 3, 4 o 5 estrellas, y la anotación *promedio de votación*, que corresponde al promedio de las estrellas por tweet.

3.2.3. Resultado de la votación

La página se difunde a través de redes sociales, como *Facebook* y *Twitter*, de manera de tener más cantidad de votaciones. Se reciben más de 23.000 votaciones durante los primeros tres días de existencia (se libera el 1º de setiembre de 2014), y para el fin del mes de setiembre de 2014 se reciben más de 46.000.

Se debe tener en cuenta también remover votaciones de anotadores vandálicos: deben quitarse aproximadamente 20.000 votaciones correspondientes a dos anotadores que han votado en todas las ocasiones como no humor, cuando entre ellas existen tweets que claramente lo son. Además las votaciones se dan en un intervalo de tiempo reducido lo que permite suponer que se trata de una máquina y no de un ser humano. Al culminar el mes de octubre se llega a un total de 33.531 votos para los 16.185 tweets (sin contar las veces que se presionó el botón “Ver otro”, y sin tener en cuenta las votaciones removidas).

En la Figura 3.3 y en la Tabla 3.3 se muestra la cantidad de votos por cada calificación. Se puede observar que la mitad de votos son de no humor y los restantes se distribuyen entre las categorías de uno a cinco estrellas. En la Figura 3.4 se muestra el histograma de cantidad de votos por tweet. De aquí se desprende que aproximadamente la mitad de los tweets tienen a lo sumo un voto.

Votos por calificación

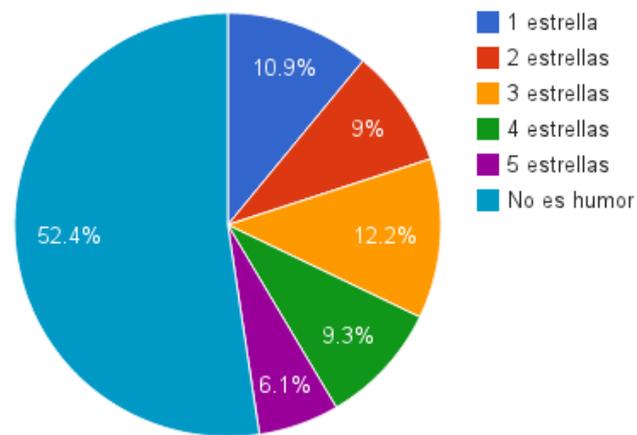


Figura 3.3: Cantidad de votos por calificación

Calificación	#votos
1 estrella	3.665
2 estrellas	3.031
3 estrellas	4.087
4 estrellas	3.132
5 estrellas	2.054
No es humor	17.562

Tabla 3.3: Cantidad de votos por calificación

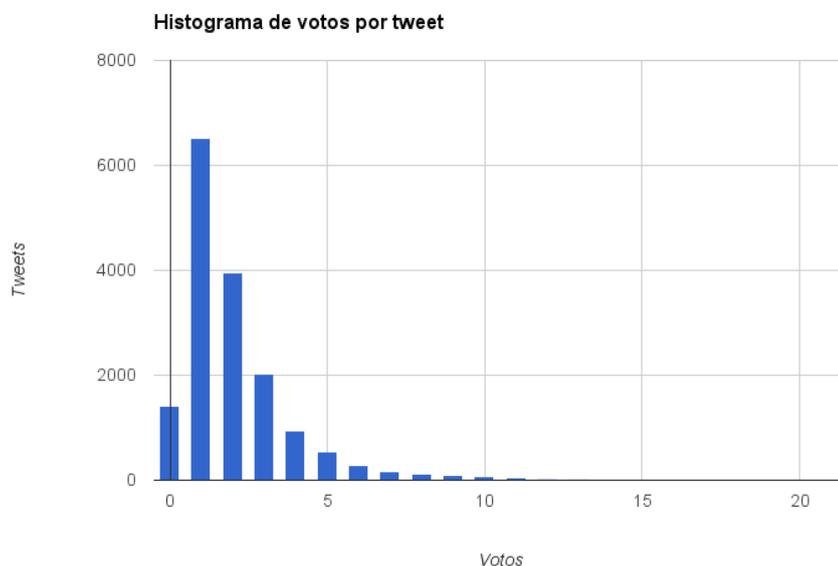


Figura 3.4: Histograma de la cantidad de votos por tweet

Sea la *fracción de votos de humor* la cantidad de votos de humor (de 1, 2, 3, 4 o 5 estrellas) dividido entre el total de votos por tweet. La Figura 3.5 representa un histograma de la fracción de votos de humor (en porcentaje). Se puede ver claramente que la gráfica se encuentra discretizada. Esto es debido a que cada tweet tiene baja cantidad de votos, hace que las posibilidades de combinación de votos no sea muy alta. Por ejemplo, con 3 votos se pueden tener 3 votos de humor y ninguno de no humor, 1 voto de humor y 2 de no humor, 2 votos de humor y uno de no humor o 0 votos de humor y 3 votos de no humor.

3.2.4. Humor según la votación

Tomando en cuenta el resultado de la votación, se decide establecer un criterio para definir qué es humor en base a las votaciones:

Un tweet debe tener al menos un 60% de sus votos como Humor para ser considerado humorístico (los votos con 1, 2, 3, 4 o 5 estrellas), menor o igual a 30% para ser considerado de No humor, y de otra manera es considerado como Dudoso (inclusive cuando no tiene votos).

Bajo este criterio se tienen las proporciones mostradas en la Tabla 3.4 e ilustradas en la Figura 3.6. Asimismo, con este criterio se puede decidir la categoría entre Humor y No humor con pocos votos: los tweets con tres votos se pueden decidir con dos de ellos iguales, los que tienen cuatro con tres y los que tienen cinco con tres también. Adicionalmente se está dejando menos margen a la categoría No humor ya que se supone que los tweets son a priori de humor al venir de cuentas de este tipo. Se debe tener en cuenta que utilizando otros márgenes se obtienen resultados similares, en donde siempre existen algunos tweets que son considerados como dudosos

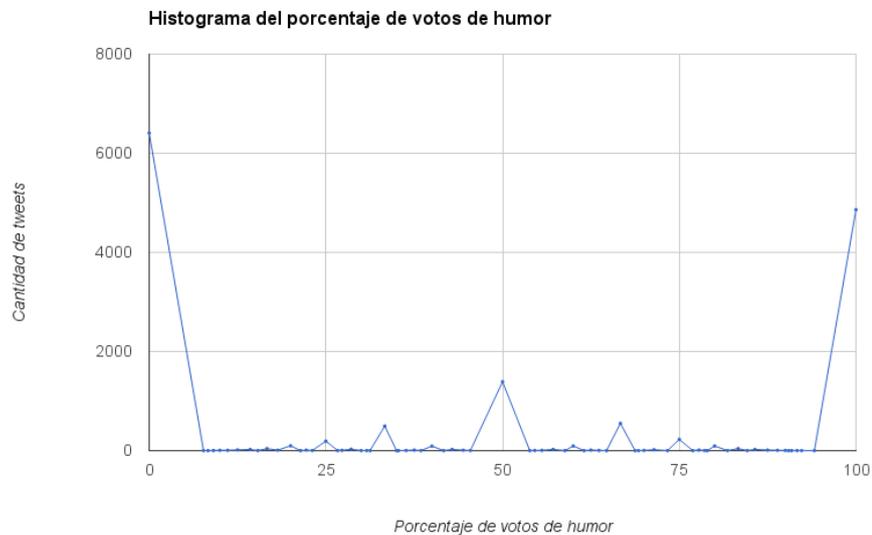


Figura 3.5: Histograma del porcentaje de votos de humor de los tweets

para la gente y es muy difícil considerarlos como humor o como no humor sin dejar mucho margen de error.

Se pueden observar ciertos grupos que se reconocen como los picos presentes en la gráfica del histograma de la cantidad de votos de la Figura 3.5. A continuación se describen los grupos encontrados, considerando solamente aquellos que agrupen más de 45 tweets:

- 0 % de votos de humor, cantidad 6.351 tweets. Son calificados como no humorísticos.
- 20 % de votos de humor, cantidad 98 tweets. Son calificados como no humorísticos.
- 25 % de votos de humor, cantidad 196 tweets. Son calificados como no humorísticos.
- 33,33 % de votos de humor, cantidad 491 tweets. Son calificados como dudosos.
- 40 % de votos de humor, cantidad 93 tweets. Son calificados como dudosos.
- 50 % de votos de humor, cantidad 1.388 tweets. Son calificados como dudosos.
- 60 % de votos de humor, cantidad 96 tweets. Son calificados como humorísticos.
- 66,67 % de votos de humor, cantidad 554 tweets. Son calificados como humorísticos.
- 75 % de votos de humor, cantidad 233 tweets. Son calificados como humorísticos.
- 80 % de votos de humor, cantidad 92 tweets. Son calificados como humorísticos.
- 100 % de votos de humor, cantidad 4.862 tweets. Son calificados como humorísticos.

Conjunto	#tweets
humor	5.952
no humor	6.755
dudoso con votos	2.066
dudoso sin votos	1.412

Tabla 3.4: Cantidad de tweets por conjunto de humor/no humor/dudoso



Figura 3.6: Repartición de tweets en grupos de humor/no humor/dudoso

En este análisis de los grupos se encuentra el 93,16% de los tweets. También vale notar que existen sólo 69 valores distintos de porcentaje de humor.

Se observa que según la votación existe una gran cantidad de tweets de no humor, aunque era esperable que los tweets extraídos de las cuentas de humor tuvieran mucho menos ruido, considerando que se eligen de manera tal de evitarlo. A modo ilustrativo se presentan algunos de los tweets que pertenecen al conjunto de los dudosos en los ejemplos (19), (20) y (21).

(19) Típico: esa alegría de saber que a tu amigo también le fue mal en la prueba.

(20) Dejar para mañana lo que ayer dejé para hoy.

(21) 2 palabras, 9 letras, 1 mentira universal: estoy bien.

3.2.5. Promedio de estrellas

Adicionalmente es atractivo reportar el promedio de estrellas obtenidas en el corpus. Se define el promedio de votación de un tweet t como:

$$PV_t = \frac{\sum_{i=0}^5 i \times v_{ti}}{v_t}$$

A su vez se define el promedio de votación positiva de un tweet t como:

$$PVP_t = \frac{\sum_{i=0}^5 i \times v_{ti}}{h_t}$$

El promedio de la medida PVP para todos los tweets con más de un voto es 2,0344, mientras que el de la medida PV es 1,4468. Si se utilizan las mismas medidas sobre el conjunto de tweets seleccionado como humor se obtiene que el promedio de la medida PVP es 2,8409 mientras que el promedio de la medida PV es 2,5810, indicando que el conjunto de tweets seleccionados como humor tiene mayor promedio de humor que el conjunto de todos los tweets, como es de esperarse.

3.2.6. Mejores chistes

Por último se pueden ver los tweets mejores votados mirando el promedio de votación. Sin embargo esto puede estar entorpecido por tweets con pocos votos; hay que tener en cuenta también cuántas votaciones tiene un tweet para hacer una lista de los mejores chistes. Se propone primero ordenar por el promedio de votación y luego según la cantidad de votos. Bajo este criterio, los mejores son los tweets (22) y (23) que tienen promedio cinco y cantidad de votos tres.

(22) — ¿Fumaste marihuana?
 — ¡No, papá! Te juro que...
 — Idiota, ¡soy tu perro!
 — ¡Jajajá Firulay me asustaste! :(

(23) — Ayer, al salir del trabajo atropellé a un unicornio.
 — No jodas, ¿tenés trabajo?

3.2.7. Concordancia entre los anotadores

Es de interés saber qué concordancia hay entre los anotadores, es decir, qué tan de acuerdo están las personas a la hora de anotar. Se propone utilizar la medida *kappa* de Fleiss, que mide la concordancia en el corpus.

La medida kappa de Fleiss (1971) sirve para evaluar la concordancia entre los anotadores de un corpus. Calcula el grado de acuerdo en la clasificación comparándose con el azar, siendo 1 el máximo valor posible y pudiendo dar menor que 0, significando ser peor que el azar. En este trabajo sirve para medir la concordancia entre los distintos anotadores sobre el corpus construido.

Tiene como ventaja sobre la medida kappa de Cohen (1960), que origina a la medida kappa de Fleiss, que acepta más de dos anotadores: supone que hay una cantidad fija de anotaciones por instancia, pudiendo eventualmente provenir de distintos anotadores.

Se define de la siguiente manera:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Donde se define \bar{P} como un promedio:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i$$

En donde N es la cantidad de ejemplos clasificados y P_i es la fracción de pares de anotadores que coinciden en una categoría en el ejemplo i -ésimo:

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1)$$

Siendo n la cantidad de anotaciones que tuvo cada instancia, k la cantidad de categorías y n_{ij} la cantidad de anotadores que asignaron el ejemplo i a la categoría j (observar que $\sum_{j=1}^k n_{ij} = n, \forall i \in \{1, \dots, N\}, j \in \{1, \dots, k\}$).

\bar{P}_e es como \bar{P} pero aplicado a obtener pares de coincidencias en categorías al azar:

$$\bar{P}_e = \sum_{j=1}^k p_j^2$$

En donde p_j es la proporción de las anotaciones que fueron a la categoría j :

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}$$

Se evalúa la medida kappa de Fleiss para las votaciones mirando si han sido consideradas de humor o de no humor, independientemente de la calificación. Para esto hay que enfrentarse a un problema: existe la suposición que tiene que haber una cantidad fija de votaciones por tweet. Entonces se procede a hacer una normalización, llevando las votaciones de cada tweet al mínimo común múltiplo de las distintas cantidades de votaciones que hay, teniendo en cuenta cada cantidad de votaciones en las categorías humor y no humor. Por ejemplo, si en un tweet se tienen 3 votos de humor y 2 de no humor y se decide llevar la cantidad de votaciones a 10, quedará como 6 votos de humor y 4 de no humor. En la Tabla 3.5 y en la Figura 3.7 se muestra la medida kappa de Fleiss para distintos conjuntos de tweets. El acuerdo para todos los tweets posibles según kappa es 0,6122, pero cuando tweets con más votos son considerados, peor es.

tweets considerados	#tweets	κ
≥ 2 votos	8.320	0,6122
≥ 3 votos	4.309	0,5226
≥ 4 votos	2.273	0,4691
≥ 5 votos	1.331	0,4341
≥ 6 votos	805	0,4056
≥ 7 votos	527	0,3883
≥ 8 votos	354	0,3810
≥ 9 votos	244	0,3586
≥ 10 votos	164	0,3225
≥ 11 votos	105	0,3086
≥ 12 votos	64	0,2931

Tabla 3.5: Medida kappa de Fleiss para los distintos conjuntos de tweets

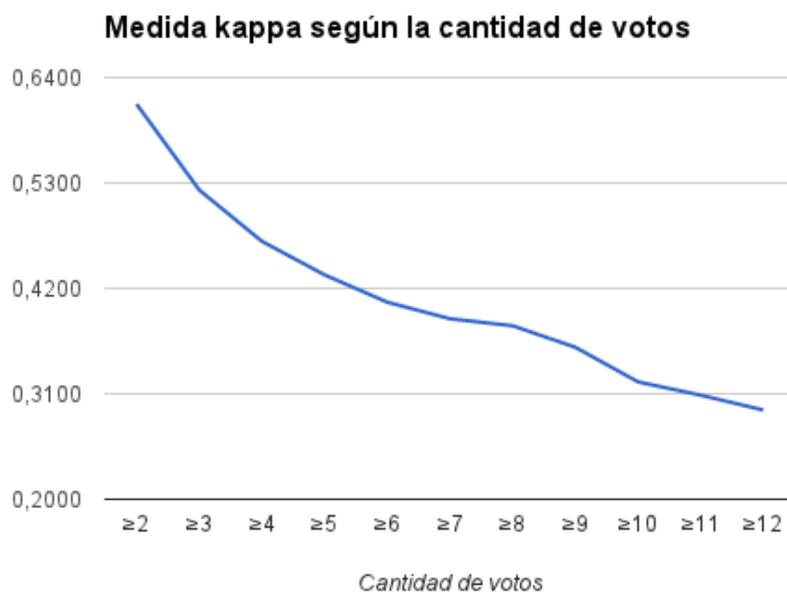


Figura 3.7: Medida kappa de Fleiss según la cantidad de votos

Existe poca determinación entre autores sobre qué valores de *kappa* son considerados buen acuerdo o no (Gwet 2014, cap. 6). Por la definición de la medida se puede asegurar que valores muy cercanos a cero o menores a él implican mal acuerdo porque son iguales o peores que el azar. Razonando de manera análoga valores altos y cercanos a 1 son considerados buenos. Se considera que los valores obtenidos en este caso indican una concordancia de nivel medio-alto, ya que se consigue ser aproximadamente un 60 % de lo mejor que se puede llegar, respecto a una anotación aleatoria.

Capítulo 4

Aprendizaje Automático

Es necesario estar un poco en contexto sobre los métodos de Aprendizaje Automático que son utilizados para construir un clasificador de detección de humor. Éstos hacen distintas suposiciones, que tomándolas como verdaderas permiten generalizar conceptos y predecir futuras instancias nunca vistas. También es necesario conocer las métricas disponibles. En el presente capítulo se abordan superficialmente ambos aspectos: si el lector considera que tiene una base de conocimientos en el área, puede saltarse este capítulo.

4.1. Métodos de Aprendizaje Automático

La idea principal del Aprendizaje Automático es hacer aprender a una computadora. En este trabajo se quiere aprender una función, que es aquella que dice si un tweet es Humor o No humor. Al tratarse de un codominio de categorías, la tarea que se quiere desempeñar se llama *clasificación*. Si por ejemplo el codominio fueran los números reales, sería *regresión*. A su vez, este trabajo se basa en clasificación estadística ya que aprende de los ejemplos (se entrena con ellos) para luego poder predecir. Al aprender de ejemplos marcados como Humor o No humor, se dice que es *aprendizaje supervisado*. Si tuviera que aprender de ejemplos no etiquetados sería *aprendizaje no supervisado*, y tendría que de alguna forma deducir las reglas que reinan al modelo.

A continuación se presentan los distintos tipos de clasificadores utilizados en este trabajo.

4.1.1. Support Vector Machine

Support Vector Machine (SVM) (scikit-learn 2015g) es un método que busca la mejor curva que separa los ejemplos en un espacio n -dimensional, siendo n la cantidad de características. Los ejemplos son representados por vectores, en donde cada componente corresponde con el valor de una de las características.

La Figura 4.1 ilustra cómo funciona SVM. La curva utilizada en este caso es un hiperplano (una recta, al ser dos dimensiones), aunque el tipo de curva es un parámetro; puede elegirse. Los ejemplos en este caso quedaron linealmente separables, y es en lo que se basa SVM: supone que las instancias son separables por una curva. Sin embargo no siempre ocurre esto: pueden quedar ejemplos mal clasificados.

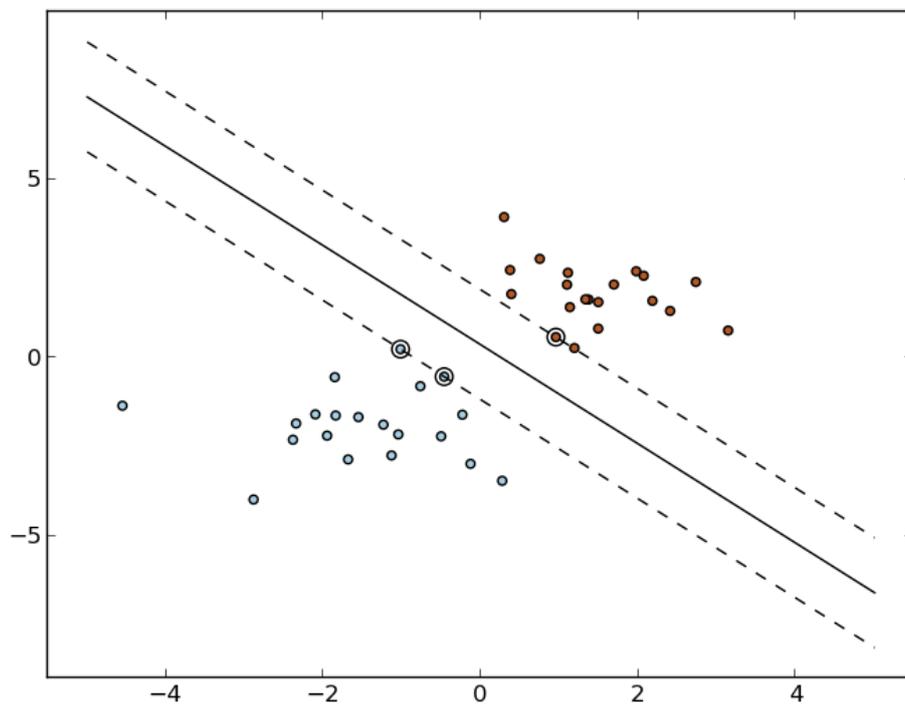


Figura 4.1: Ejemplo de curva que separa los ejemplos. En este caso quedan todos separables, pero eventualmente puede no suceder esto. Los puntos distinguidos son los “vectores de soporte” que le dan nombre a esta técnica. Ellos cumplen ser los puntos más cercanos a la curva, y que por tanto la definen.

4.1.2. Árbol de decisión

Un Árbol de decisión o en inglés Decision Tree (DT) (Mitchell 1997, cap. 3; scikit-learn 2015a) es una estructura que basa su clasificación en forma de preguntas sobre el valor de los atributos. En sus nodos internos tiene atributos, en donde cada rama adyacente corresponde con un valor posible del atributo, y en las hojas tiene un valor de clasificación elegido. En la Figura 4.2 se muestra un ejemplo.

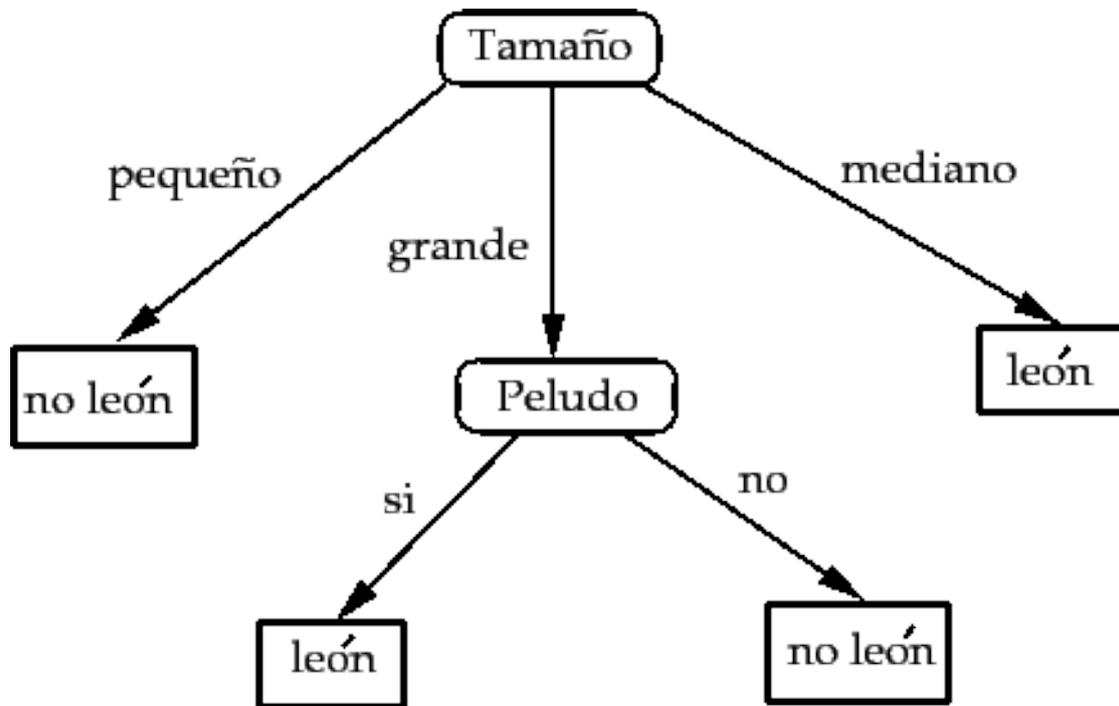


Figura 4.2: Ejemplo de árbol de decisión para discriminar animales en si son leones o no.

La forma de generar un árbol de decisión, en la biblioteca scikit-learn (2014), se basa en el algoritmo ID3 (*Iterative Dichotomiser 3*). El mismo comienza desde la raíz, elige el atributo más discriminante (según algún criterio, como *Ganancia de información* o *Gini*) para el nodo actual, arma tantas ramas como valores distintos acepte el atributo (o hay que separar en intervalos si el atributo es continuo) y continúa recursivamente con aquellos ejemplos que tengan dicho valor del atributo, hasta que llega a un nodo con todos los ejemplos pertenecientes a la misma clase, o hasta que no queden atributos disponibles.

Este algoritmo, al tomar los atributos más discriminantes al comienzo, genera árboles poco profundos pero anchos. Esto se traduce en que esta forma de aprendizaje supone que el concepto que se quiere encontrar es de alguna forma simple, ya que los árboles de decisión se corresponden con expresiones lógicas¹ que resultan ser más cortas y simples.

Esta forma de armar los árboles de decisión puede llevar a *sobreajuste* (en inglés *overfitting*). Esto significa que se elige una *hipótesis* que se comporta muy bien en el conjunto de entrenamiento

¹Notar que un árbol de decisión se puede expresar en forma lógica, dado un valor de la clase objetivo, como la disyunción de las conjunciones de los valores de los atributos de todos los caminos posibles desde la raíz hasta las hojas que contienen dicho valor de la clase objetivo (Forma normal disyuntiva). En el ejemplo de la Figura 4.2, la clase león se puede expresar como: (tamaño = grande Y peludo = si) O (tamaño = mediano)

y tiene baja predicción en el de evaluación, pero existen otras hipótesis que tienen ligeramente peor predicción en el entrenamiento pero mejor rendimiento en la evaluación. En la Figura 4.3 se muestra un ejemplo de sobreajuste. Algunas condiciones de parada pueden ser impuestas para evitar esto, como establecer el número mínimo de ejemplos en una hoja, o la máxima profundidad del árbol. Existe también la idea de *sobregeneralización* (en inglés *underfitting*) que ocurre cuando se generaliza demasiado un concepto que se busca definir.

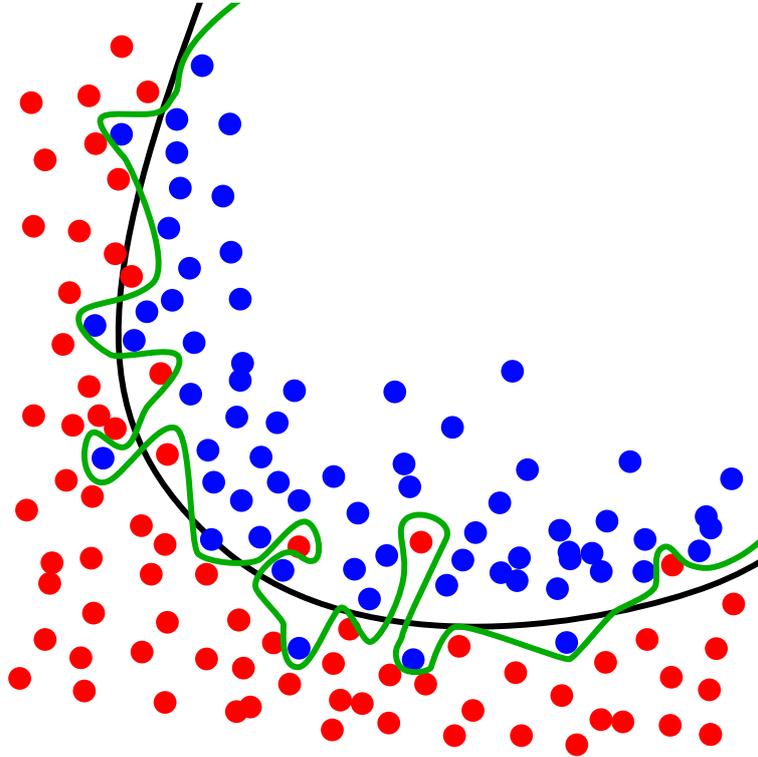


Figura 4.3: Ejemplo de sobreajuste. Se quiere buscar la curva que separa los puntos, pero se ajusta demasiado al conjunto de entrenamiento y se pierde poder de generalización.

4.1.3. Naïve Bayes

Un clasificador bayesiano (Mitchell 1997, cap. 6; scikit-learn 2015b) se basa en Teorema de Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Donde A y B son dos sucesos. $P(A)$ es conocida como la probabilidad de A *a priori* y $P(A|B)$ es la probabilidad *a posteriori*.

Siendo C el atributo objetivo con sus valores posibles y siendo a_i el valor de cada atributo A_i , se puede obtener la clase más probable de un ejemplo:

$$\begin{aligned}
c &= \arg \max_{c \in C} P(c|a_1, \dots, a_n) \\
&= \arg \max_{c \in C} \frac{P(a_1, \dots, a_n|c)P(c)}{P(a_1, \dots, a_n)} \\
&= \arg \max_{c \in C} P(a_1, \dots, a_n|c)P(c)
\end{aligned}$$

Es decir, el valor de clase más probable es aquel que maximiza la probabilidad. Notar que calcular la probabilidad conjunta de los valores de los atributos requiere de muchos ejemplos para cada combinación posible de valores, lo que no siempre ocurre. Por lo tanto, el método de Naïve Bayes supone interdependencia entre atributos, y postula que:

$$c = \arg \max_{c \in C} P(c) \prod_{i=1}^n P(a_i|c)$$

Estas probabilidades se pueden aproximar en base a las frecuencias de los valores en el conjunto de entrenamiento. Eventualmente se puede suponer que los atributos siguen una distribución, más aún si sus conjuntos son no acotados. A la versión de este clasificador que supone que los atributos son discretos y siguen una distribución multinomial se le llama Multinomial Naïve Bayes (MNB). A la que supone que los atributos son continuos y siguen una distribución normal, se le llama Gaussian Naïve Bayes (GNB).

4.1.4. k Nearest Neighbors

Esta técnica de clasificación (Mitchell 1997, cap. 13; scikit-learn 2015c) se basa en instancias conocidas. Es decir, clasifica directamente según el valor de ejemplos vistos considerados parecidos. En este caso se ubican a las instancias en un espacio vectorial, de forma análoga a SVM, y se clasifica en base a los vecinos más cercanos. Se puede ponderar la clasificación también en base a la distancia de los mismos. En la Figura 4.4 se ilustra la clasificación que realiza k Nearest Neighbors (kNN).

4.2. Medidas de evaluación

Es necesario definir medidas para poder evaluar a un clasificador, y también así poder compararlo. En la clasificación binaria se distingue a los valores de la clase objetivo como positivo o negativo. En este documento se trata a la clase Humor como positivo y al No humor como negativo. A partir de esto pueden definirse cuatro categorías para los ejemplos clasificados:

- **Verdadero positivo:** ejemplo correctamente clasificado como positivo.
- **Falso positivo:** ejemplo incorrectamente clasificado como positivo.
- **Verdadero negativo:** ejemplo correctamente etiquetado como negativo.
- **Falso negativo:** ejemplo incorrectamente clasificado como negativo.

Se define la medida *precisión* de la siguiente manera:

$$Precision = \frac{vp}{vp + fp}$$

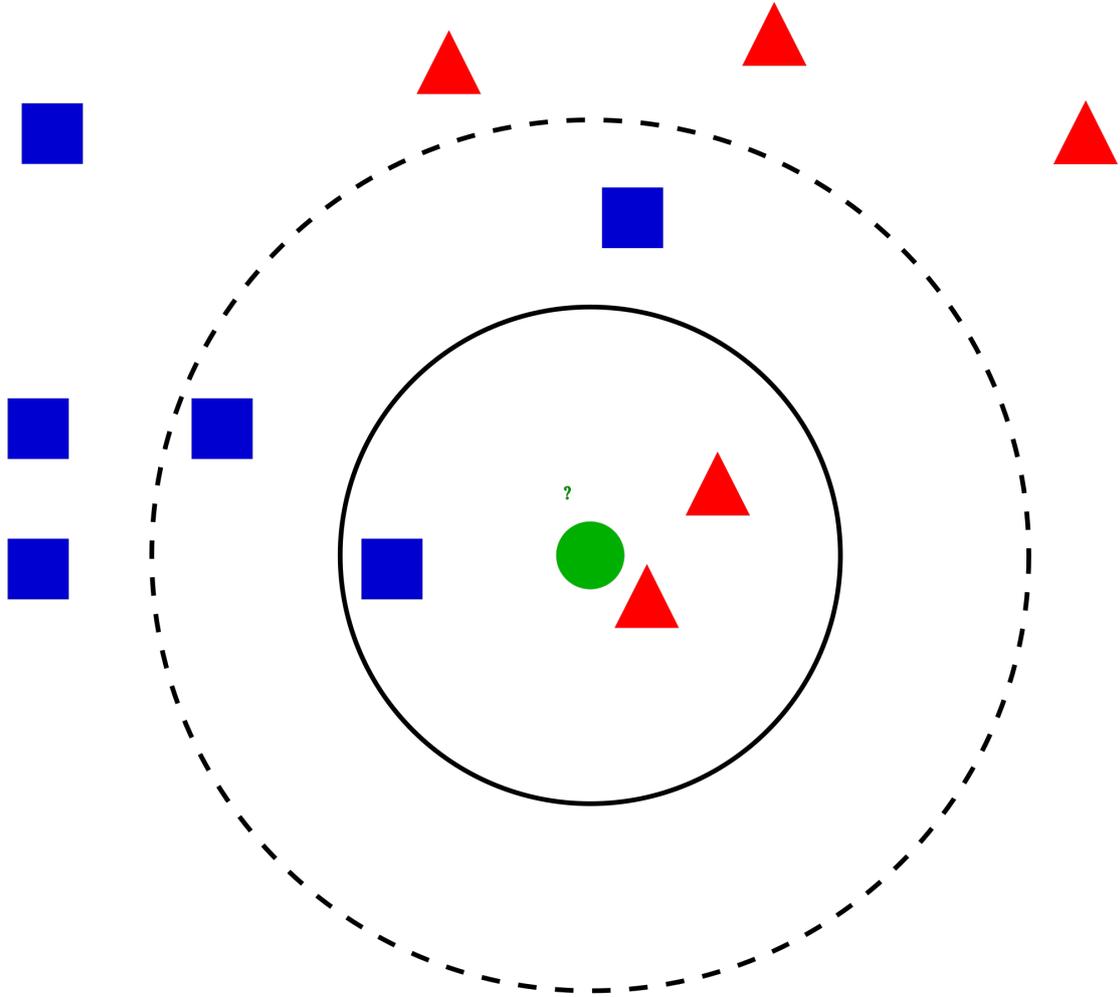


Figura 4.4: Ejemplo de clasificación en base a los vecinos más cercanos. Si consideramos los tres vecinos más cercanos, se clasificaría como rojo, pero si consideramos cinco sería azul.

Siendo vp la cantidad de verdaderos positivos y fp la de los falsos positivos. La precisión indica cuán certero es el clasificador con aquellas instancias que considera positivas. Dicho de otra manera, le da peso a los falsos positivos.

La medida *recall* (o *exhaustividad*, aunque se prefiere el término en inglés) también es definida:

$$Recall = \frac{vp}{vp + fn}$$

Donde fn es la cantidad de falsos negativos. Esta medida por su parte le da peso a los falsos negativos. Indica cuánto el clasificador llega a encontrar los ejemplos positivos.

También se define la medida F_1 :

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{vp}{vp + \frac{fp+fn}{2}}$$

La misma puede ser vista como un tipo de promedio entre las medidas anteriores, aunque es mejor vista como una medida que le da igual peso a los falsos positivos y a los falsos negativos.

Observar que estas medidas toman en cuenta todo lo que ocurre alrededor de los positivos. Se define la medida *acierto*, que se enfoca también en los ejemplos negativos:

$$Acierto = \frac{vp + vn}{vp + vn + fp + fn}$$

El acierto mira la fracción de ejemplos clasificados correctamente. La Figura 4.5 ilustra las medidas mencionadas.

Cuál medida es referente depende de la tarea que se quiera desempeñar. Por ejemplo, si se tiene un clasificador que es utilizado para mostrar aquellos ejemplos positivos, se preferirá una buena precisión ya que al tener una baja tasa de falsos positivos no se mostrarían instancias incorrectamente clasificadas. Si el recall es bajo se pierden instancias positivas, que eventualmente podrían ser de interés para la tarea en cuestión.

Por otro lado, si se está buscando un documento con un buscador, tal vez valga la pena arriesgar un poco de precisión a cambio de un buen recall. Es más importante poder encontrar un documento buscado que la molestia de perder tiempo mirando un documento que no es relevante, al menos mientras no sean demasiados.

Cuando lo que importa son tanto los falsos negativos como positivos, se puede mirar el acierto así como la medida F_1 .

Hay que notar también cómo los ejemplos disponibles influyen sobre las medidas. Si el clasificador no se desempeña de igual manera con los ejemplos positivos que con los negativos y además las proporciones de ambos no son similares, se ven afectadas el acierto, la precisión y por tanto también la medida F_1 . El acierto se ve afectado ya que mira la clasificación de ambas clases, estando más afectada por la que tenga más cantidad de ejemplos de las dos. La precisión se ve afectada por los negativos ya que, cuanto más haya, más posibilidades hay de encontrar falsos positivos. A su vez influye qué ejemplos son considerados. Si los ejemplos tienden a sesgarse sobre cierta categoría, los resultados también. Hay que tener esto en cuenta según la tarea que se quiera realizar. Por ejemplo, si se quiere ver cómo se comporta el clasificador en la realidad, sería bueno poder contar con una muestra aleatoria de ejemplos. Si se desea estudiar cómo se comporta el clasificador según distintos tipos de categorías o tipos, se pueden conseguir varios ejemplos de cada uno y tomar eso como base.

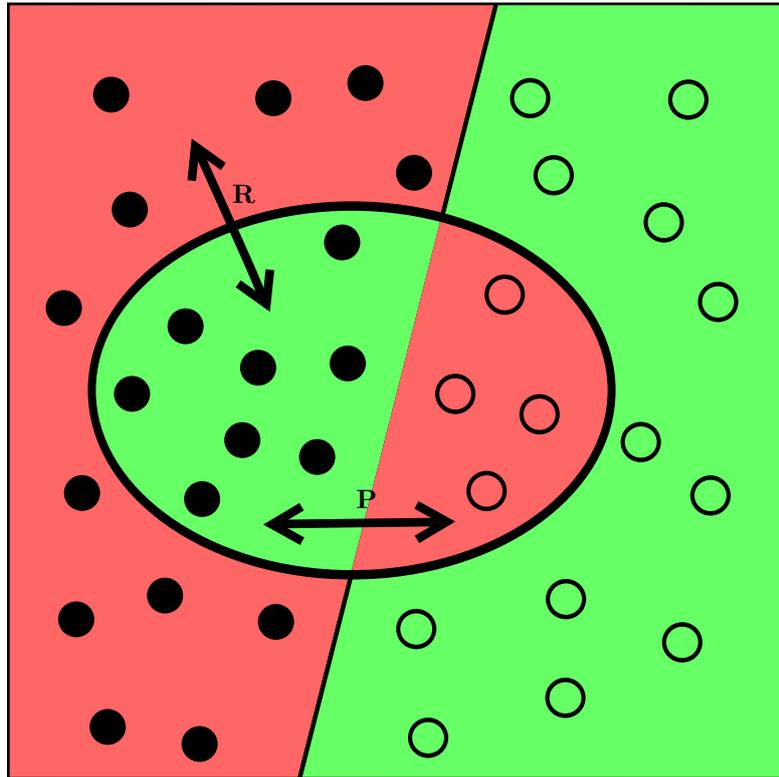


Figura 4.5: En este ejemplo se quieren clasificar los puntos rellenos. Las instancias consideradas como positivas por el clasificador están encerradas en la elipse. La línea separa a las instancias según su clase. Las que tienen fondo verde son aquellas correctamente clasificadas. La *precisión* es el cociente entre la parte verde dentro de la elipse y la cantidad de instancias en toda la elipse. El *recall* es la cantidad de puntos rellenos en la elipse dividido la cantidad de puntos rellenos. El *acierto* es la cantidad de puntos con fondo verde sobre el total de puntos.

4.3. Escalado de características

En algunos tipos de clasificadores, como SVM y kNN, hay una función de distancia implícitamente definida, en la cual los distintos valores de cada característica imponen un peso en la misma. Por ejemplo, si se tiene una característica binaria que indica con 0 o 1 la presencia de cierta palabra en un tweet y otra característica que cuenta la cantidad de letras, la última toma valores muchos más grandes y una diferencia de dos letras en un tweet representarían una distancia mayor que dos tweets con distintos valores en la primera característica, aunque no necesariamente es cierto que una distancia sea mayor que la otra. A métodos que estandaricen los datos de esta manera se dice que hacen *Escalado de características* (Shalev-Shwartz y Ben-David 2014, sección 25.2.1).

Se puede resolver este problema realizando un *Reescalado*, dejando los valores de todas las características en el mismo rango, como $[0, 1]$, por ejemplo. Pero adicionalmente, cuando se entrena con SVM con ciertos tipos de curvas, se asumen que los datos tienen la misma varianza y están centrados en cero. Entonces se preprocesan los valores de los atributos para que queden con promedio cero y desviación estándar uno. A esto se le conoce como *Estandarización*.

Para el caso puntual de MNB no tiene sentido hacer escalado ya que supone que las características tienen distribución multinomial, es decir de un dominio discreto. A su vez el clasificador no permite valores negativos de características, por lo tanto un centrado tampoco aplica.

Capítulo 5

Clasificador

Una vez obtenido el estado del arte, el corpus de tweets y un conocimiento básico de los métodos de aprendizaje automático se procede a implementar un clasificador que permita detectar humor en tweets en español. Para esto, a partir del corpus se definen el conjunto de entrenamiento y evaluación y se definen líneas base de los clasificadores construidos. Luego se realizan características que intentan definir el humor y se muestran los resultados de los primeros clasificadores construidos. Por último, se realiza selección de características para obtener las más relevantes y se realizan ciertos análisis de interés. Se utilizan los clasificadores SVM, DT, kNN, MNB y GNB.

5.1. Conjunto de entrenamiento/evaluación

Del corpus construido en la etapa se toma aleatoriamente un 20 % como conjunto de evaluación y el resto se utiliza para entrenar.

Al evaluar el clasificador buscando resultados finales se utiliza el conjunto de evaluación, mientras que al computar resultados intermedios se utiliza la técnica de validación cruzada con cinco particiones sobre el conjunto de entrenamiento, buscando no sesgarse al conjunto de evaluación elegido inicialmente. Adicionalmente se puede utilizar un 20 % aleatorio del conjunto de entrenamiento para que oficie de conjunto de evaluación si se quiere estudiar ejemplos de tweets mal clasificados.

La validación cruzada se realiza partiendo el conjunto aleatoriamente una cantidad de veces determinada k , en donde se hacen k iteraciones entrenando con $k - 1$ particiones y evaluando con la restante. Tiene la ventaja que no se sesga fácilmente a haber elegido determinada división en entrenamiento y evaluación ya que esto se hace más de una vez, evitando sobreajustarse. Es indicado también cuando se dispone de pocos datos (y así poder aprovecharlos para entrenar tanto como para evaluar).

5.1.1. Subcorpus

Es importante tener en cuenta que a partir de la votación se obtiene que hay tweets que resultan ser de no humor. Por ese motivo se hace la siguiente distinción: a la parte del corpus que proviene de cuentas de humor se le denomina *Subcorpus de cuentas de humor*, y dentro de ella se le denomina *Subcorpus de humor* a la porción que se toma como de humor según la votación. También se define la parte del corpus que proviene de cuentas no humorísticas como *Subcorpus de cuentas no humorísticas*.

5.2. Línea base

Para poder comparar este trabajo y determinar una calidad mínima, se definen dos líneas base. La primera utiliza MNB combinado con la técnica Bag of Words (BoW). Para tokenizar el texto se busca por palabras de al menos 3 caracteres de largo que tengan alguna letra y sean alfanuméricas o con guiones o puntos. Se descartan algunas palabras comunes del idioma español de una lista de 275 palabras, como “de”, “el” y “la”. El resultado de la clasificación se reporta en la Tabla 5.1, en donde el acierto es de 83,51 %. Se busca mirar principalmente la medida F_1 de Humor en este caso para ponderar de igual manera los falsos positivos y negativos de la clase Humor.

BOW	precisión	recall	F_1
No humor	97,72	86,06	88,80
Humor	63,12	75,42	68,73

Tabla 5.1: Métricas para el clasificador BOW sobre el corpus de evaluación

La segunda es un clasificador que predice cada ejemplo como proveniente de la clase más probable a priori, no humor, cuya frecuencia es aproximadamente 83 % (tanto en el corpus de entrenamiento como en el de evaluación). Interesa sobretodo evaluar las medidas que aplican sobre el corpus de humor, pero se puede utilizar esta línea base para comparar el acierto.

5.3. Características

Se implementan características en base a lo estudiado en el Estado del arte y analizando patrones en los distintos tweets del conjunto de entrenamiento. A continuación se hace una categorización de las características implementadas, en donde luego se detallan en la presente subsección:

- **Temas típicos de chistes:** teniendo en cuenta que hay ciertos temas que son frecuentes en el humor, se puede buscar por ellos para intentar reconocerlo. Las características que se implementan bajo esta categoría son: Presencia de animales, Jerga sexual, Distancia temática y Palabras frecuentes.
 - **Temas que generan tensión:** siguiendo la Teoría del alivio (Subsección 2.1.2) hay ciertos temas que generan tensión y el humor los libera. Se puede intentar buscarlos para identificar al humor. Se implementa la característica Jerga sexual.
- **Orientado a personas:** los chistes encontrados en general cumplen que son conversaciones o hablan de personas, así como también lo sugiere el Estado del arte. Se implementan las características: Primera persona, Segunda persona y Diálogo.
- **Formato:** se encuentra que muchos tweets humorísticos cumplen con ciertos formatos particulares. Teniendo en cuenta esto, se elaboran las características: Diálogo, Preguntas-Respuestas y Links.
- **Contradicción — Negatividad:** siguiendo la idea de algunos autores que en el humor hay contradicciones y se encuentran muchas negaciones, y que además existen oposiciones en el humor (según SSTH, Subsección 2.1.6), se implementan las características Antónimos y Negación.

- **Informalidad:** en textos humorísticos suelen haber contenido menos formales que en textos no humorísticos. Se implementan entonces las características: Exclamación, Palabras mayúsculas, Palabras fuera del vocabulario, Palabras no españolas, Hashtags.

5.3.1. Presencia de animales

La característica *Presencia de animales* intenta captar la mención de animales en los tweets. Para lograr el cometido se conforma un diccionario de animales extraído de la fuente Chistes.com (2015), se toman los chistes anotados con la categoría *Animales* y se identifican manualmente las palabras que referencian a animales. Se conforma un diccionario de 103 nombres de animales, incluyendo errores típicos de ortografía como es la omisión de tildes. Luego, para cada tweet, se toma la medida *PresenciaAnimales*, donde DIC_A es el anteriormente mencionado y *tweet* es representado como un conjunto de palabras:

$$PresenciaAnimales(tweet) = \frac{|tweet \cap DIC_A|}{\sqrt{|tweet|}}$$

La intersección denotada en la fórmula es una intersección de multi-conjuntos. Por ejemplo, si un nombre de animal se encuentra más de una vez en un tweet, al realizar la intersección se cuenta tantas veces como aparezca. Se normaliza con la raíz cuadrada del largo del tweet dado que cuanto más palabras tenga el tweet más nombres de animales podrá tener.

5.3.2. Jerga sexual

La característica *Jerga sexual* intenta captar referencias a temas sexuales que sean tabú en la sociedad. Para lograr el cometido se conforma un diccionario de palabras consideradas como jerga sexual. El diccionario construido en una primera instancia se basa en 21 palabras intuitivas. Luego, se realiza una extracción de tweets buscando por las palabras de a pares y se cuenta la cantidad de veces que ocurre una palabra cualquiera en cada tweet, quitando palabras frecuentes del idioma español. Por último se seleccionan manualmente de la lista resultante aquellas que hacen referencia a jerga sexual, logrando así obtener una total de 132 palabras. La técnica de obtener más cantidad de ejemplos a partir de un conjunto reducido se conoce como *Bootstrapping*.

Para cada tweet se toma la medida *JergaSexual*, donde DIC_{JS} es el diccionario de palabras referentes a jerga sexual y *tweet* es representado como un conjunto de palabras:

$$JergaSexual(tweet) = \frac{|tweet \cap DIC_{JS}|}{\sqrt{|tweet|}}$$

La intersección denotada en la fórmula es una intersección de multi-conjuntos.

5.3.3. Primera persona

La característica *Primera persona* intenta captar verbos, determinantes y pronombres que hacen referencia a texto en primera persona. Para esto se utiliza la herramienta Freeling (Padró y Stanilovsky 2012), que por cada token indica la categoría a la que pertenece. Luego para cada tweet se toma la medida *PrimeraPersona* como la cantidad de tokens en primera persona y se divide entre la raíz cuadrada de la cantidad de tokens del tweet.

5.3.4. Segunda persona

La característica *Segunda persona* intenta captar verbos, determinantes y pronombres que hacen referencia a texto en segunda persona. Se utiliza la herramienta Freeling, que por cada token indica la categoría a la que pertenece. Para cada tweet se toma la medida *SegundaPersona* como la cantidad de tokens en segunda persona y se divide entre la raíz cuadrada de la cantidad de tokens del tweet.

5.3.5. Distancia temática

La característica *Distancia temática* intenta captar la cercanía a los temas potenciales de chistes. Para esto se utiliza la fuente *chistes.com* que contiene chistes clasificados según una categoría: Machistas, Borrachos, Matrimonios, etc. Para cada categoría se contruye un clasificador que discrimina entre la categoría en cuestión y texto seleccionado del corpus *Wikicorpus* al azar. Para los clasificadores se utiliza el MNB con la técnica BoW, de igual manera que en la primera Línea base. El clasificador seleccionado además de retornar la clase a la cual pertenece un ejemplo retorna la probabilidad que el texto que se clasifica pertenezca a la categoría en cuestión.

Para cada categoría se tiene una característica igual a la probabilidad de la clase que predice el clasificador. Sólo se consideran aquellas categorías que tengan más de 300 ejemplos, dado que con menos cantidad se considera que una categoría no está bien representada. Por lo tanto se implementan las siguientes cinco características: *Chistes cortos*, *Adivinanzas*, *Animales*, *Atlantes* y *Otros....*

5.3.6. Diálogo

Esta característica intenta indicar la presencia de diálogo en el tweet, buscando si comienza con un guion de diálogo, en base a varios símbolos utilizados comúnmente para indicar su presencia, como el símbolo de resta, viñetas, etc.

5.3.7. Preguntas-respuestas

Intenta contar la cantidad de preguntas seguidas de respuestas que hay en el tweet. Lo hace buscando la presencia de partes del texto encerradas entre signos de preguntas, seguidas por texto en forma no interrogativa. Se basa en que existen muchos chistes que tienen preguntas y respuestas.

5.3.8. Palabras frecuentes

La característica *Palabras frecuentes* intenta captar la presencia de palabras frecuentes en tweets. Para esto se conforma, de manera intuitiva, una lista de 43 palabras frecuentes en chistes.

Luego para cada tweet se toma la medida *PalabrasFrecuentes*, donde DIC_{PF} es el diccionario de palabras frecuentes y *tweet* es representado como un conjunto de palabras:

$$PalabrasFrecuentes(tweet) = \frac{|tweet \cap DIC_{PF}|}{\sqrt{|tweet|}}$$

La intersección denotada en la fórmula es una intersección de multi-conjuntos.

5.3.9. Links

La característica *Links* cuenta la cantidad de enlaces a sitios web en un tweet. Generalmente el humor no necesita de un hipervínculo, por lo tanto la presencia de enlaces puede ser un buen indicador de que el tweet no es de humor. Inclusive el enlace podría tratarse de una imagen, pero al ser este trabajo sobre humor verbal no deberían necesitarse.

5.3.10. Antónimos

Cuenta la cantidad de pares de antónimos que hay en el tweet, dividido entre la raíz cuadrada de la cantidad de tokens que resultan de tokenizarlo usando Freeling:

$$Antonimos(tweet) = \frac{|\{pares\ de\ antonimos\}|}{\sqrt{|tweet|}}$$

5.3.11. Hashtags

La característica *Hashtags* cuenta la cantidad de *hashtags* de un tweet.

5.3.12. Exclamación

La característica *Exclamación* cuenta la cantidad de signos de exclamación de apertura o de clausura de un tweet, dividiendo entre la raíz de la cantidad de tokens:

$$Exclamacion(tweet) = \frac{|\{signos\ de\ exclamación\}|}{\sqrt{|tweet|}}$$

5.3.13. Palabras mayúsculas

La característica *Palabras mayúsculas* cuenta la cantidad de palabras totalmente en mayúsculas, dividido entre la raíz de la cantidad de tokens:

$$PalabrasMayusculas(tweet) = \frac{|\{palabras\ mayúsculas\}|}{\sqrt{|tweet|}}$$

Las palabras totalmente en mayúscula pueden denotar informalidad.

5.3.14. Negación

Cuenta la cantidad de veces que aparece la palabra “no” y se divide entre la raíz cuadrada de la cantidad de tokens del tweet.

5.3.15. Palabras fuera del vocabulario

Las características *Palabras fuera del vocabulario*, u *OOV*, intenta captar la presencia de palabras fuera del vocabulario que generalmente se encuentran en textos de humor debido a su originalidad. Existen cuatro diferentes características para captar las palabras fuera de vocabulario:

- **OOV Freeling-Google:** por cada token de un tweet se consulta si se encuentra en el diccionario brindado por la herramienta Freeling, y en caso de no estar se consulta al buscador Google (2015). Si no se encuentran resultados o el buscador realiza una sugerencia de corrección ortográfica del token buscado, se incrementa un contador de tokens fuera del vocabulario. Una vez procesados todos los tokens se normaliza el contador dividiendo entre la raíz cuadrada de la cantidad de tokens en el tweet.
- **OOV Freeling:** por cada token de un tweet se consulta si se encuentra en el diccionario brindado por la herramienta Freeling y en caso de no estar se incrementa un contador de tokens fuera del vocabulario. Una vez procesado todos los tokens se normaliza el contador dividiendo entre la raíz cuadrada de la cantidad de tokens en el tweet.
- **OOV Freeling-Wiktionary:** por cada token de un tweet se consulta si se encuentra en el diccionario brindado por la herramienta Freeling, en caso de no estar se consulta al diccionario Wiktionary (2015). Si no se encuentra en el diccionario se incrementa un contador de tokens fuera del vocabulario. Una vez procesados todos los tokens se normaliza el contador dividiendo entre la raíz cuadrada de la cantidad de tokens en el tweet.
- **Wiktionary:** por cada token de un tweet se consulta si se encuentra en el diccionario Wiktionary y en caso de no estar se incrementa un contador de tokens fuera del vocabulario. Una vez procesado todos los tokens se normaliza el contador dividiendo entre la raíz cuadrada de la cantidad de tokens en el tweet.

5.3.16. Palabras no españolas

Esta característica cuenta la cantidad de palabras que tienen caracteres fuera del alfabeto español y luego se normaliza dividiendo por la raíz cuadrada de la cantidad de tokens en el tweet.

5.4. Resultados iniciales

Luego de implementadas las características se procede a ejecutar cada clasificador. En la Tabla 5.2 se muestran los resultados obtenidos por los diferentes clasificadores realizando escalado de características si corresponde y ajuste de parámetros. En el Anexo A se encuentra la arquitectura del clasificador y en el Anexo B se encuentra el ajuste de parámetros. El Anexo C detalla las opciones del clasificador y muestra un ejemplo de ejecución.

Clasificador/Medida	precisión positiva	recall positivo	F_1 positivo	precisión negativa	recall negativo	F_1 negativo	acierto
SVM	84,60	66,23	74,29	93,53	97,59	95,52	92,37
DT	65,03	65,19	65,10	93,04	92,98	93,01	88,53
GNB	59,68	76,78	66,49	95,06	89,17	92,01	87,04
MNB	84,57	58,83	69,37	92,24	97,85	94,97	91,35
kNN	82,19	64,26	71,12	93,16	97,21	95,14	91,76

Tabla 5.2: Métricas de los clasificadores inicialmente

Se puede ver que los clasificadores SVN, MNB y kNN superan la primera línea base ya que F_1 positivo es mayor a 68,73%. Además todos los clasificadores superan la segunda línea base dado que tienen un acierto superior a 83%.

5.5. Selección de características

Interesa saber si existen características irrelevantes, y en caso afirmativo eliminarlas. La presencia de características irrelevantes puede causar sobreajuste, lo que implica una pérdida en la capacidad de generalizar el concepto subyacente que se quiere encontrar y describir, que en este caso es el humor. Como consecuencia directa se tiene un rendimiento predictivo bajo a la hora de clasificar nuevas instancias. De la misma manera es deseado remover características que sean redundantes, ya que esto permite acelerar el proceso de entrenamiento y de predicción. Por consiguiente, se quiere valorar cada característica.

De forma ilustrativa para las pruebas, se agrega una característica de nombre *ALEATORIA*, cuyo valor asignado dado un tweet es un número aleatorio entre 0 y 1, y también se introduce una que devuelve exactamente el mismo valor que la clase objetivo, de nombre *CLASE*.

5.5.1. Extremely Randomized Trees

Extremely Randomized Trees (Extra Trees) (scikit-learn 2014) es una técnica que entrena una cantidad fija de árboles de decisión aleatorios, en este caso 1000, con varios subconjuntos de los datos, promediando la capacidad de cada uno de predecir. En base a cómo quedan formados los árboles y qué capacidad de predicción tienen, se decide qué características son mejores discriminantes y les asigna un porcentaje de importancia. En la Tabla 5.3 se muestran los resultados obtenidos.

5.5.2. Eliminación recursiva de atributos

La Eliminación recursiva de atributos (Guyon et al. 2002) (en inglés *Recursive Feature Elimination* o RFE) consiste en asignar pesos a las características con el clasificador para luego quitar recursivamente una a una volviendo a entrenar y evaluar para poder encontrar el número óptimo de características. Se procede junto con validación cruzada, de manera de tener más certeza sobre los resultados. Esta estrategia tiene la ventaja de validar empíricamente los atributos que deja de lado.

Al realizar esta técnica se encuentra que el óptimo se da al quitar las características Negación, Palabras no españolas y Antónimos. Se decide entonces eliminar estas características, quedando veinte. El puntaje obtenido, según el número de características se puede ver en la Figura 5.1. La diferencia entre todas las características y eliminar estas tres es muy sutil.

5.6. Resultados finales

En esta sección se detallan los resultados de los diferentes tipos de clasificadores, sin considerar las características descartadas en la etapa selección de características (Sección 5.5) sobre el conjunto de evaluación. En la Tabla 5.4 se presentan los resultados obtenidos.

Los mejores resultados obtenidos son utilizando SVM, mientras que kNN da resultados ligeramente inferiores. GNB y DT tienen muy baja precisión, aunque GNB tiene mucho más alto recall que SVM. Sin embargo esto sirve de poco para la tarea de saber si un tweet es humor, ya que predice más tweets como humorísticos haciendo subir el recall y bajando la precisión. Se puede ver que los clasificadores SVN, MNB y kNN superan la primera línea base ya que tienen una medida F_1 de humor mayor a 68,73%. Además todos los clasificadores superan la segunda línea base dado que tienen un acierto superior a 83%. En la Tabla 5.5 se muestra la matriz de confusión para el clasificador SVM y el corpus de evaluación.

Característica	Valor
CLASE	74,05
Diálogo	10,86
Distancia temática: Otros...	03,07
Distancia temática: Atlantes	02,65
Distancia temática: Chistes cortos	02,53
Preguntas-respuestas	01,98
Distancia temática: Adivinanzas	0,95
Distancia temática: Animales	0,87
Palabras clave	0,57
Exclamación	0,57
Hashtags	0,55
Links	0,54
Primera persona	0,21
Segunda persona	0,15
OOV Freeling	0,08
Palabras mayúsculas	0,06
ALEATORIA	0,06
Jerga sexual	0,06
Negación	0,05
OOV Wiktionary	0,04
OOV Freeling Wiktionary	0,03
Presencia de animales	0,03
OOV Freeling Google	0,03
Antónimos	0,02
Palabras no españolas	0,00

Tabla 5.3: Resultados obtenidos para Extra Trees

Clasificador/Medida	precisión positiva	recall positivo	F_1 positivo	precisión negativa	recall negativo	F_1 negativo	acierto
SVM	83,61	68,85	75,52	93,84	97,24	95,51	92,45
DT	66,51	67,54	67,02	93,33	93,03	93,18	88,85
GNB	57,49	78,17	66,25	95,17	88,16	91,53	86,46
MNB	84,76	60,02	70,27	92,27	97,79	94,95	91,37
kNN	81,26	66,31	73,03	93,35	96,87	95,08	91,67

Tabla 5.4: Métricas finales de los clasificadores

son/clasificados	Humor	No humor
Humor	842	381
No humor	165	5805

Tabla 5.5: Matriz de confusión del clasificador SVM y el corpus de evaluación.

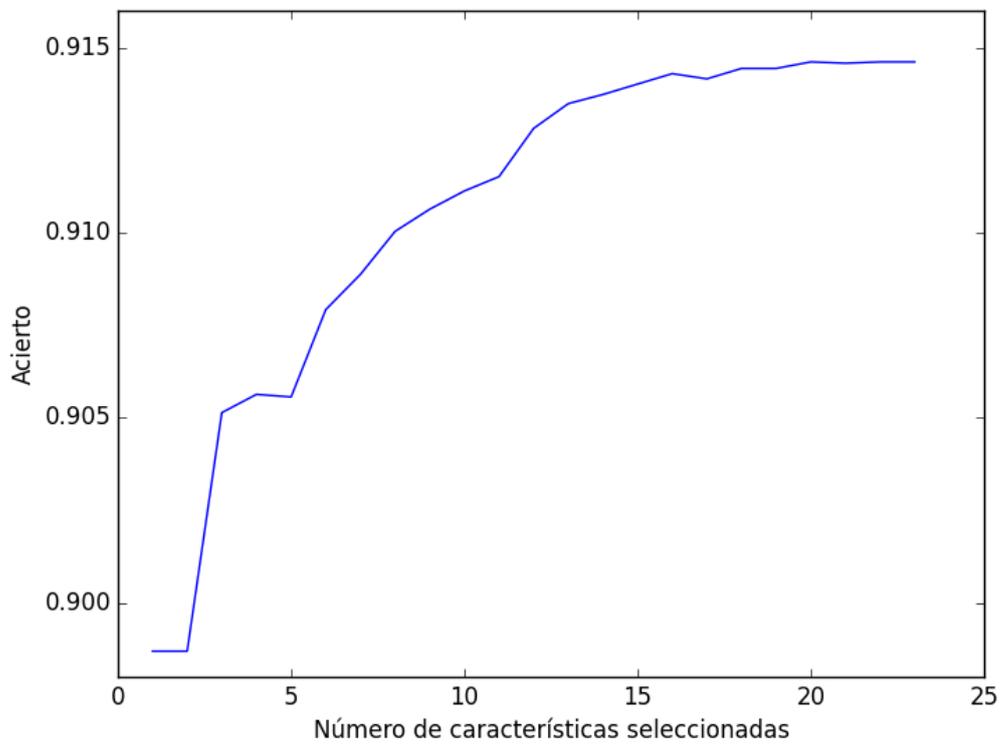


Figura 5.1: Acierto según la cantidad de características consideradas en RFE

5.7. Evaluación en el conjunto de entrenamiento

Es de interés saber cómo se comporta el clasificador evaluando con el mismo conjunto con el que se entrena, ya que no obtener muy buenos resultados de esta manera es un indicador de que no hay que esperar buenos resultados del clasificador. En la Tabla 5.6 se muestran los resultados de evaluar sobre el conjunto de entrenamiento para los distintos tipos de clasificadores.

Clasificador/Medida	precisión positiva	recall positivo	F_1 positivo	precisión negativa	recall negativo	F_1 negativo	acierto
SVM	87,46	69,61	77,52	94,16	98,01	96,05	93,28
DT	99,96	98,82	99,38	99,76	99,99	99,88	99,80
GNB	58,06	77,65	66,44	95,21	88,79	91,89	86,94
MNB	84,56	58,93	69,46	92,26	97,85	94,97	91,67
kNN	86,98	71,47	78,47	94,49	97,86	96,15	93,47

Tabla 5.6: Métricas de los clasificadores sobre el conjunto de entrenamiento

5.7.1. Inconsistencias en el corpus

A diferencia de otros tipos de clasificadores, DT no tiene *sesgo restrictivo* (Mitchell 1997, sec. 3.6.1). Esto quiere decir que para construir un clasificador DT parte del espacio de hipótesis completo y se ajusta al conjunto de entrenamiento. Por los parámetros utilizados, no hay profundidad máxima en el árbol y el mínimo de ejemplos por hoja es uno, por tanto la única condición de parada de la construcción de este tipo de clasificador se da cuando ya no hay más atributos por discriminar. Por lo tanto, si falla la clasificación en DT sobre el conjunto con el que se entrena, es debido a que existen instancias con los mismos valores en las características, pero pertenecen a distintas clases. Esto puede ocurrir porque las características no discriminan correctamente a las clases, o porque hay errores en el corpus.

Se buscan primero inconsistencias en el corpus. Se busca por aquellos tweets que sean muy parecidos y tengan distinto valor de humor. Para esto, se utiliza la medida distancia de edición de Levenshtein (Jurafsky y Martin 2008, sec. 3.11), pero en lugar de hacerlo a nivel de letra se hace con una granularidad a nivel de token. Esto quiere decir que añadir un token, eliminar uno o sustituir uno son operaciones que tienen distancia uno. Se buscan aquellos pares de tweets que tengan hasta una distancia del 20% de la cantidad de tokens del tweet que tiene más tokens del par, y que uno esté etiquetado con humor y el otro no.

Se encuentran 367 pares de ejemplos que cumplen las condiciones mencionadas. Tras una revisión manual, se consta que efectivamente todos los pares tienen tweets que refieren a lo mismo y que no deberían tener distinto valor de humor. Todos estos ejemplos refieren a errores en la anotación del subcorpus de cuentas de humor. En los ejemplos (24) y (25), (26) y (27), y (28) y (29) se muestran pares de tweets que son considerados muy parecidos bajo este esquema y que fueron votados por los anotadores con distinto valor de humor. En estos y otros casos las diferencias suelen ser si es un *retweet* o no, quién hace el *retweet*, los enlaces o su cantidad, los espacios o los tipos de guiones. En algunos casos se sustituyen palabras por sinónimos o por cohipónimos (palabras que representan conceptos que son de un mismo tipo, así como naranja y pera son frutas). Se vuelve a entrenar con el corpus de entrenamiento y clasificar con el corpus de evaluación luego de remover estos ejemplos y se obtienen los resultados de la Tabla 5.7. Se obtiene una leve mejora sobre lo que originalmente se obtuvo, sobretudo en la precisión del humor.

- (24) RT @ElDatoDelDia: Video de israelíes en Tel Aviv: “¡Mañana no hay escuela, ya no quedan niños en Gaza!”: <http://t.co/eiivZhxkj2> <http://t.co...>
- (25) RT @ElDatoDelDia: Video de israelíes en Tel Aviv: “¡Mañana no hay escuela, ya no quedan niños en Gaza!”: <http://t.co/YeneLgg8Uk>

- (26) Típico :
- 1- No hacer nada.
 - 2- Darte cuenta que tenés mucho que hacer.
 - 3- Intentar organizarte.
 - 4- Agobiarte.
 - 5- Volver al paso 1.
- (27)
- 1- No hacer nada.
 - 2- Darte cuenta que tenés mucho que hacer.
 - 3- Intentar organizarte.
 - 4- Agobiarte.
 - 5- Volver al paso 1.
- (28) Limpiar tu cuarto = 1% limpieza. 30% quejarse. 69% jugar con lo que vas encontrando.
- (29) Limpiar tu cuarto:
- 1% limpieza.
- 30% quejarse.
- 69% jugar con lo que vas encontrando.

SVM	precisión	recall	F_1
No humor	94,18	97,75	95,93
Humor	85,71	69,09	76,51

Tabla 5.7: Resultados del clasificador sobre el corpus de evaluación quitando todos los tweets considerados muy parecidos pero con distinto valor de humor

Luego se procede a buscar los pares de tweets con mismos valores de características y diferentes clases, sin contar los tweets considerados parecidos anteriormente. Se encuentran treinta pares de ejemplos que cumplen las condiciones. De éstos, dieciocho son tweets parecidos al igual que los anteriores. Se decide quitarlos. El resto son pares de tweets distintos, pero de forma tal que los hacen tener las mismas características. Algunos de estos ejemplos parecen mostrar errores en la votación, pero para no interferir se dejan. Se muestran los pares de ejemplos (30) y (31), (32) y (33), (34) y (35), y (36) y (37). Se clasifica nuevamente y se obtienen los resultados de la Tabla 5.8, que son muy similares a los anteriores.

- (30) Selfie <http://t.co/MaXNX3hQZZ>
- (31) Héroe <http://t.co/uNx048NvSD>
- (32) #SabiasQue Es imposible estornudar con los ojos abiertos.

- (33) #PreguntasSinRespuesta ¿Por qué es imposible estornudar con los ojos abiertos?
- (34) #TerminarUnaNotaDeSuicidioCon Soy Darks.
- (35) #SiYoMeLlamaraKevinRoldan Me suicidaria.
- (36) RT @JohannStevonn: Yo <http://t.co/M9XZhr69ec>
- (37) Necesito a alguien que... <http://t.co/BcBxpq2DCP>

SVM	precisión	recall	F_1
No humor	94,20	97,75	95,94
Humor	85,71	69,21	76,58

Tabla 5.8: Resultados del clasificador sobre el corpus de evaluación quitando todos los tweets considerados muy parecidos pero con distinto valor de humor.

Se puede concluir que existen errores en el corpus anotado manualmente que influyen en el rendimiento del clasificador, aunque el clasificador debería poder tolerar cierta cantidad de ruido en el conjunto de entrenamiento. Se podría intentar contar la tasa de errores, haciendo una muestra por ejemplo, sin embargo implicaría interferir en el proceso de definición de qué es humor al afirmar que un ejemplo está o no bien anotado.

5.7.2. SVM

En la Tabla 5.9 se muestran los resultados de evaluar con el corpus de entrenamiento, habiendo removido todos los ejemplos encontrados que tienen un tweet muy parecido pero con distinto valor de humor (385 pares, que representan 649 tweets en total). El acierto es 93,98%. En teoría esto es lo mejor que se puede esperar del clasificador. El corpus no está logrando ser separable completamente por una curva de tipo RBF. Esto probablemente es por los errores en el corpus, y porque las características utilizadas no discriminan totalmente al humor, sino que de alguna manera lo hacen parcialmente.

SVM	precisión	recall	F_1
No humor	94,72	98,31	96,48
Humor	88,96	71,27	79,13

Tabla 5.9: Resultados del clasificador sobre el corpus de entrenamiento quitando todos los tweets considerados muy parecidos pero con distinto valor de humor

5.8. Restricción al subcorpus de cuentas de humor

Como se explica en la Construcción del Corpus (Sección 3.1), existen tweets que son extraídos de cuentas que generalmente tienen chistes y cuentas que no. Al conjunto de tweets extraído de cuentas que generalmente contienen chistes se los somete a una votación y luego se divide el corpus en conjuntos que son *humor*, *dudosos* y *no humor*. Resulta interesante evaluar el clasificador solamente tomando en consideración los tweets sacados de estas cuentas dado que

si no son de humor, como pertenecen a cuentas que generalmente tienen contenido de humor, serán parecidos a tweets humorísticos. Los resultados obtenidos restringidos al corpus de cuentas de humor para entrenar y evaluar se presentan en la Tabla 5.10. El conjunto de evaluación además es un subconjunto del conjunto de evaluación original. Podemos observar que aunque las medidas son en general inferiores a las obtenidas en los experimentos anteriores se puede ver que la precisión y el recall promedio en el clasificador con mejores resultados ronda el 80 %.

Clasificador/Medida	precisión positiva	recall positivo	F_1 positivo	precisión negativa	recall negativo	F_1 negativo	acierto
SVM	81,87	73,83	77,64	78,47	85,36	81,77	79,92
DT	74,49	75,48	74,98	72,20	71,14	71,66	74,12
GNB	78,72	78,55	78,64	76,10	76,29	76,19	77,48
MNB	68,52	85,45	76,06	83,27	64,86	72,9	74,58
kNN	79,24	73,02	76,00	77,43	82,87	80,06	78,14

Tabla 5.10: Métricas de los clasificadores sobre el conjunto de entrenamiento

5.8.1. Métricas ponderadas según calificación

Al restringirse al subcorpus humorístico todos los tweets tienen el atributo Promedio de votación (Subsección 3.2.5). Se desea obtener un resultado de la medida recall, que mide cuánto se acierta en una clase, de forma ponderada según la calificación. Tener en cuenta que la precisión carece de sentido en este caso ya que depende de los falsos positivos (recordemos que Humor es considerada la clase positiva), y por ende se vería afectada al haber menos cantidad de negativos, no pudiéndose comparar con la precisión sin ponderar. Se calcula el recall ponderado como:

$$recall_{ponderado} = \frac{\sum_{t \in vp} PV_t}{\sum_{t \in vp} PV_t + \sum_{t \in fn} PV_t} = \frac{\sum_{t \in vp} PV_t}{\sum_{t \in H} PV_t}$$

Donde vp representa a los tweets verdaderos positivos y fn a los tweets falsos negativos. Se puede ver que cuanto mayor promedio de votación tengan los tweets pertenecientes a fn , más penalizan el valor del recall.

Se puede comparar este recall con el que se obtendría originalmente e interpretar el cociente:

$$\frac{recall_{ponderado}}{recall} = \frac{\frac{\sum_{t \in vp} PV_t}{\sum_{t \in H} PV_t}}{\frac{|vp|}{|H|}} = \frac{|H| \sum_{t \in vp} PV_t}{|vp| \sum_{t \in H} PV_t} = \frac{|H| |prom_{vp}| |vp|}{|vp| |prom_H| |H|} = \frac{prom_{vp}}{prom_H}$$

Es interesante observar que cuánto aumenta esta medida respecto a la original indica cuánto el clasificador encuentra los ejemplos que tienen más estrellas que el promedio. En el caso de que estas medidas fueran independientes, esta comparación debería dar cerca de uno.

Calculando la medida para el conjunto de evaluación restringida al corpus de tweets humorísticos se obtiene el valor de 70,68 %, dos por ciento más que el *recall* obtenido sin ponderar. En la Tabla 5.11 se presenta la matriz de confusión mostrando los valores de promedio de votación para cada categoría. Se puede observar que el promedio de votación de los tweets considerados como verdaderos positivos es mayor que el promedio de votación de los tweets considerados como falsos negativos. De los tweets no humorísticos, los que son clasificados como humorísticos tienen menor promedio de votación que los correctamente clasificados como no humorísticos, sin embargo, la diferencia entre los porcentajes de humor es reducida, por lo tanto no se puede concluir algo significativo.

son/clasificados	Humor	No humor
Humor	2,7227	2,4961
No humor	0,0256	0,0300

Tabla 5.11: Matriz de confusión del promedio de votación.

5.9. Tweets censurados

Al momento de realizar la anotación del subcorpus de humor a través de la página web y de la aplicación móvil se eliminan los tweets que contienen palabras sexualmente explícitas, no quedando considerados en el uso clasificador. Por lo tanto, se procede a incluir los tweets removidos y anotarlos manualmente. Al ejecutar el algoritmo con los tweets censurados con el clasificador SVM se obtienen los resultados denotados en la Tabla 5.12. El acierto es 92,32 %.

SVM	precisión	recall	F_1
No humor	93,74	97,17	95,43
Humor	84,01	69,59	76,13

Tabla 5.12: Métricas para el clasificador SVM agregando los tweets censurados.

Sin embargo, al realizar nuevamente los estudios de importancia de características se puede ver que no varía la importancia de la característica *Jerga Sexual* al agregar los tweets censurados. En conclusión, las mejoras en precisión y recall son el resultado de agregar nuevos chistes que tienen otro tipo de característica además de Jerga sexual, que son clasificados correctamente.

5.10. Diferentes subconjuntos no humorísticos

Como se explica en la conformación del corpus (Sección 3.1) los tweets extraídos de cuentas no humorísticas se dividen en Noticias, Reflexiones y Curiosidades. En esta parte se evalúa los clasificadores entre los tweets anotados como humorísticos y cada subconjunto de tweets perteneciente a cada una de las categorías. De esta forma se puede saber cuál es la categoría que más cuesta diferenciar de los tweets de humor. Los resultados obtenidos se presentan en la Tabla 5.13. El acierto con Noticias es 96,97%, con Reflexiones es 88,92% y con Curiosidades es 90,74%. Se puede observar que la categoría que más se diferencia de los humorísticos es Noticias. Respecto al conjunto Reflexiones y Curiosidades no se puede encontrar una diferencia notoria. Sin embargo, mirando el recall, los resultados restringidos a cada subconjunto son mejores que los del clasificador utilizando todo el corpus.

SVM	precisión	recall	F_1
No humor	96,95	98,12	97,53
Humor	97,00	95,18	96,08

(a) Métricas restringidas a *Noticias*

SVM	precisión	recall	F_1
No humor	83,95	95,27	89,25
Humor	94,95	82,99	88,57

(b) Métricas restringidas a *Reflexiones*

SVM	precisión	recall	F_1
No humor	88,24	96,26	92,08
Humor	94,64	83,73	88,85

(c) Métricas restringidas a *Curiosidades*

Tabla 5.13: Métricas del clasificador para los subconjuntos: Noticias, Reflexiones, Curiosidades

5.11. *Tweets* dudosos

Resulta atractivo reportar también la cantidad de tweets que el clasificador considera humor en los ejemplos considerados dudosos en la etapa de construcción del corpus. El clasificador predice que hay un 21% de los tweets de humor dentro de los dudosos. Recordemos que son 3478 la cantidad de ejemplos dudosos, incluyendo 1412 tweets provenientes de cuentas de humor que no tienen votos.

5.12. Independencia y combinación de características

Los distintos tipos de clasificadores dan diferentes resultados. Los clasificadores Naïve Bayes suponen por definición que los atributos son independientes entre sí, cuando se cree que no lo son. Por ejemplo, la probabilidad del valor de la característica *Diálogo* aumenta si el valor de la característica *Preguntas-respuestas* de un tweet es alto. Los demás clasificadores tratan mejor la interdependencia entre atributos. SVM, por ejemplo, al buscar la curva que mejor separe las instancias indirectamente las está agrupando, y las características tienen la capacidad de actuar juntas. kNN al mirar las instancias cercanas entre sí, también hace lo suyo. DT, por otra parte, al separar en ramas los distintos valores de atributos, también permite que valores de distintos atributos actúen juntos. Un ejemplo de que acción colaborativa es que valores altos de la característica Segunda persona pueden asegurar con menor margen de error si un tweet en cuestión es realmente un diálogo, indicando más informalidad, y aumentando la probabilidad de que sea humor.

Más en general, hay características que parecen inútiles, pero actuando juntas revelan patrones interesantes. Un ejemplo simple es el de tratar de aproximar la función XOR (“o” exclusivo). Se puede tener un atributo que indique el valor de una de las variables, y otro atributo que indique el de la otra, y que por separado carezcan de valor. Más aún, si se hace una prueba de independencia dará que son independientes a la función que se quiere aproximar, ya que el valor de las características por separado no discrimina el resultado, aunque sean las únicas variables consideradas por la verdadera función. Sin embargo, actuando juntas es que se obtiene dicha función.

Capítulo 6

Conclusiones

Se logra relevar un estado del arte, estudiando trabajos anteriores similares. Han habido algunos pocos trabajos que han intentado identificar al humor, y lo han logrado pero con versiones simplificadas de la problemática, de manera análoga que el presente trabajo. A su vez no hay una definición única aceptada sobre qué es el humor; hay varias visiones que ninguna cubre la totalidad pero se complementan bien.

Muchas votaciones permiten haber elaborado un corpus que se utiliza para el clasificador, en base a tweets considerados no dudosos para los anotadores.

Se construye un clasificador con resultados que superan la línea base trazada. El mejor clasificador construido tiene como valor de precisión 83,61% y un valor de recall de 68,85%, más allá de los errores que pueda tener el corpus. El aprendizaje automático supervisado estadístico parece haber sido una buena técnica para encarar el problema. Sin embargo, hay que notar que el corpus utilizado no tiene demasiada diversidad de humor, y por lo tanto funcionan características que no definen en su esencia al humor y que en otros contextos no servirían tanto.

La característica Diálogo parece ser la más relevante para la discriminación de humor por todos los métodos de selección de característica utilizados. Las características que peor discriminan el humor son la presencia de palabras no españolas, las palabras fuera de vocabulario y la presencia de antónimos.

Se cree que para las características construidas, la hipótesis de independencia entre ellas no se cumple. Por lo tanto, como muestran los resultados, los clasificadores que no suponen independencia y son capaces de vincular diferentes características dan mejor resultado. En el presente proyecto se combinan características simples referentes al formato del texto y la temática de su contenido. Se presume que la buena combinación de características más complejas de formato de texto y que identifiquen temas típicos en textos humorísticos, llevan a mejores resultados.

Se comprueba que en general existe una leve mejora del clasificador cuando se clasifica un tweet con promedio de votación mayor. La categoría que más se logra diferenciar de los tweets humorísticos son los tweets extraídos de medios de prensa y los más difíciles son los tweets extraídos de cuentas de reflexiones y curiosidades.

La selección de características resulta en este caso poco cambio en las medidas arrojadas. Se cree que todas las características aportan algo a la discriminación del humor, por tanto sacarlas en general empeora los resultados.

Por otro lado hay inconsistencias en el corpus: se encuentran cientos de pares que implican incoherencias. No se le puede pedir un gran desempeño al clasificador si llegara a tener demasiados errores. Tampoco si no hay una definición clara de lo que se quiere predecir. Obviamente más cantidad de votos hubieran sido de mayor utilidad.

El Reconocimiento de humor no es un área del Procesamiento del Lenguaje Natural en el que abunden recursos especializados como corpus de humor, menos teniendo en cuenta que se está trabajando para el idioma español. Se cree que en este trabajo se han realizado diferentes tipos de recursos que son de gran valor para futuros proyectos como el corpus etiquetado según humor o no por una votación con más de 30.000 votos y la creación de algunos diccionarios, además de la implementación en sí.

Al igual que los trabajos estudiados, las noticias son las más fáciles de distinguir del humor respecto a otros conjuntos considerados. A diferencia de los trabajos anteriores, Naïve Bayes y SVM sí dan diferencias significativas. Como se menciona en la Sección 5.12, parte de esto es atribuido a que no se cumplen las suposiciones de los clasificadores bayesianos en este caso. Otra diferencia que se tiene con los autores es que Palabras clave no funciona tan bien como a ellos.

6.1. Trabajo futuro

Se presume que las características elaboradas son en su mayoría simples. Como trabajo a futuro se podrían hacer características más complejas, como intentar detectar juegos de palabras, ambigüedad, perplejidad de un modelo de lenguaje como lo son los n-gramas, etc.

GNB y MNB suponen cada uno que los atributos siguen una distribución gaussiana y multinomial respectivamente, y esto implica que se trate a los datos como continuos o como discretos, cuando en realidad existen atributos pertenecientes a ambos mundos. *Links* es discreto, ya que cuenta la cantidad de hipervínculos, por ejemplo. *Jerga sexual* está normalizado según la cantidad de tokens del tweet, por lo tanto es continuo. Se podría construir un clasificador que trate a los atributos discretos con distribución multinomial y a los continuos con distribución gaussiana.

A su vez sería de interés ver si se puede hacer regresión en el promedio de estrellas de un tweet. Es decir, si se puede aproximar la función que dado un tweet diga el promedio de estrellas que tiene. Esto indicaría de cierta manera qué tan gracioso le parece un tweet a la gente en general. Podría utilizarse la cantidad de *retweets* o favoritos que tiene, aunque a priori parece ser algo inútil, ya que ejemplos que no tienen nada que ver con el humor pueden tener toda la variedad posible de favoritos y *retweets*, y los que sí tienen que ver con el humor también.

Asimismo sería de interés poder estudiar cómo influye la clasificación de humor en los distintos estratos sociales, ya sea por personas de distinto género, de distinta edad, diferentes áreas de interés, o inclusive estudiar si varía en distintos momentos para la misma persona. Se cree que la edad es un factor muy determinante en la clasificación de este corpus, ya que por ejemplo tweets de humor verde pueden ser considerados muy graciosos por adolescentes y tal vez muy malos, o peor aún como no humorísticos, por personas de edad mayor.

Glosario

- **Bag of Words** Forma de representar las instancias, en donde cada ejemplo es convertido en un vector de conteo de las apariciones de sus palabras, tomándose por lo tanto como características a la cantidad de veces que aparece cada palabra. 36, 52
- **BoW** Bag of Words. 36, 52, *Glosario*: Bag of Words
- **característica** *Característica* (o *atributo*, o en inglés *feature* o *attribute*) es una propiedad medible de una instancia que se observa. 8
- **corpus** Conjunto habitualmente muy amplio de ejemplos reales de uso de una lengua. 2
- **DT** Decision Tree. 28
- **ejemplo** Instancia perteneciente a un conjunto de datos que se quiere estudiar. 23, 52
- **GNB** Gaussian Naïve Bayes. 30
- **histograma** Representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. 17
- **instancia** Elemento perteneciente al dominio definido de un tópico que es estudiado. 23, 52
- **kNN** k Nearest Neighbors. 30
- **MNB** Multinomial Naïve Bayes. 30
- **SVM** Support Vector Machine. 26
- **tokenización** Proceso de separar un texto en *tokens* (o símbolos). 36
- **tweet** Mensaje de Twitter, que tiene un máximo de 140 caracteres. 1

Apéndice

Apéndice A

Arquitectura del clasificador

En el presente anexo se detalla la arquitectura de software utilizada para la construcción del clasificador. Se describen los componentes que conforman la arquitectura y las principales clases de software que componen al clasificador.

Componentes

En la Figura A.1 se presenta un diagrama de componentes del clasificador. A continuación se describen las componentes:

- **Servidor Freeling (Tokenizador-Etiquetador):** es un servidor Freeling con la responsabilidad de tokenizar y etiquetar gramaticalmente texto.
- **Servidor Freeling (Diccionario):** es un servidor Freeling con la responsabilidad de indicar si una palabra se encuentra en el diccionario.
- **Google:** servidor que brinda servicio de búsqueda.
- **Wiktionary:** servidor que brinda servicio de búsqueda de palabras en un diccionario.
- **Clasificador:** lógica principal del clasificador. Calcula las características y es el encargado de ejecutar los algoritmos de Aprendizaje Automático.
- **Base de datos (corpus):** base de datos que almacena el corpus extraído.
- **Base de datos (chistes.com):** base de datos que almacena los chistes de Chistes.com (2015), en donde están agrupados por categoría (*Machistas*, *Borrachos*, etc.).

Clasificador

En la Figura A.2 se presenta un diagrama de clases para la componente Clasificador. Como principales clases se puede distinguir:

- **Feature:** clase abstracta que encapsula el concepto de característica. Toda característica del clasificador debe heredar de esta clase abstracta e implementar el método `calcular-feature(tweet) : double`, que calcula el valor de la característica a partir de un tweet.
- **Persistencia:** clase encargada de consumir la información de las bases de datos. Las funciones principales son traer el corpus de *tweets* extraídos y el corpus de Chistes.com a memoria.

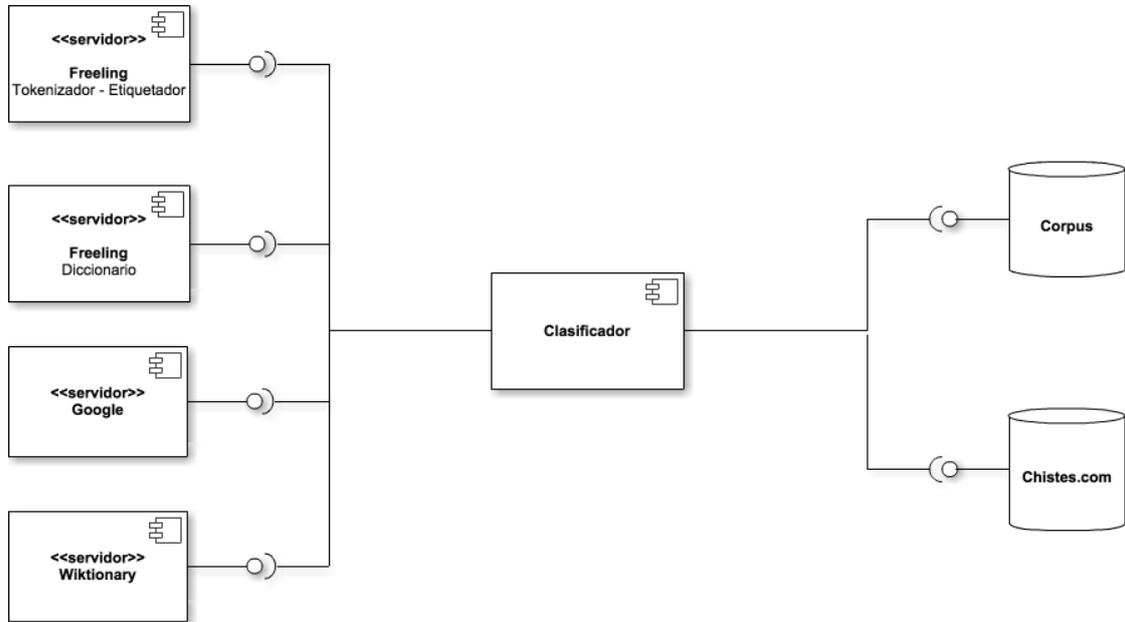


Figura A.1: Diagrama de componentes

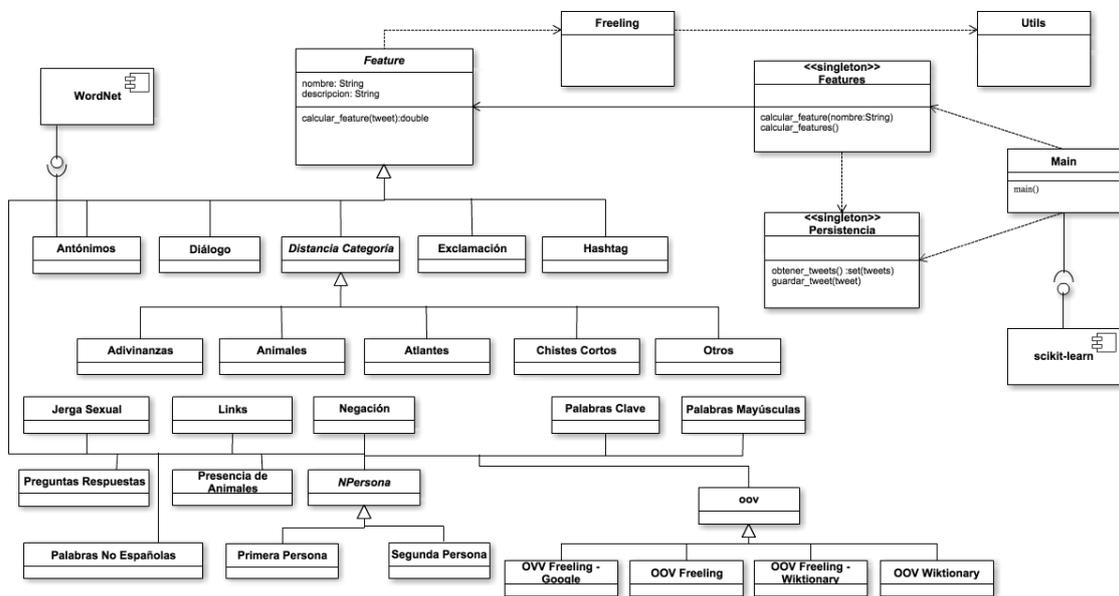


Figura A.2: Diagrama de clases

- **Freeling:** se trata de una clase encargada de enviar pedidos a los servidores Freeling y de procesar las respuestas.
- **Utils:** módulo que brinda funcionalidades auxiliares para diferentes partes del código. Entre las funcionalidades que implementa se encuentran: obtener diccionarios, ejecutar comandos en consola, entre otras.
- **Features:** clase *singleton* encargada de manejar las diferentes características. Las funciones principales son calcular todas las características para todos los tweets y calcular una característica particular para cada tweet.
- **Main:** módulo encargado de ejecutar los métodos de Aprendizaje Automático, utilizando la biblioteca scikit-learn (2014). Procesa las diferentes opciones del clasificador como el tipo de clasificador a utilizar, realizar validación cruzada, entre otras acciones.

Apéndice B

Ajuste de parámetros

En el presente anexo se detalla el proceso de ajuste de parámetros a cada método de Aprendizaje Automático, a excepción del algoritmo gaussiano de Naïve Bayes que no acepta ninguno. Para esto se utiliza el algoritmo *Grid Search*, que se basa en entrenar y clasificar con cada combinación de parámetros deseada, utilizando fuerza bruta.

Se obtienen los mismos parámetros que los que se establecen por defecto, para todos los tipos de clasificadores, no pudiendo mejorar los resultados.

Support Vector Machine

En este método se pueden seleccionar los siguientes parámetros (scikit-learn 2015g):

- **Núcleo:** transformación que se le hace al espacio utilizado por SVM para lograr el tipo de curva deseado, modificando el hiperplano original (representado por $\langle \mathbf{x}, \mathbf{x}' \rangle$, siendo \mathbf{x} y \mathbf{x}' dos vectores).
 - **RBF (*radial basis function*):** una función de la forma $e^{-\gamma\|\mathbf{x}-\mathbf{x}'\|^2}$
 - **Polinomial:** $(\gamma\langle \mathbf{x}, \mathbf{x}' \rangle + r)^d$, polinomio en \mathbb{R}^n , siendo n la cantidad de características.
 - **Lineal:** $\langle \mathbf{x}, \mathbf{x}' \rangle$, hiperplano en \mathbb{R}^n .
 - **Sigmoid:** $\tanh(\gamma\langle \mathbf{x}, \mathbf{x}' \rangle + r)$
- **Gamma:** parámetro γ en el núcleo, si aplica.
- **Término independiente:** parámetro r en el núcleo, si aplica.
- **Grado:** parámetro d en el núcleo, si aplica.
- **Parámetro de penalización C:** Este parámetro influye en cuánto la curva se ajusta al conjunto de entrenamiento. Si el parámetro toma un valor grande se aumenta la penalización de los puntos que quedan mal clasificados según una curva candidata. Notar que ejemplos erróneos afectarían negativamente la curva si este parámetro tiene valores altos.

Se realiza el algoritmo Grid Search con los parámetros:

- Núcleo = $\{RBF, Polinomial, Lineal, Sigmoid\}$
- Gamma = $\left\{ \frac{1}{\#\{características\}}; 0,03; 0,04; 0,05; 0,06 \right\}$

- Grado = {3; 4; 5}
- Término independiente = {-1, 0, 1}
- Parametro de penalización C = {0,9; 1,0}

Obteniendo como mejores parámetros el núcleo *RBF*, parámetro Gamma = $\frac{1}{\#\{caractersticas\}}$ y parámetro C = 1,0.

Árbol de decisión

En este método se pueden seleccionar los siguientes parámetros (scikit-learn 2015f):

- **Criterio de partición:** función que mide la efectividad de la partición.
 - **Gini:** se utiliza la Impureza de Gini como función.
 - **Entropía:** se utiliza la entropía como función.
- **Estrategia de separación:** estrategia utilizada para elegir la partición en cada nodo.
 - **Óptima:** se elige la partición óptima en cada nodo, según el criterio de partición.
 - **Aleatoria:** se elige la mejor partición de un conjunto de particiones aleatorias.
- **Máxima cantidad de características:** indica la máxima cantidad de características a considerar.
- **Máxima profundidad:** altura máxima del árbol de decisión generado.
- **Cantidad mínima de ejemplos para una partición:** indica el mínimo número de ejemplos requerido para dividir un nodo.
- **Mínimo de ejemplos por hoja:** indica el mínimo número de ejemplos requeridos para que el nodo sea considerado una hoja.
- **Máxima cantidad de hojas:** limita la cantidad de hojas que puede tener el árbol generado. En caso de ser fijado se ignora el parámetro de altura máxima.

Se ejecuta el algoritmo Grid Search con los parámetros:

- Criterio de partición = {*Gini*, *Entropia*}
- Estrategia de separación = {*Óptima*, *Aleatoria*}
- Máxima cantidad de características = {17, 18, 19, 20, 21}
- Máxima profundidad = {25, 26, 27, 28, 29, 30, +∞}
- Cantidad mínima de ejemplos para una partición = {1, 2, 3, 4, 5}
- Mínimo de ejemplos por hojas = {1, 2, 3, 4}
- Máxima cantidad de hoja = {49, 50, 51, +∞}

Obteniendo como mejores parámetros Criterio de partición = *Gini*, Estrategia de separación = *Óptima*, Máxima cantidad de características = +∞, Máxima profundidad = +∞, Cantidad mínima de ejemplos para una partición = 2, Mínimo de ejemplos por hojas = 1, Máxima cantidad de hojas = +∞.

Naïve Bayes multinomial

En este método se pueden seleccionar los siguientes parámetros (scikit-learn 2015d):

- **Alfa:** parámetro aditivo de suavizado, para valores de características no vistos.
- **Ajustar probabilidades a priori:** si se calculan las probabilidades a priori de las clases a partir de los datos o no.

Se realiza el algoritmo Grid Search con los parámetros:

- Alfa = $\{0; 0,05; 0,1; 0,15; 0,2; 0,25; \dots; 0,8; 0,85; 0,9; 0,95; 1\}$
- Ajustar probabilidades a priori = $\{No, Si\}$

Obteniendo como mejores parámetros Alfa = 1 y Ajustar probabilidades a priori = *Si*.

k Nearest Neighbors

En este método se pueden seleccionar los siguientes parámetros (scikit-learn 2015e):

- **Número de vecinos:** cantidad de vecinos más cercanos a considerar.
- **Pesos:** ponderación que se le asigna a los vecinos.
 - **Uniforme:** todos los ejemplos vecinos tienen el mismo peso.
 - **Distancia:** el peso asignado es el inverso del cuadrado de la distancia a un ejemplo dado.
- **Métrica:** medida de distancia utilizada. Por defecto es la medida de Minkowski ($\sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p}$), pudiendo cambiarla por una personalizada.
- **p:** parámetro p en la métrica. Con $p = 2$ se utiliza la distancia euclideana.

Se realiza el algoritmo Grid Search con los parámetros:

- Número de vecinos = $\{4, 5, 6, 7, 8\}$
- Pesos = $\{Uniforme, Distancia\}$
- Métrica = $\{Minkowski\}$
- $p = \{1, 2, +\infty\}$

Obteniendo como mejores parámetros Números de vecinos = 5, Pesos = *Uniforme*, Métrica = *Minkowski* y $p = 2$.

Apéndice C

Uso del programa clasificador

En el presente anexo se presenta como utilizar programa clasificador. Al ejecutar el clasificador con la bandera *help* se obtiene la sugerencia de como utilizar el clasificador que se presenta a continuación.

```
uso: main.py [-h] [-a] [-c {DT,GNB,kNN,LinearSVM,MNB,SGD,SVM}]
            [-x] [-D] [-e] [-b] [-j] [-k] [-g] [-G {1,2,3,4,5}]
            [-i] [-z] [-l LIMITE] [-m] [-q] [-p] [-n] [-s]
            [-f NOMBRE_FEATURE] [-d] [-r] [-E] [-S]
            [-N {Noticias,Curiosidades,Reflexiones}]
            [-t THREADS] [-o]
```

Clasifica humor de los tweets almacenados en la base de datos.

optional arguments:

- h, --help
muestra esta ayuda y termina.
- a, --calcular-features-faltantes
calcula el valor de todas las features para los tweets a los que les falta calcularla
- c {DT,GNB,kNN,LinearSVM,MNB,SGD,SVM},
--clasificador {DT, GNB, kNN, LinearSVM, MNB, SGD, SVM}
establece qué tipo de clasificador es usado, que por defecto es SVM
- x, --cross-validation
para hacer validación cruzada
- D, --dudosos
clasifica los tweets dudosos
- e, --evaluar
para evaluar con el corpus de evaluación

-b, --explicar-features
muestra las features disponibles y termina el programa

-j, --feature-aleatoria
agrega una feature con un valor binario aleatorio

-k, --feature-clase
agrega una feature cuyo valor es igual a la clase objetivo

-g, --grid-search
realiza el algoritmo grid search para el ajuste de hiperparámetros

-G {1,2,3,4,5}, --grupo-de-calificacion {1,2,3,4,5}
establece a qué grupo de promedio de humor restringir el corpus

-i, --importancias-features
reporta la importancia de cada feature

-z, --incluir-chistes-sexuales
incluye en el entrenamiento y en la evaluación los chistes con contenido sexual

-l LIMITE, --limite LIMITE
establece una cantidad límite de tweets a procesar

-m, --mismas-features-distinto-humor
imprime los tweets que tienen los mismos valores de features pero distinto valor de humor

-q, --medidas-ponderadas
imprime las medidas precisión, recall y F1 ponderadas según el promedio de humor

-p, --parametros-clasificador
lista los parámetros posibles para un clasificador

-n, --ponderar-segun-votos
en la clasificación pondera los tweets según la concordancia en la votación. Funciona sólo para DT y SVM

-s, --recalcular-features
recalcula el valor de todas las features

-f NOMBRE_FEATURE, --recalcular-feature NOMBRE_FEATURE

recalcula el valor de una feature

-d, --rfe
habilita el uso de Recursive Feature Elimination antes de clasificar

-r, --servidor
levanta el servidor para responder a clasificaciones

-E, --sin-escalar
establece si no deben escalarse las características

-S, --solo-subcorpus-humor
entrena y evalúa solamente en el corpus de humor

-N {Noticias, Curiosidades, Reflexiones},
--subconjunto-no-humor {Noticias, Curiosidades, Reflexiones}
selecciona solamente el subconjunto pasado como parámetro del corpus no humorístico

-t THREADS, --threads THREADS
establece la cantidad de threads a usar al recalcular las features

-o, --tweets-parecidos-distinto-humor
busca y quita los tweets que son parecidos pero tienen distinto valor de humor

A continuación se presenta la ejecución del clasificador evaluando en el corpus de entrenamiento y en el de evaluación (se ejecuta: ./main.py --evaluar):

Entrenando clasificador...

Evaluando clasificador con el conjunto de entrenamiento...

Acierto: 0.9328

	precision	recall	f1-score	support
No humor	0.9416	0.9801	0.9605	23660
Humor	0.8746	0.6961	0.7752	4729
avg / total	0.9081	0.8381	0.8678	28389

Matriz de confusión:

		(clasificados como)	
		Humor	No humor
(son)	Humor	3292	1437
(son)	No humor	472	23188

Evaluando clasificador...

Acierto: 0.9241

	precision	recall	f1-score	support
No humor	0.9384	0.9724	0.9551	5970
Humor	0.8361	0.6885	0.7552	1223
avg / total	0.8873	0.8304	0.8551	7193

Matriz de confusión:

		(clasificados como)	
		Humor	No humor
(son)	Humor	842	381
(son)	No humor	165	5805

Bibliografía

- Basili, Roberto y Fabio Massimo Zanzotto (2002). «Parsing Engineering and Empirical Robustness». En: *Natural Language Engineering* 8.3, págs. 97-120.
- Chistes.com (2015). *Chistes.com | Los mejores chistes de la web*. URL: <http://www.chistes.com> (accedido en febrero de 2015).
- Cohen, Jacob (1960). «A Coefficient of Agreement for Nominal Scales». En: *Educational and Psychological Measurement* 20.1, págs. 37-46. ISSN: 0013-1644.
- Fleiss, Joseph L (1971). «Measuring nominal scale agreement among many raters». En: *Psychological bulletin* 76.5, pág. 378.
- Freud, S. y J. Strachey (1905). *Jokes and Their Relation to the Unconscious*. Complete Psychological Works of Sigmund Freud. ISBN: 9780393001457.
- Google (2015). *Google*. URL: <https://www.google.com.uy> (accedido en enero de 2015).
- Gruner, C.R. (2000). *The Game of Humor: A Comprehensive Theory of Why We Laugh*. Transaction Publishers. ISBN: 9780765806598.
- Guyon, Isabelle et al. (2002). «Gene Selection for Cancer Classification using Support Vector Machines». En: *Machine Learning* 46.1-3, págs. 389-422.
- Gwet, Kilem (2014). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC.
- International Journal of Humor Research (1988). *HUMOR*. URL: <http://www.degruyter.com/view/j/humr> (accedido en enero de 2015).
- Jurafsky, Dan y James Martin (2008). *Speech & Language Processing*. 2.^a ed. Prentice Hall.
- Mihalcea, Rada y Stephen Pulman (2007). *Characterizing Humour: An Exploration of Features in Humorous Texts*.
- Mihalcea, Rada y Carlo Strapparava (2005). «Making Computers Laugh: Investigations in Automatic Humor Recognition». En: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. HLT '05. Vancouver, British Columbia, Canada: Association for Computational Linguistics, págs. 531-538.
- (2010). «Learning to Laugh (automatically): Computational Models for Humor Recognition». En: *Computational Intelligence* 22.2, págs. 126-142. URL: <http://dblp.uni-trier.de/db/journals/ci/ci22.html#MihalceaS06>.
- Mitchell, Tom (1997). *Machine Learning*. McGraw Hill.
- Mulder, M.P. y Antinus Nijholt (2002). *Humour Research: State of Art*. Imported from CTIT. Enschede, the Netherlands. URL: <http://doc.utwente.nl/63066>.
- Padró, Lluís y Evgeny Stanilovsky (2012). «FreeLing 3.0: Towards Wider Multilinguality». En: *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. ELRA. Istanbul, Turkey.
- Raskin, Victor (1985). *Semantic Mechanisms of Humor*. Studies in Linguistics and Philosophy. Springer. ISBN: 9789027718211.

- Real Academia Española (2001). *Diccionario de la lengua española*. URL: <http://www.rae.es/recursos/diccionarios/drae> (accedido en julio de 2014).
- Reyes, Antonio, Davide Buscaldi y Paolo Rosso (2009). «An Analysis of the Impact of Ambiguity on Automatic Humour Recognition». En: *TSD*. Ed. por Václav Matousek y Pavel Mautner. Vol. 5729. Lecture Notes in Computer Science. Springer, págs. 162-169. ISBN: 978-3-642-04207-2.
- Reyes, Antonio, Paolo Rosso et al. (2009). «Características y rasgos afectivos del humor: un estudio de reconocimiento automático del humor en textos escolares en catalán». En: *Procesamiento del Lenguaje Natural* 43, págs. 235-243.
- Ritchie, Graeme (1999). «Developing the Incongruity-Resolution Theory». En: *Proceedings of the AISB Symposium on Creative Language: Stories and Humour*, págs. 78-85.
- Ruch, Willibald, Salvatore Attardo y Victor Raskin (1993). «Toward an Empirical Verification of the General Theory of Verbal Humor». En: *HUMOR: the International Journal of Humor Research*.
- Rutter, Jason (1997). «Stand-up as Interaction: Performance and Audience in Comedy Venues». Tesis doct. Citeaser.
- scikit-learn (2015a). *Decision Trees*. URL: <http://scikit-learn.org/stable/modules/tree.html> (accedido en febrero de 2015).
- (2015b). *Naive Bayes*. URL: http://scikit-learn.org/stable/modules/naive_bayes.html (accedido en febrero de 2015).
- (2015c). *Naive Bayes*. URL: <http://scikit-learn.org/stable/modules/neighbors.html> (accedido en febrero de 2015).
- (2014). *scikit-learn*. URL: <http://scikit-learn.org> (accedido en julio de 2014).
- (2015d). *sklearn.naive_bayes.MultinomialNB - scikit-learn documentation*. URL: http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html (accedido en febrero de 2015).
- (2015e). *sklearn.neighbors.KNeighborsClassifier - scikit-learn documentation*. URL: <http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html> (accedido en febrero de 2015).
- (2015f). *sklearn.tree.DecisionTreeClassifier - scikit-learn documentation*. URL: <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> (accedido en febrero de 2015).
- (2015g). *Support Vector Machine*. URL: <http://scikit-learn.org/stable/modules/svm.html> (accedido en febrero de 2015).
- Shalev-Shwartz, S. y S. Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. ISBN: 9781107057135.
- Sjöbergh, Jonas y Kenji Araki (2007). «Recognizing Humor Without Recognizing Meaning». En: *WILF*. Ed. por Francesco Masulli, Sushmita Mitra y Gabriella Pasi. Vol. 4578. Lecture Notes in Computer Science. Springer, págs. 469-476. ISBN: 978-3-540-73399-7. URL: <http://dblp.uni-trier.de/db/conf/wilf/wilf2007.html#SjoberghA07>.
- Suslov, I.M. (1992). «Computer Model of a “Sense of Humor” I. General Algorithm». En: *Biofizika* 37.2, págs. 318-324.
- Veatch, Thomas (1998). «A Theory of Humor». En: *HUMOR: The International Journal of Humor Research* 11-2, págs. 161-215.
- Wiktionary (2015). *Wiktionary*. URL: <https://www.wiktionary.org> (accedido en enero de 2015).