CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR DE ENSENADA, BAJA CALIFORNIA

PROGRAMA DE POSGRADO EN CIENCIAS EN CIENCIAS DE LA COMPUTACIÓN

Diseño de algoritmos bioinspirados para la selección de características en el análisis de sentimientos de documentos en español

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Maestro en Ciencias

Presenta:

Rosa Alejandra Ortega del Castillo

Ensenada, Baja California, México

2015

Tesis defendida por

Rosa Alejandra Ortega del Castillo

y aprobada por el siguiente comité

Dr. Carlos Alberto Brizuela Rodríguez
Codirector del Comité

Dr. Hugo Homero Hidalgo Silva
Codirector del Comité

Dra. Mónica Elizabeth Tentori Espinosa
Miembro del Comité

Dr. Israel Marck Martínez Pérez
Miembro del Comité

Dr. Miguel Ángel Alonso Arevalo
Miembro del Comité

Dra. Ana Isabel Martínez García
Coordinador del Programa de Posgrado en Ciencias de la Computación

Dr. Jesús Favela Vara
Director de Estudios de Posgrado

Febrero, 2015

Resumen de la tesis que presenta Rosa Alejandra Ortega del Castillo como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la Computación.

Diseño de algoritmos bioinspirados para la selección de características en el análisis de sentimientos de documentos en español

Resumen elaborado por:

Rosa Alejandra Ortega del Castillo

El análisis de sentimientos o minería de opiniones consiste en la clasificación de documentos que expresan una opinión, separándolos según el sentimiento que expresan. Para realizar esta clasificación, bajo un enfoque de aprendizaje de máquina, es necesario definir un conjunto de características que se usarán para representar a cada documento. En general, el número de características que se puede extraer de los documentos es elevado y manejarlas para la tarea de clasificación se vuelve un problema computacionalmente costoso. Aunado a esto, usar todas las características posibles no necesariamente garantizará una máxima precisión en la clasificación ya que varias características podrían no estar relacionadas con la clase que se supone debe definir.

En este trabajo se propone un enfoque bioinspirado para la selección de características con la finalidad de encontrar el subconjunto óptimo del conjunto total de características que dé la mejor precisión de clasificación. Para esto se utiliza un algoritmo genético (AG) para la generación y búsqueda de los posibles subconjuntos de características aunado a una máquina de soporte vectorial (SVM) para evaluar la calidad del subconjunto seleccionado. El algoritmo recibe como entrada las características presentes en un conjunto de documentos seleccionado y el número de generaciones que se desea que itere el algoritmo. La salida del algoritmo es el mejor subconjunto de características encontrado que brinda la mejor precisión de clasificación. Se ha diseñado la representación de los individuos y los operadores genéticos que resuelven este problema en particular.

Para la realización de los distintos experimentos se ha utilizado un corpus en idioma español de críticas de cine recogidas de la web *muchocine*. Este cuenta con un conjunto de 2624 documentos, 1274 con una opinión positiva y 1350 con opinión negativa.

Bajo este enfoque se ha logrado obtener un subconjunto de características que da una precisión de 91.5 % con los documentos de prueba en español. Sin embargo, con documentos nuevos esta precisión baja a 77 %. Aprovechando 11 modelos diferentes que genera el algoritmo genético en diferentes corridas se construyó un meta clasificador por consenso con el cual se logró mejorar la precisión de cada modelo por separado.

Palabras Clave: **Análisis de sentimientos, minería de opiniones, selección de características, clasificación, documentos de opinión**

Abstract of the thesis presented by Rosa Alejandra Ortega del Castillo as a partial requirement to obtain the Master of Science degree in Master in Sciences in Computer Science.

Design of Bio-inspired Algorithms for Feature Selection on Sentiment Analysis of Documents in Spanish

Abstract by:

Rosa Alejandra Ortega del Castillo

Sentiment analysis or opinion mining involves the classification of documents expressing an opinion, separating them according to the sentiment expressed. To perform this classification, with a focus on machine learning, it is necessary to define a set of features that will be used to represent each document. Overall, the number of features that can be extracted from the documents is high and to manage them for the classification task becomes a computationally expensive problem. Added to this, using all possible features not necessarily guarantee a high classification accuracy as several features may not be related to the class that is supposed to define.

This paper presents a bio-inspired approach to feature selection in order to find the optimal subset of the total set of features that give the best classification accuracy. For this a genetic algorithm (GA) for generating and searching for possible subsets of features coupled with a support vector machine (SVM) to evaluate the quality of the selected subset is used. The algorithm receives as input the features present in a set of selected documents and the number of generations you want to iterate the algorithm. The output of the algorithm is the best subset of features found which gives the best classification accuracy. We have designed the representation of individuals and genetic operators that solve this particular problem.

To carry out the various experiments we used a corpus in Spanish movie reviews collected from the web *muchocine*. This has a set of 2624 documents, 1274 with a positive review and 1350 with negative opinion.

Under this approach it has managed to obtain a subset of features that gives an accuracy of 91.5 % with the test documents in Spanish. However, with new documents this precision decreases down to 77 %. Taking advantage of 11 different models generated by the genetic algorithm in different runs a meta classifier by consensus was constructed with which was possible to improve the accuracy of each model separately.

Keywords: **Sentiment analysis, opinion mining, feature selection, classification, opinion documents**

Dedicatoria

A mis padres Ana y Ricardo.

A mi novio Esteban.

Agradecimientos

A mis padres Ana Del Castillo y Ricardo Ortega por su apoyo incondicional, su cariño y su paciencia, por ser una inspiración para seguir estudiando y motivarme a hacerlo.

A mi novio Esteban Ontiveros por escucharme, ser paciente, por cuidarme, por no dejar que me saliera de mi camino y ayudarme en cada aspecto de mi vida profesional y personal.

A mis hermanos, a mis primos, a mis abuelitos y a toda mi familia por sus bendiciones y buenos deseos.

A mis asesores de tesis Carlos Brizuela y Hugo Hidalgo por guiarme durante este trabajo de tesis, del mismo modo al resto de mi comité por sus valiosas observaciones.

A mis roomies, a mis compañeros del taller de teatro y a mis compañeros de generación por todos eso gratos momentos y escapes del estrés.

A mi amiga Lyla Morales por darme ánimos y mostrar siempre interés enviándome enlaces de artículos similares a mi trabajo.

Al Centro de Investigación Científica y de Educación Superior de Ensenada.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de maestría.

Tabla de contenido

Página

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de tablas

Lista de tablas (continuación)

Capítulo 1. Introducción

1.1. Antecedentes y motivación

En la actualidad es muy fácil hacer todo tipo de consultas en la red de Internet, más aún con el gran crecimiento de los servicios de *microblogging* y redes sociales. Podemos asesorarnos de las opiniones de gente de distintas partes del mundo para tomar una decisión a la hora de realizar una compra, visitar un nuevo lugar, ver una película, etcétera. Este tipo de consultas también es útil para los políticos que desean conocer las opiniones de sus votantes potenciales o para las empresas que desean conocer las opiniones de sus clientes potenciales.

Internet se ha convertido en una fuente masiva y continua de opiniones. Estas opiniones, que están en lenguaje natural, son de fácil entendimiento para las personas, pero sería imposible leer todas estas para darse una idea genérica de lo que el público en general opina. Hay sitios como *Twitter.com, Foursquare.com, Labutaca.net, Ciao.es, Dooyoo.es* que son más comunes para hacer este tipo de consultas, sin embargo, solo tenemos tiempo para leer unas cuantas reseñas y tomar una decisión con base en ellas.

Es necesario utilizar técnicas de clasificación automatizadas que permitan manejar toda esta información, y así poder organizarla para los usuarios. El aprovechamiento y explotación de estas nuevas fuentes de opinión ofrece un gran atractivo para las empresas y los clientes. El análisis de sentimientos nos ayuda en esta tarea, ya que consiste en realizar una clasificación de los textos por su opinión, y se podría distinguir entre comentarios positivos o negativos sin tener que leerlos.

Fue en el año 2002, cuando Pang *et al.* (2002) y Turney (2002) propusieron clasificar los documentos basados en la opinión. Turney (2002) aplicó una técnica de aprendizaje no supervizada con un enfoque linguístico, recopilando estadísticas en un motor de búsquedas. Por otro lado, Pang *et al.* (2002) utilizaron métodos de aprendizaje de máquina supervisados.

A partir de estos dos enfoques se han desarrollado distintos trabajos tratando de clasificar las opiniones.

Algunos autores como Li *et al.* (2012); Taboada *et al.* (2011); Yu *et al.* (2008); Liu y Lai

(2010) han optado por el enfoque ligüístico intentando mejorar los diccionarios semánticos, donde las precisiones de clasificación en sus documentos están entre 79 % y 86 %.

Otros autores como Go *et al.* (2009); Koncz y Paralic (2011); Duric y Song (2011); Smith y Lee (2012) han optado por los métodos de aprendizaje de máquina tratando de mejorar los clasificadores, aquí las precisiones de clasificación en sus documentos están entre 83 % y 88 %.

1.2. Análisis de sentimientos

El problema del análisis de sentimientos consiste en asignar la clase positiva o negativa a un conjunto de documentos que expresan una opinión. Por ejemplo, si tengo una crítica hacia una película, con el análisis de sentimientos se le puede asignar el valor de positivo si la crítica hacia la película fue buena, y negativo si fue mala.

El análisis de sentimientos o minería de opiniones es el estudio de las opiniones de las personas, apreciaciones y emociones hacia entidades, eventos y sus atributos. La investigación en el campo comenzó con la clasificación de sentimientos y subjetividad, lo cual se trata como un problema de clasificación de texto. La clasificación de sentimientos determina si un documento dogmático (v.g. la revisión de un producto) o una oración expresa una opinión positiva o negativa (Liu, 2010).

El léxico utilizado por los usuarios de Internet, quienes suelen escribir las críticas, es muy extenso, y crear clasificadores con tantas características puede ser muy tardado e ineficiente. Además, a diferencia de los autores que trabajan con procesado de lenguaje natural (quienes elijen características con carga afectiva), en el aprendizaje de máquina se toman todas las características o combinaciones de ellas para construir los clasificadores. Algunos autores (Pang *et al.*, 2002; Annett y Kondrak, 2008) sí manejan estas combinaciones para ver qué tipo de características son más útiles para mejorar la precisión de los clasificadores, mas no pueden elegir entre ellas cuáles son las que realmente se necesitan y cuáles sólo proporcionan ruido. De aquí tenemos que la selección de características es un paso importante para la construcción de mejores clasificadores en sentido de precisión, eficiencia y reducción de espacio de almacenamiento.

1.2.1. Selección de características

El problema de selección de características consiste en encontrar un subconjunto de características con el que se pueda construir un mejor clasificador en sentido de precisión y eficiencia.

En los métodos para análisis de sentimientos basados en aprendizaje de máquina se trabaja con un número elevado de características. Los métodos para seleccionar características juegan un papel muy importante en el análisis de sentimientos. El uso apropiado de estos métodos ayuda a entender la relevancia de ciertos atributos según su clase, además de que incrementa la precisión de clasificación (Koncz y Paralic, 2011).

Abbasi *et al.* (2008) son de los primeros autores en hacer una selección de características en el contexto de análisis de sentimientos proponiendo un algoritmo genético ponderado con entropías (EWGA) y obteniendo un 91.7 % de precisión clasificando críticas de grupos extremistas. O'Keefe y Koprinska (2009) proponen otros métodos menos costosos computacionalmente y obtienen 87.15 % de precisión utilizando diferencia proporcional categórica (PD). Koncz y Paralic (2011) proponen su propio método para la selección de características usando valores de frecuencias de términos en categorías particulares obteniendo 87.65 % de precisión, pero no obtienen mejores resultados que utilizando ganancia de información. En Duric y Song (2011) proponen un método basado en contenido y modelos de sintaxis y obtienen una precisión de 87.5 % en la clasificación.

1.2.2. Trabajos en español

Los trabajos en análisis de sentimientos en español son escasos. La gran mayoría de la investigación relacionada con el análisis de sentimientos se centra casi exclusivamente en textos escritos en inglés. Sin embargo, el idioma español tiene cada vez más presencia en Internet[1] por lo que la investigación en análisis de sentimientos debería centrarse también en este idioma.

Cruz *et al.* (2008) utilizan el concepto de orientación semántica para clasificar críticas de cine en español. La arquitetura de su clasificador se basa en la de Turney (2002), donde

[1] http://www.internetworldstats.com/stats7.htm visto en agosto 2014

hace uso de búsquedas en la web para obtener relaciones entre sus términos y las palabras *excelente* y *malo*. Rosá *et al.* (2010); Rosá (2011); Montejo-Ráez *et al.* (2012) también utilizan un enfoque semántico para la clasificación de texto en español, mientras que Martínez *et al.* (2011) utilizan máquinas de soporte vectorial (SVM) y Naïve Bayes (NB), que son clasificadores de aprendizaje de máquina.

Banea *et al.* (2008) y Brooke *et al.* (2009) utilizan traductores de texto español a inglés para utilizar recursos disponibles en inglés. R-Moreno *et al.* (2013); Montejo-Ráez *et al.* (2013) se enfocan en obtener un corpus de *Twitter* en español para clasificar polaridad en masa.

1.2.3. Sistemas existentes

Entre algunos sistemas propuestos en la literatura, podemos encontrar un sistema llamado Opinion Observer por Liu *et al.* (2005), SentiWordNet por Esuli y Sebastiani (2006), SO-Cal por Taboada *et al.* (2011), TOES por Cruz Mata (2011), entre otros. La mayoría de estos trabajan con el enfoque lingüístico, expandiendo sus diccionarios de términos con orientación semántica.

Existe un sistema para el análisis de sentimientos en español, llamado Sentitext (Moreno *et al.*, 2010), el cual está basado en conocimiento. Similar a la orientación semántica, utilizan un sistema de valencias (-2,-1,1 y 2) para indicar la carga afectiva de la palabra (muy negativa, negativa, positiva y muy positiva).

1.3. Planteamiento del problema

El enfoque que se desea tomar en el problema del análisis de sentimientos es la selección de características para documentos en español. Utilizando un algoritmo bioinspirado, con la ayuda de un clasificador en la función de evaluación, se optimizará la precisión de clasificación teniendo como elemento de decisión el subconjunto de características seleccionado.

1.4. Objetivos

1.4.1. Objetivo general

Diseñar, implementar y analizar el desempeño de un algoritmo genético para la selección de características para el análisis de sentimientos de documentos en español. El algoritmo

genético utilizará como función de aptitud la precisión de clasificación de una SVM lineal.

1.4.2. Objetivos específicos

Se plantean los siguientes objetivos específicos:

- Establecer un corpus o corpora de entrenamiento.
- Seleccionar los parámetros del algoritmo genético que ayudarán a reducir el número de características para determinar la polaridad de un texto en español.
- Diseñar la estructura de los individuos que representarán los subconjuntos de características.
- Diseñar los operadores evolutivos para este problema en particular.
- Diseñar la función de evaluación que mejore la precisión reduciendo el número de características.
- Evaluar la confiabilidad de los resultados con textos nuevos.
- Analizar el desempeño logrado con el consenso de varios clasificadores con respecto al logrado por cada clasificador por separado.

1.5. Metodología de solución

Para cumplir con el objetivo de este trabajo, se propone el uso de un algoritmo genético para seleccionar un conjunto de características con el cual se construya un clasificador que resuelva el problema de análisis de sentimientos en español. El clasificador, que se utiliza en la función de evaluación del algoritmo genético, es un SVM (Support Vector Machine) con kernel lineal, ya que ha demostrado obtener buenos resultados en la clasificación de documentos (Pang *et al.*, 2002).

Ya que se ha elegido un corpus o corpora, se extraen todas las características de los documentos de opinión obteniendo así el conjunto total de características a utilizar. De aquí, se seleccionan subconjuntos de características aleatoriamente. Además se selecciona un

subconjunto usando la ganancia de información. Estos conjuntos de características se utilizan para clasificar los documentos.

Se utiliza un enfoque de envoltura el cual consiste en combinar un método de búsqueda y un clasificador. La búsqueda se realiza con la ayuda de los operadores genéticos. El clasificador nos da la precisón la cual nos indica qué tan bueno es un subconjunto de características. Dependiendo de la precisión que brindan los conjuntos de características, estos se van modificando con los operadores genéticos hasta tener un conjunto que brinde una mejor precisión en la clasificación de los documentos. De este modo nos quedamos con los mejores subconjuntos de características encontrados.

Una vez que se tiene el conjunto de características y el clasificador generado por este conjunto, se prueba con documentos desconocidos (documentos que no se tomaron en cuenta para la extracción de características) para ver la precisión que se genera a partir de estos.

1.6. Organización de la tesis

En el Capítulo 2 se introducen los conceptos básicos para el entendimiento de este trabajo, así como ejemplos sencillos para facilitar la comprensión de estos.

En el Capítulo 3 se describe detalladamente el algoritmo propuesto, así como los módulos principales que son la creación de los individuos (la representación de los conjuntos de características), las operaciones, la selección y la función de evaluación.

En el Capítulo 4 se presentan los experimentos elaborados a lo largo de este trabajo, así como los parámetros que se tomaron para el algoritmo. También se presentan los resultados obtenidos a partir de estos experimentos.

En el Capítulo 5 se presenta un sumario del trabajo así como las conclusiones y algunas ideas de trabajo a futuro.

En el Apéndice A se encuentra una lista de las características obtenidas por el mejor individuo en una corrida del algoritmo genético, así como un ejemplo de un documento verdadero positivo, uno verdadero negativo, uno falso positivo y uno falso negativo dados por el clasificador generado por este individuo.

Capítulo 2. Marco teórico

En este capítulo se presentan los conceptos básicos que se necesitan para entender tanto el problema que se aborda como el procedimiento que se llevó a cabo para resolverlo.

2.1. Análisis de sentimientos

Sea $D = \{d_1, d_2, ..., d_n\}$ un conjunto de documentos de opinión y $C = \{negativa, positiva\}$ el conjunto de las clases a las que están asociadas los documentos de D. La tarea del análisis de sentimientos, también conocida como minería de opiniones, consiste en asignar a cada uno de los documentos de D una clase de C, según el carácter negativo o positivo de las opiniones vertidas en los mismos (Cruz *et al.*, 2008).

El análisis de sentimientos trata de identificar y analizar las opiniones y emociones (Abbasi *et al.*, 2008). En el análisis de sentimientos se trabaja con información subjetiva lo que supone una dificultad en comparación con la clasificación de textos según su tema.

2.1.1. Selección de características

El problema de selección de características en términos de aprendizaje inductivo supervisado se define como: dado un conjunto de características candidatas, seleccionar un subconjunto definido por uno de los tres siguientes enfoques (Molina *et al.*, 2002):

- El subconjunto de un tamaño específico que optimice una medida de evaluación.
- El subconjunto de menor tamaño que satisfaga cierta restricción de medida de evaluación.
- El subconjunto con el mejor compromiso entre su tamaño y el valor de su medida de evaluación (caso general).

Formalmente, el problema de selección de características se define como, sea X el conjunto original de características con cardinalidad $|X| = n$ y sea $J(X')$ una medida de evaluación que se desea optimizar (digamos maximizar) definida por $J : X' \to \mathbb{R}$. La selección de un subconjunto de características se da bajo las tres consideraciones (Molina *et al.*, 2002):

- Establecer $|X'| = m < n$. Encontrar $X' \subset X$, tal que $J(X')$ es máximo.
- Establecer un valor J_0, esto es, el mínimo J que será tolerado. Encontrar el $X' \subseteq X$ con menor $|X'|$, tal que $J(X') \geq J_0$.
- Encontrar un compromiso entre minimizar $|X'|$ y maximizar $J(X')$ (caso general).

Nótese que, con estas definiciones, un subconjunto óptimo de características no es necesariamente único.

2.1.1.1. Selección de características en el contexto de análisis de sentimientos

Nuestro problema se enfoca en encontrar un subconjunto de características, del conjunto total de características que se extraen de los mismos documentos de entrenamiento y prueba, para la construcción de un buen clasificador de documentos según su sentimiento.

La medida de evaluación que indica qué tan bueno es un subconjunto de características elegido es la precisión que brinda el clasificador con los documentos de prueba.

2.1.2. Documento de opinión

Se define como un documento de opinión a cualquier unidad de texto en la que se recoja un análisis crítico sobre algún objeto, pudiendo ser ese objeto un producto comercial, una película, una ley o cualquier otra entidad susceptible de ser sometida a crítica (Cruz *et al.*, 2008).

Cada documento de opinión tiene su propio estilo, lo que crea desorden y alta diversidad ya que las opiniones se expresan en un lenguaje informal, el cual puede depender del tipo de comunidad que lo emplea (Westerski, 2007).

2.1.3. Corpus

El corpus es la recopilación de los documentos de opinión, los cuales serán utilizados para el entrenamiento del clasificador. Estos documentos se etiquetan con la clase a la que pertenecen. La utilización de un corpus de críticas de cine es conveniente debido a las grandes colecciones en línea y porque suelen venir acompañadas de una puntuación en estrellas que

resume el sentimiento del crítico. A la vez, es interesante trabajar con un corpus de críticas de cine porque resulta difícil (Turney, 2002), ya que existen dos aspectos en una película, los elementos de la película (i.e. actores, eventos), y el estilo y arte de la película.

Dentro de los trabajos que hay en español, Banea *et al.* (2008) utilizan el corpus MPQA por Wiebe y Cardie (2005), el cual habla sobre noticias en inglés, pero lo traducen al español obteniendo una precisión de 68.5 % con un enfoque de aprendizaje de máquina. Brooke *et al.* (2009) utilizan dos corpora en español obtenidos de *Ciao.es* y *Dooyoo.es*, estos corpora contienen críticas sobre cualquier objeto, y obtienen precisión de 74.5 % y 73.5 %, respectivamente bajo el enfoque lingüístico, también realizan pruebas con enfoque de aprendizaje de máquina obteniendo precisiones de 72.25 % y 69.75 %. Vilares *et al.* (2013) utiliza un corpus compuesto por tuits en español (TASS 2012) con el cual obtienen una precisión de 67.6 % con un enfoque lingüístico. Montejo-Ráez *et al.* (2013) también trabajan con tuits con un corpus al cual llaman *MeSiento* y obtienen precisión de 68.25 % con un enfoque lingüístico. Cruz *et al.* (2008) proponen un corpus que obtienen de Muchocine.com, que es sobre críticas de cine, y obtienen una precisión de 77.5 % con un enfoque lingüístico, otros autores utilizan este mismo corpus, como Martínez *et al.* (2011) y Martín-Valdivia *et al.* (2013) quienes obtienen precisión de 86.4 % y 88.57 %, respectivamente con este mismo corpus con un enfoque de aprendizaje de máquina.

2.1.4. Características

Las características que se extraen de los documentos son las que se utilizan para entrenar un clasificador. Estas pueden ser n-gramas, lemas, etiquetado morfosintáctico (POS-tag), POS-bigramas, signos de puntuación, entre otras. Además el valor que cada característica puede tomar puede indicar su presencia (con 0 y 1), frecuencia (con 0,1,2,...), o valor semántico (con un número real). A continuación se describen los más relevantes de esta lista.

2.1.4.1. N-gramas

Los n-gramas que se utilizan comúnmente son los "unigramas" y "bigramas". Los "unigramas" consisten en una palabra o un conjunto de palabras con un solo significado (v.g. por_arte_de_magia); los "bigramas", en pares consecutivos de "unigramas"; "trigramas", en tres "unigramas". Se recomienda utilizar hasta "trigramas", pues de cuatro en adelante se

presenta el fenómeno conocido como sobreajuste (Abbasi *et al.*, 2008). Tomemos los siguientes documentos cortos como ejemplo para extraer sus características.

d_1**:** La mejor película que he visto.

d_2**:** Es una película vacía.

En la Tabla 1 se observa, en las primeras tres columnas, los "unigramas", "bigramas" y "trigramas" correspondientes a estos documentos.

2.1.4.2. Lemas

El lema de una palabra es la forma base de una palabra. Es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. Es decir, el lema de una palabra, es la palabra que nos encontraríamos como entrada en un diccionario tradicional: singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos. En la Tabla 1 podemos ver los lemas correspondientes a nuestros documentos de ejemplo de la Sección 2.1.4.1 en la cuarta columna.

2.1.4.3. Etiquetado morfosintáctico

El etiquetado morfosintáctico (part-of-speech tagging (POS-tag)) es el proceso de asignar a cada una de las palabras de un texto los rasgos morfológicos asociados a ella (categoría gramatical, género, número, persona, etc.) en función del contexto en que aparecen. Se utilizan las etiquetas EAGLES (Gibbon *et al.*, 1997) para representar esta información, estas presentan ciertos códigos simbolizando cada rasgo morfológico, los cuales indican el género y número de la palabra, así como si esta es adjetivo, adverbio, pronombre, entre otros. En la Tabla 1, en la quinta columna se muestran los POS-tag correspondientes a los "unigramas" de los documentos de ejemplo de la Sección 2.1.4.1. Los signos de puntuación están incluidos en esta categoría, ya que cada signo siempre tendrá el mismo código (v.g. Fp representa '.', Fit representa '?' y Fat representa '!').

2.1.4.4. POS-bigramas

Los POS-bigramas se obtienen de las etiquetas morfosintácticas de dos palabras consecutivas, o bien de cada bigrama. Por lo que cada POS-bigrama consiste de dos etiquetas separadas por '_'. En la Tabla 1 se muestra el ejemplo de los POS-bigramas obtenidos de los documentos d_1 y d_2 del ejemplo de la Sección 2.1.4.1.

Tabla 1: Ejemplo de características extraídas de los documentos d_1 y d_2.

unigramas	**bigramas**	**trigramas**	**lemas**	**POS-tag**	**POS-bigramas**
la	la mejor	la mejor película	el	DA0FS0	DA0FS0_AQ0CS0
mejor	mejor película	mejor película que	mejor	AQ0CS0	AQ0CS0_NCFS000
película	película que	película que he	película	NCFS000	NCFS000_PR0CN000
que	que he	que he visto	que	PR0CN000	PR0CN000_VAIP1S0
he	he visto	es una película	haber	VAIP1S0	VAIP1S0_VMP00SM
visto	es una	una película vacía	ver	VMP00SM	VMP00SM_Fp
es	una película		ser	VSIP3S0	VSIP3S0_DI0FS0
una	película vacía		uno	DI0FS0	DI0FS0_NCFS000
vacía			vacío	AQ0FS0	NCFS000_AQ0FS0
				Fp	AQ0FS0_Fp

2.1.5. Métodos para la selección de características

El conjunto de características utilizado suele ser bastante grande y no todas las características son relevantes. Para obtener el mejor subconjunto de características hay que eliminar aquellas que no sean útiles o sean redundantes. Existen dos enfoques para realizar la selección de características (García Castellano, 2009): mediante filtrado, donde se seleccionan las variables basándose en medidas sobre los datos; y por envoltura, donde las variables de seleccionan utilizando el algoritmo de aprendizaje como si fuese caja negra. En el enfoque mediante filtrado se utilizan métodos basados en umbral y por envoltura, algoritmos bio-inspirados utlizando un clasificador en su función de evaluación.

A continuación se describen dos métodos basados en umbral: la ganancia de información y el estadístico χ^2; y un método basado en algoritmos bio-inspirados: el algoritmo genético.

2.1.5.1. Métodos para la selección de características basados en umbral

Estos métodos se basan en la evaluación de criterios sobre las características para con base en ellas seleccionar un subconjunto de las mismas para entrenar un clasificador.

Ganancia de Información. La Ganancia de Información (GI) es una medida basada en la entropía del sistema, es decir, en el grado de desorden del sistema. Esta medida nos indica cuánto se reduce la entropía de todo el sistema si conocemos el valor de un atributo determinado y nos sirve para extraer ciertas características de un conjunto dado a partir de fijar un cierto umbral para este criterio. La expresión para calcularla es la siguiente (Yang y Pedersen, 1997)

$$GI(C|E) = H(C) - H(C|E), \tag{1}$$

donde

$GI(C|E)$: es la ganancia de información de la etiqueta o característica E para la clase C,

$H(C)$: es la entropía del sistema, y

$H(C|E)$: es la entropía condicional del sistema conocido el valor de la etiqueta E.

La entropía del sistema nos indica el grado de desorden y viene dada por:

$$H(C) = -\sum_{i=1}^{|C|} p(c_i) \log_2(p(c_i)), \tag{2}$$

donde $p(c_i)$ es la probabilidad de la clase c_i sin conocimiento a priori. La entropía condicional se calcula de la siguiente forma:

$$H(C|E) = \sum_{j=1}^{|E|} p(e_j)(-\sum_{i=1}^{|C|} p(c_i|e_j) \log_2(p(c_i|e_j))), \tag{3}$$

donde $p(e_i)$ es la probabilidad de que la característica E tome el valor e_i, y $p(c_i|e_j)$ es la probabilidad condicional de tener la clase c_i dado el valor e_j.

Como ejemplo, en la Tabla 2 se muestra un conjunto de entrenamiento donde vemos tres características que definen si un objeto pertenece a la clase positiva $(+)$ o a la clase negativa $(-)$ denotando con uno que el objeto contiene esa característica y con cero que carece de esta. Para obtener la entropía del sistema vemos que $p(+) = \frac{1}{2}$ y $p(-) = \frac{1}{2}$, por lo que $H(C) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$. Ahora falta obtener la entropía condicional de las características rojo,

Tabla 2: Ejemplo de conjunto de entrenamiento.

Rojo	Cuadrado	Grande	Clase
1	1	1	+
0	1	1	+
1	0	0	−
0	1	0	−
1	0	1	+
0	1	1	−

cuadrado y grande. Si tomamos la característica grande, por ejemplo, tenemos $p_G(1) = \frac{2}{3}$ y $p_G(0) = \frac{1}{3}$, y las probabilidades condicionales $p_G(+|1) = \frac{3}{4}$, $p_G(-|1) = \frac{1}{4}$, $p_G(+|0) = 0$ y $p_G(-|0) = 1$, por lo que $H(C|grande) = \frac{2}{3}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) + \frac{1}{3}(-0\log_2 0 - 1\log_2 1) = 0.5408$. De aquí que $GI(C|grande) = 1 - 0.5408 = 0.4591$. Realizando lo mismo para las características rojo y cuadrado tenemos que $GI(C|rojo) = 0.0817$ y $GI(C|cuadrado) = 0$. Así que la característica grande es la que me da más información sobre el objeto.

Yang y Pedersen (1997) encuentran que trabajar con GI es efectivo en la reducción de dimensionalidad para la categorización de texto. Generalmente se utiliza como método base para evaluar un algoritmo de selección de características.

Estadístico χ^2. El estadístico χ^2 mide la falta de independencia entre un término t y una categoría c. Utilizando la tabla de contingencia de dos vías de t y c, donde A es el número de veces que t y c ocurren, B es el número de veces que t ocurre sin c, C es el número de veces que ocurre c sin t, D es el número de veces que ni c ni t ocurren y N es el número total de documentos, se define esta medida como (Yang y Pedersen, 1997):

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}. \tag{4}$$

Una vez que se computa el estadístico χ^2 de cada categoría con cada término, se obtienen los valores de $\chi^2_{avg}(t)$ y/o $\chi^2_{max}(t)$.

Utilizando el ejemplo del conjunto de entrenamiento en la Tabla 2 obtenemos las medidas de falta de independencia para las características rojo, cuadrado y grande. Ya que hay el mismo número de objetos pertenecientes a la clase positiva y negativa, solo es necesario

obtener el estadístico χ^2 para una clase. Por lo que $\chi^2(rojo) = \frac{2}{3}$, $\chi^2(cuadrado) = 0$ y $\chi^2(grande) = 3$, donde la característica con mayor valor de estadístico χ^2 es la característica grande.

Se conoce que la estadística χ^2 no es fiable para términos de baja frecuencia (Dunning, 1993).

2.1.5.2. Algoritmos bioinspirados

Los algoritmos bioinspirados son conocidos por resolver problemas de optimización tratando de emular la evolución natural de las especies. Su naturaleza es aleatoria y dependen de distintos parámetros.

Un algoritmo bioinspirado consta de una población (un conjunto de datos candidatos a ser solución), un mecanismo de evolución (procedimientos para modificar la población), el mecanismo de calificación (procedimientos para asignar una calificación a cada individuo de la población) y condiciones de parada.

Algoritmos genéticos. Los principios básicos de los Algoritmos Genéticos (AG) fueron establecidos por Holland (1975). Están basados en el proceso genético de los organismos vivos. Los AG son capaces de ir creando soluciones para problemas del mundo real, usando los principios de la selección natural y supervivencia de los más fuertes. Un tratado más profundo del tema se puede encontrar en Holland (1975); Goldberg (1989); Eiben y Smith (2003); Koza (1992).

2.2. Clasificación

Se conoce como clasificación a la tarea de asociar los valores que toma un conjunto de variables (también denominadas características) con un conjunto discreto de valores, denominados clases. La idea es construir un modelo, denominado clasificador, a partir de un conjunto de datos, y luego utilizarlo para asignar clases a nuevos datos que no se usan en la construcción del modelo.

2.2.1. Clasificación de documentos

La clasificación de documentos es una instancia importante de la clasificación con desafíos y requerimientos únicos, lo cual consiste en clasificar un documento en una o más C posibles clases. Un gran reto con la clasificación de documentos es la representación de los documentos. La forma más fácil de hacer esto es con la representación llamada bolsa de palabras, la cual consiste en un vector con el conteo de palabras que aparecen en él (Boulis y Ostendorf, 2005).

Otra dificultad de la clasificación de documentos es la alta dimensionalidad del espacio de características. Es altamente deseable reducir el espacio de características sin sacrificar la precisión de clasificación y de forma automática (Yang y Pedersen, 1997). Para esto existen los métodos de selección de características como la ganancia de información (Lewis y Ringuette, 1994), información mutua y el estadístico χ^2 (Wiener *et al.*, 1995), análisis de componentes principales (Yang, 1995), técnicas de agrupamiento (Yang y Wilbur, 1996), entre otros.

2.2.2. Clasificación de sentimientos

La clasificación de sentimientos no es exclusiva para documentos ya que se pueden distinguir emociones en el discurso hablado (El Ayadi *et al.*, 2011), en las imágenes (Qiao *et al.*, 2011), en los rostros de las personas (Humphries y McDonald, 2011), en la música (Lin *et al.*, 2011) y hasta en las finanzas y la economía (Ahmad *et al.*, 2011).

2.2.3. Clasificación de documentos para el análisis de sentimientos

La clasificación de documentos para el análisis de sentimientos consiste en determinar la clase a la que corresponde cada documento de opinión, ya sea asignar la clase negativa o positiva. La clasificación de un documento de opinión puede verse como una clasificación binaria, ya que las clases positiva y negativa se pueden representar por uno y cero, respectivamente. En otros casos se incluyen otro tipo de sentimientos, muy positivo, positivo, neutral, negativo, muy negativo, o sin opinión (Vilares *et al.*, 2013), o sentimientos de enojo, vergüenza, empatía, miedo, orgullo, alivio, o tristeza (Xu *et al.*, 2012), que son multiclase.

Existen dos enfoques populares dentro del análisis de sentimientos para hacer la clasificación. Uno es el enfoque lingüístico y el otro aprendizaje de máquina (Brooke *et al.*, 2009).

Con el enfoque lingüístico se considera la orientación semántica de las palabras, utilizando expertos o sistemas propuestos para realizar una ponderación de las palabras.

2.2.3.1. Aprendizaje de máquina

Los clasificadores basados en aprendizaje de máquina aprenden un conjunto de reglas de los datos de entrenamiento. Los datos de entrenamiento suelen ser una sección del corpus previamente etiquetados. Esto indica que la clasificación debe ser supervisada y una vez que el clasificador ha sido entrenado se podrá probar con otros datos no vistos (Waila *et al.*, 2012).

Para conocer más sobre los temas de patrones y aprendizaje de máquina se pueden consultar los libros de Bishop (2006); Duda *et al.* (2001); Mitchell (1997).

Los algoritmos de aprendizaje de máquina que más se utilizan dentro del análisis de sentimientos son SVM (Support Vector Machine) y NB (Naïve Bayes). Martínez *et al.* (2011) hace una comparativa de estos dos métodos de clasificación con un corpus en español y obtiene mejores resultados con SVM.

2.2.3.2. Orientación semántica

La orientación semántica de una característica se define como un valor real que siendo positivo indica que este tiene implicaciones subjetivas positivas, y siendo negativo indica lo contrario (Cruz *et al.*, 2008). Además distintos valores absolutos de la medida informan distintos grados de intensidad en dichas implicaciones.

Para hacer una clasificación con orientación semántica, basta con realizar una suma sobre todos los valores de orientación semántica de las características contenidas en un documento de opinión, si la suma es mayor o igual a un umbral, el documento es positivo, y si es menor, negativo. El umbral suele ser cero (Turney, 2002).

2.2.4. Clasificación dentro de la metodología del trabajo

Para este trabajo se utiliza el enfoque de aprendizaje de máquina utilizando un clasificador SVM, que es el clasificador que se encuentra en el estado del arte y además ha brindado buenos

resultados para clasificar documentos de opinión tanto en inglés como en español (Smith y Lee, 2012; Martínez *et al.*, 2011).

Cuando se ha seleccionado un conjunto de características, este se pasa al clasificador que se utiliza como caja negra y obtenemos una precisión que se utiliza para comparar con la precisión obtenida por otros conjuntos de características y así poder seleccionar el mejor conjunto de características en forma iterativa a través de las generaciones del algoritmo evolutivo.

2.2.5. Trabajo previo relevante

En el uso de los métodos de aprendizaje de máquina surge el problema de selección de características. Pang *et al.* (2002) hacen una comparativa utilizando unigramas, unigramas+bigramas, bigramas, unigramas+POS, adjetivos, y otros; obteniendo su mejor precisión de clasificación de 82.9 % utilizando unigramas y clasificando con SVM, esto para críticas de cine en inglés. Sin embargo, no se hace una selección específica de estos unigramas.

Entre los autores que abordan el problema de análisis de sentimientos utilizando métodos de aprendizaje de máquina, están Duric y Song (2011); Koncz y Paralic (2011); O'Keefe y Koprinska (2009); Smith y Lee (2012); Xu *et al.* (2012) que hacen su selección de características utilizando la Ganancia de Información, proponen sus propios métodos (dando ponderaciones a las características en función de la frecuencia con que aparecen) o sólo eliminan las características que no aparezcan más de 5 veces. Por otro lado, Banea *et al.* (2008); Go *et al.* (2009); Mudinas *et al.* (2012); Pak y Paroubek (2010) utilizan n-gramas y/o combinaciones de estos como características. Las precisiones utilizando métodos de aprendizaje de máquina están entre el 73 % y 88 % con corpora misceláneos desde tuits hasta noticias.

Una aplicación diferente al de las críticas de cine, como lo hacen la mayoría de los autores, Smith y Lee (2012) utilizan un dominio clínico de críticas de pacientes hacia el servicio y Xu *et al.* (2012) detectan *bullying* en redes sociales, ambos trabajan con SVM ya que ha mostrado los mejores resultados (Annett y Kondrak, 2008).

En la Tabla 3 se aprecia un condensado de los trabajos más relevantes dentro del análisis de sentimientos que utilizan métodos de aprendizaje de máquina.

Tabla 3: Trabajos relevantes dentro del análisis de sentimientos utilizando métodos de aprendizaje de máquina.

Autor	**Corpus**	**Idioma**	**Clasificador**	**Selección de características**	**Mejor precisión**
Pang *et al.* (2002)	Críticas de cine	Inglés	Naïve Bayes y SVM	No	82.9 %
Abbasi *et al.* (2008)	Críticas de cine	Inglés	SVM	EWGA	91.7 %
Banea *et al.* (2008)	MPQA corpus (noticias traducidas con Google translator)	Rumano y español	Naïve Bayes y SVM	No	71.83 %
Go *et al.* (2009)	Twitter	Inglés	Naïve Bayes, Entropía máxima y SVM	No	82.2 %
O'Keefe y Koprinska (2009)	Críticas de cine	Inglés	Naïve Bayes y SVM	PD, SWNSS y SWNPD	87.15 %
Duric y Song (2011)	Críticas de cine	Inglés	Máxima entropía	HMM-LDA	87.5 %
Koncz y Paralic (2011)	Críticas de cine	Inglés	SVM	Método propuesto con frecuencias de documento	87.64 %
Martínez *et al.* (2011)	Muchocine	Español	Naïve Bayes y SVM	*Stopper* y *Stemmer*	86.84 %
Mudinas *et al.* (2012)	Críticas de software y de cine	Inglés	SVM	No	89.64 %
Smith y Lee (2012)	Textos clínicos	Inglés	Naïve Bayes y SVM	Únicamente las que aparezcan más de 5 veces	83.53 %
Martín-Valdivia *et al.* (2013)	Muchocine	Español	SVM	TF-IDF, TF, TO y BTO	87.69 %

2.2.6. EWGA

El algoritmo genético con ponderación de entropías (EWGA por sus siglas en inglés *entropy weighted genetic algorithm*) incorpora la heurística de ganancia de información en el algoritmo genético para mejorar el rendimiento de selección de características (Abbasi *et al.*, 2008). Se utiliza la ganancia de información para ponderar varios atributos de sentimiento, los pesos para las ponderaciones se incorporan en la población inicial del algoritmo así como en los operadores de cruzamiento y mutación.

Capítulo 3. Algoritmo propuesto

En este capítulo se describe la metodología propuesta, la cual consiste en utilizar un enfoque de envoltura para seleccionar las características que maximicen la precisión de clasificación utilizando un algoritmo genético.

3.1. Enfoque de envoltura

En el enfoque de envoltura, mostrado en la Figura 1, la selección del subconjunto de características se realiza utilizando un algoritmo genético donde la aptitud de los individuos está dada por la precisión de un clasificador como caja negra (i.e., no se necesita conocimiento del algoritmo, solo la interfaz). El algoritmo de selección de características conduce a la búsqueda de un buen subconjunto utilizando el clasificador como función de evaluación (Kohavi y John, 1997). La búsqueda requiere de un espacio de soluciones, una población de soluciones inicial, una condición de terminación y una máquina de búsqueda. La búsqueda es un proceso iterativo donde en cada iteración va cambiando el espacio de soluciones para llegar a la solución óptima.

3.2. Obtención de características

Para trabajar con el algoritmo genético, las características de los documentos se representan mediante un vector, el cual llamaremos vector de características. Por cada documento se genera un nuevo vector, con el cual se representa dicho documento, y este indica con un uno si la característica se encuentra presente y con un cero si la característica no está según la posición en el vector de características. De este modo, se forma una matriz de $m \times n$ elementos donde m es el número de documentos y n el número total de características. Esta matriz está compuesta de unos y ceros, y la llamaremos matriz de características, la cual varía según los documentos que se utilicen, ya que las características dependen de estos documentos.

Se tienen matrices de características para distintos conjuntos de documentos de opinión de en archivos de texto. s

Si tomamos los documentos de ejemplo vistos en la Subsección 2.1.4.1, y ordenamos las características de la Tabla 1, primero todos los unigramas, después bigramas, trigramas,

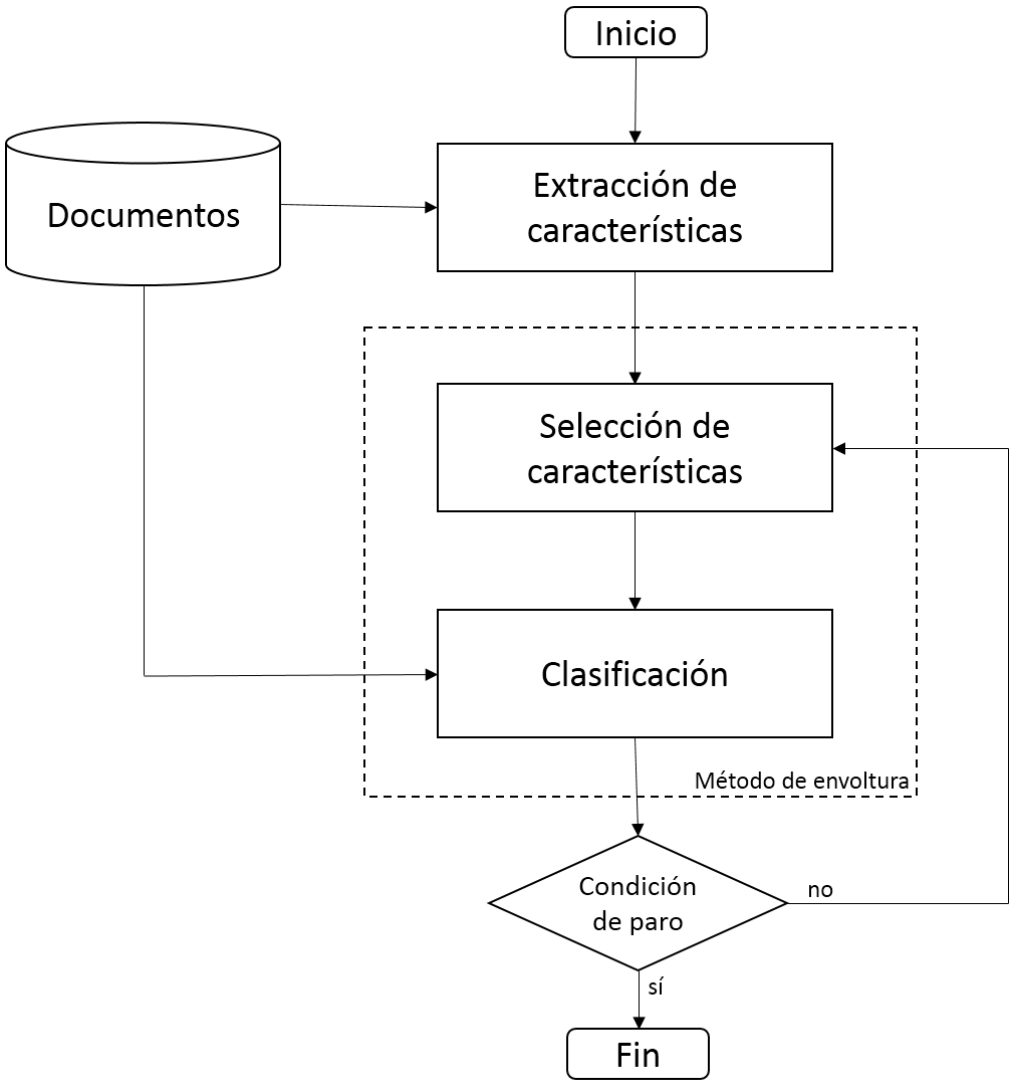
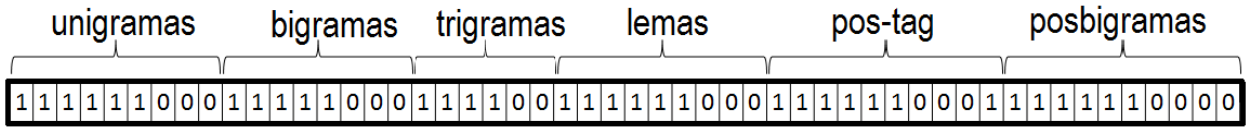
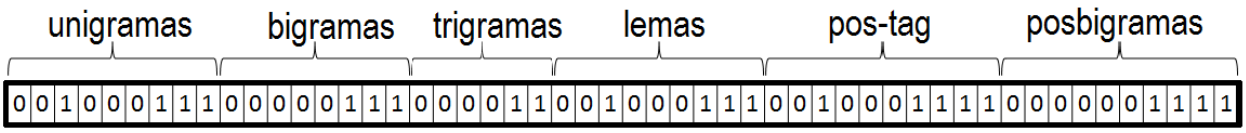


Figura 1: Diagrama de enfoque de envoltura.

lemas, pos-tag y posbigramas, tendremos que el vector de características para d_1 es el que se presenta en la Figura 2(a) y para d_2 es el que se presenta en la Figura 2(b).



(a) d_1



(b) d_2

Figura 2: Vectores de características para los documentos de ejemplo de la Sección 2.1.4.1.

De estos vectores de características queremos obtener los índices de las características que nos darán un mejor clasificador. Con esto esperamos que se reduzca el número de características para disminuir el espacio de almacenamiento y mejorar la eficiencia del clasificador.

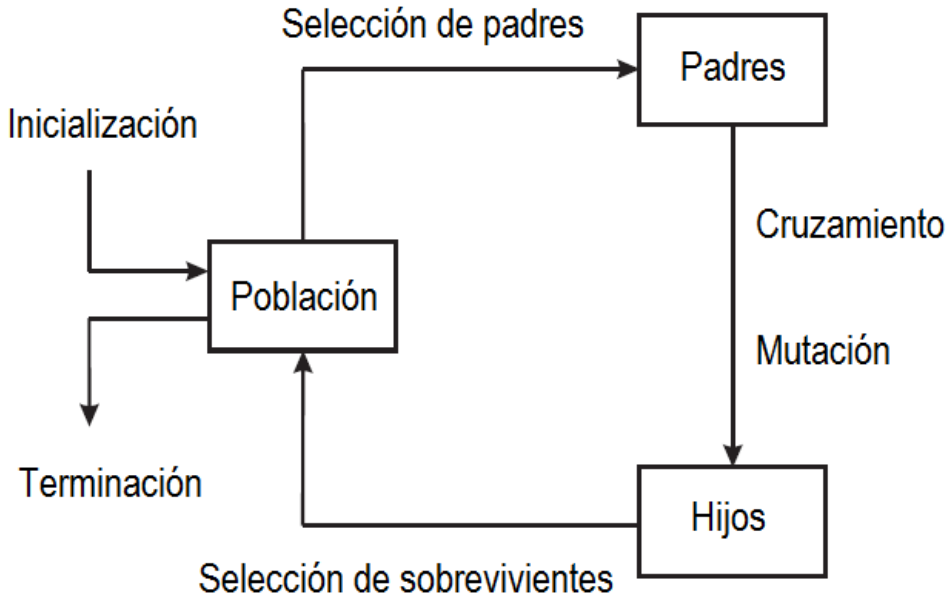


Figura 3: Diagrama del AG.

3.3. Algoritmo genético

El algoritmo genético (AG) propuesto por Holland (1975) genera soluciones a un problema buscando de forma estocástica la mejor solución. Se inicializa con una población, donde se

hace una selección de padres para que a través de operaciones de cruzamiento y mutación se genere una descendencia, después se elijen sobrevivientes para ser la población de la siguiente generación. En la Figura 3 se observa un diagrama esquemático del AG.

En el Algoritmo 1 se presenta el pseudocódigo del algoritmo genético que se utilizó para resolver el problema de selección de características en el análisis de sentimientos (AGpAS), así como los módulos principales; la generación de la población inicial de individuos se presenta en el Algoritmo 2, la selección de padres en el Algoritmo 3 y las operaciones genéticas en el Algoritmo 4.

Algoritmo 1 AGpAS

Entrada: generaciones, matriz_de_caracteristicas
Salida: poblacion

```
1: poblacion ← Genera_poblacion_inicial(matriz_de_caracteristicas, tamaño_poblacion)
2: Ordena(poblacion)
3: para i : 0 → generaciones hacer
4:     padres ← Selecciona_padres(poblacion)
5:     hijos ← Operadores_geneticos(padres)
6:     nueva_poblacion ← población ∪ hijos
7:     Ordena(nueva_poblacion)
8:     población ← Selecciona_sobrevivientes(nueva_poblacion)
9: fin para
```

Algoritmo 2 Genera_poblacion_inicial

Entrada: matriz_de_caracteristicas, tamaño_poblacion
Salida: poblacion

```
1: población ← ∅
2: ind ← Obten_individuo_GI(matriz_de_caracteristicas)
3: poblacion ← poblacion ∪ ind
4: mientras longitud(poblacion) < tamaño_poblacion hacer
5:     ind ← NumerosAleatorios()
6:     si ind ∉ poblacion entonces
7:         poblacion ← poblacion ∪ ind
8:     fin si
9: fin mientras
10: Evaluar(poblacion)
```

3.3.1. Representación del individuo y generación de la población inicial

Los individuos se representan por un vector de números enteros desde 0 hasta $n-1$, donde cada entero representa la posición dentro del vector de características, siendo 0 la primer característica y $n-1$ la última.

Algoritmo 3 Selecciona

Entrada: poblacion
Salida: padres
1: padres $\leftarrow \emptyset$
2: **para** i : 0 $\rightarrow$ tamaño_poblacion/4 **hacer**
3: elegidos $\leftarrow$ Elige5Individuos(poblacion)
4: Ordena(elegidos)
5: padre1 $\leftarrow$ Ruleta(elegidos)
6: padre2 $\leftarrow$ Ruleta(elegidos)
7: padres $\leftarrow$ padres $\cup$ (padre1,padre2)
8: **fin para**

Algoritmo 4 Operadores_geneticos

Entrada: padres
Salida: hijos
1: **para** (padre1,padre2) $\in$ padres **hacer**
2: **si** random$< P_c$ **entonces**
3: hijos $\leftarrow$ hijos $\cup$ Cruzamiento(padre1,padre2)
4: **fin si**
5: **fin para**
6: **para** hijo $\in$ hijos **hacer**
7: **si** random$< P_m$ **entonces**
8: hijo $\leftarrow$ Mutar(hijo)
9: **fin si**
10: **fin para**
11: Evaluar(hijos)

Los individuos se crean en la línea 1 del Algoritmo 1 en la función presentada en el Algoritmo 2 la cual se describe a continuación.

El primer individuo que se genera es el de ganancia de información. Una vez que se obtiene la ganancia de información de cada característica, se descartan las características que tienen menor ganancia de información que el umbral pre-especificado, y se genera un vector con las posiciones de las características dentro del vector de características que utilizaremos para posteriormente clasificar los documentos. A este individuo le llamaremos individuo con ganancia de información.

El resto de los individuos se generan aleatoriamente, tomando r números dentro del conjunto $0, 1, ..., n-1$ para cada individuo, con una distribución uniforme y bajo un muestreo sin reemplazo.

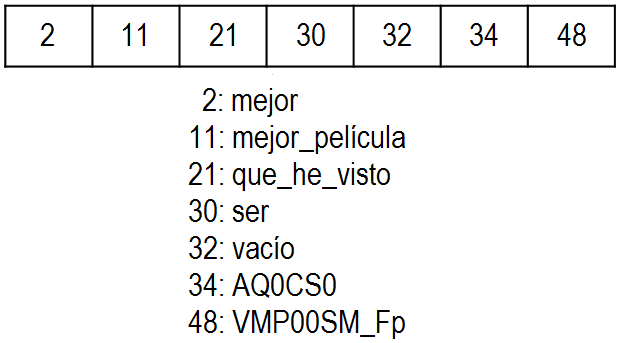


Figura 4: Ejemplo de representación de un individuo.

Como ejemplo, tenemos en la Figura 4 la representación de un individuo y las características que simbolizan los valores del individuo según el ejemplo de la Tabla 1. Por ejemplo, la característica 2 es el unigrama *mejor*, la característica 30 es el lema *ser*, la 34 es el pos-tag *AQ0CS0*, y así cada índice del individuo representa una característica.

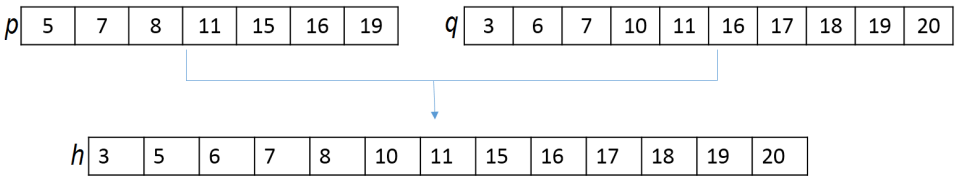
3.3.2. Operadores

Los operadores que se utilizan en el algoritmo genético son el cruzamiento y la mutación. Los operadores se utilizan después de que se seleccionaron los padres que tendrán descendencia en la línea 5 del Algoritmo 1, y las operaciones de cruzamiento y mutación se encuentran en el Algoritmo 4.

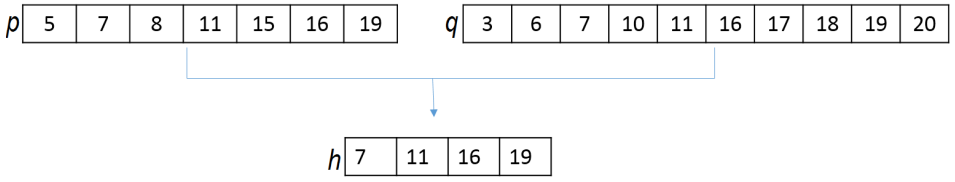
Usualmente en la literatura se encuentran trabajos de algoritmos genéticos para resolver problemas con permutaciones, en nuestro caso, por el hecho de trabajar con índices, se vuelve un problema de combinaciones, ya que deseamos seleccionar ciertos índices de nuestro conjunto total de características sin importar el orden. Los operadores genéticos encontrados en la literatura para permutaciones no sirven para nuestro problema particular por lo que se han propuesto nuevos operadores genéticos. Estos operadores genéticos fueron inspirados en las operaciones básicas de la teoría de conjuntos.

3.3.2.1. Cruzamiento

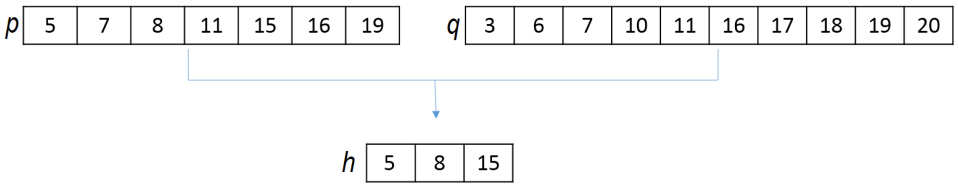
El cruzamiento consiste en obtener un nuevo individuo a partir de dos individuos llamados padres que se cruzan con una probabilidad de P_c. Se definen tres formas de obtener este nuevo individuo: unión, intersección y diferencia. En la unión, el nuevo individuo será un



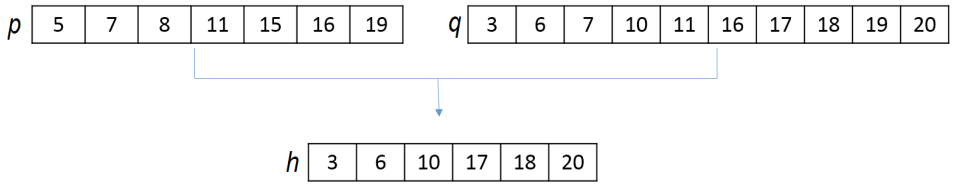
(a) Unión ($h = p \cup q$)



(b) Intersección ($h = p \cap q$)

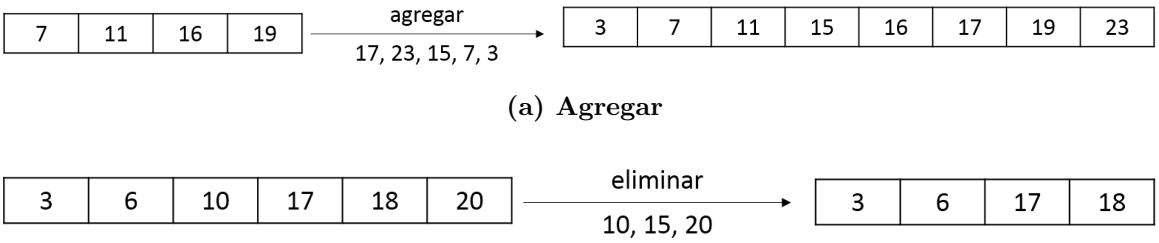


(c) Diferencia ($h = p - q$)



(d) Diferencia ($h = q - p$)

Figura 5: Operaciones de cruzamiento en el algoritmo genético.



(a) Agregar

(b) Eliminar

Figura 6: Operaciones de mutación en el algoritmo genético.

vector con todas las características de ambos padres; en la intersección, un vector con solo las características que están en ambos padres; y en la diferencia, un vector con las características de un padre que no estén contenidas en el otro. Estas operaciones se pueden apreciar en la Figura 5. Escrito de otro modo, si p y q son los padres, entonces el hijo h será $p \cup q$, $p \cap q$, $p - q$ o $q - p$. En la diferencia, el padre minuendo puede ser cualquiera de los dos padres, se elige aleatoriamente.

Cuando se han elegido dos padres para el cruzamiento, se elige aleatoriamente entre unión, intersección y diferencia para crear al nuevo individuo.

3.3.2.2. Mutación

Cuando generamos un individuo nuevo, hay una probabilidad de P_m de que ocurra una mutación. Cuando un individuo muta, se definen dos formas para llevar a cabo esta operación: agregar características o eliminar características. Aleatoriamente se elige únicamente una de las dos formas para relizar la mutación, además se elige un número al azar r entre 1 y n para agregar o eliminar esa cantidad de índices al individuo nuevo. Si se elige agregar características, se van agregando los r números enteros al vector que representa al nuevo individuo, y si se elige eliminar, se quitan los r números. Nótese que se trabaja con conjuntos, por lo que si se desea agregar un índice que ya está contenido en el individuo, no pasa nada, lo mismo si se va a eliminar un índice que no está contenido en el individuo.

En la Figura 6(a) podemos ver un ejemplo de un individuo antes y después de ser mutado con la opción de agregar $r = 5$ índices, y en la Figura 6(b) tenemos un ejemplo de un individuo con la opción de eliminar índices con $r = 3$. El orden de los índices en los ejemplos es de carácter ilustrativo ya que el orden de los índices elegidos en realidad no importa.

3.3.3. Función de evaluación

La función de evaluación o función objetivo es la que nos dice qué tan bueno es un individuo con respecto a otro. Para esto es necesario clasificar los documentos utilizando únicamente las características contenidas en el individuo. La clasificación se realiza tomando el $\alpha \times 100\,\%$ de los documentos para entrenar la SVM y se prueba con el $(1-\alpha) \times 100\,\%$ restante, esto último da la precisión usada como función objetivo. Además, se desea minimizar el número de características de modo que si dos individuos presentan una precisión similar, será mejor el individuo con menos características.

La evaluación se realiza una sola vez para cada individuo y se guarda su valor para compararlo con otros individuos. La evaluación se realiza en la línea 9 del Algoritmo 2, que es cuando se ha creado la población inicial, y en la línea 11 del Algoritmo 4, que es cuando se ha creado la descendencia.

Para tomar en cuenta la clasificación y el número de características se define un peso p de modo que la función de evaluación regresa para cada individuo:

$$f_e = (1-p) \times error\_de\_clasificación + p \times \bar{nc} \tag{5}$$

el número de características en la función se normaliza dividiéndolo entre el número total de características, i.e. $\bar{nc} = número\_de\_características / número\_total\_de\_características$.

En particular, para la clasificación se utiliza una SVM con kernel lineal que usualmente se utiliza para clasificar elementos de dos clases, y el cual utiliza un parámetro de penalización del término de error (C).

La evaluación de cada individuo se realiza una sola vez y se guarda el resultado con el individuo. De esta manera, en cada generación, sólo se evalúa la función objetivo en los nuevos individuos o hijos.

3.3.4. Selección de padres y sobrevivientes

La selección se realiza al principio de cada generación para elegir a los padres que tendrán descendencia, tal como se ve en la línea 4 del Algoritmo 1. En el Algoritmo 3 se realiza la

selección de padres que se describe a continuación.

La selección de padres para obtener a los nuevos individuos para la siguiente generación se elige por medio de una combinación de torneo y ruleta explicado a continuación (Myszkowski y Bicz, 2010).

Se eligen 5 individuos del total de la población. Se ordenan estos 5 individuos y se les asigna una probabilidad de ser elegidos. El padre con mejor valor de evaluación tiene 5/15 de probabilidad de ser elegido, el segundo mejor tiene 4/15, el tercero 3/15, el cuarto 2/15 y el último 1/15. Ya que se les ha asignado una probabilidad se toma un número aleatorio y obtenemos el primer padre. Se repite lo mismo para elegir un segundo padre.

Se ha elegido repetir la selección de padres $ni/4$ veces, donde ni es el número de individuos o tamaño de la población. Esto con el objetivo de disminuir el tiempo de ejecución al intentar reproducir toda la población, así, tenemos que solo se intentará reproducir la mitad de la población.

Una vez que se generaron los hijos, queda elegir a los sobrevivientes de la siguiente generación. Para esto, se ordena toda la población anterior más los hijos. Se eligen los primeros ni individuos por lo que se descartan los últimos individuos hasta que el tamaño de la población sea ni y ya tenemos la nueva población para la siguiente generación.

Capítulo 4. Experimentos y resultados

En este capítulo se describe de manera detallada los experimentos realizados así como los resultados obtenidos con el algoritmo propuesto.

4.1. Consideraciones preliminares

Se decidió trabajar con Python[1] por ser un lenguaje de programación de alto nivel, interpretado, multiplataforma, orientado a objetos y recursivo, con una sintaxis simple que además incluye bibliotecas científicas como Numpy, Scipy y MatPloLib. Python cuenta con Sklearn (Pedregosa *et al.*, 2011) que contiene herramientas simples y eficientes para minería y el análisis de datos, construida sobre NumPy, SciPy y MatPlotLib. Es de código abierto y con licencia BSD.

Sklearn cuenta con herramientas de clasificación, regresión, agrupamiento, entre otras. Se utilizó el SVM[2] de Sklearn que cuenta con parámetros como C (penalización del término de error), *kernel* (que puede ser lineal, polinomial, función de base radial (rbf), sigmoide o programado por el usuario), *degree* (grado del kernel polinomial), *gamma* (coeficiente utilizado por kernel rbf, polinomial y sigmoide), *coef0* (coeficiente utilizado en kernel polinomial y sigmoide), *probability*, *shrinking*, *tol*, *cache_size*, *class_weight*, *verbose*, *max_iter* y *random_state*.

Para comenzar con la clasificación de documentos de opinión se necesita un conjunto de documentos que expresen una opinión. Es necesario también realizar una sintonización de los parámetros involucrados en los algoritmos utilizados.

4.1.1. Obtención del corpus

Se utilizó un corpus en español de críticas de cine (Cruz *et al.*, 2008) para la realización de los experimentos. La colección está compuesta por 3878 documentos puntuados con una o hasta cinco estrellas, indicando con uno que fue una película mala y cinco una película buena; 1274 documentos fueron puntuados con una o dos estrellas, tomados como críticas negativas; 1350 fueron puntuados con cuatro o cinco estrellas, tomados como críticas positivas; el resto

[1] https://www.python.org/
[2] http://scikit-learn.org/stable/modules/svm.html

fue puntuado con 3 estrellas y fueron descartados por considerarse neutrales. Bajo estos criterios el total de documentos con críticas positivas o negativas es de 2624. Se decidió tomar las primeras 2000 críticas para trabajar con el algoritmo genético (1008 positivas y 992 negativas) y las 624 restantes se dejaron a un lado para hacer pruebas con los resultados obtenidos.

Cruz *et al.* (2008) utilizan la herramienta Freeling (Atserias *et al.*, 2006) para tokenizar y separar las oraciones, de aquí extraen los unigramas, lemas y etiquetado morfológico (POS-tag) de los documentos de opinión del corpus. A partir de estas características se formaron los bigramas, trigramas y POS-bigramas. Se tomaron todas estas características del conjunto de documentos de opinión que se utiliza para el AG. Se les aplicó un filtro a estas características para solo tomar en cuenta las características que aparezcan en al menos k documentos con $k > 1$.

4.1.2. Selección de características mediante el cálculo de ganancia de información

Se realizaron algunas pruebas para decidir el umbral de ganancia de información a utilizar, a partir de este umbral γ, se descartan las características que tengan una menor ganancia de información y se procede a clasificar los documentos con las características que sí tienen una ganancia de información mayor a γ. Dependiendo de la precisión de clasificación y del número de características que se eligieron, se desea elegir el umbral que dé los mejores resultados para cada conjunto de prueba.

Se consideró un conjunto de entrenamiento con 200, 300, 500, 800, 1000 y hasta 2000 documentos. Se varió el umbral a partir de 0 y con incrementos de 0.001 hasta que no se obtuvieran características. Esto para determinar el intervalo en el umbral de GI en que se obtiene la mejor precisión en el clasificador. Se consideró un clasificador SVM con kernel lineal, y la selección fue mediante una validación cruzada de tres pliegues.

En la Figura 7 se puede apreciar que la mayor precisión se presenta en el intervalo (0, 0.02). La figura ilustra la precisión que se obtuvo con las características que se seleccionaron según el umbral que se eligió. Cada curva representa el comportamiento de la precisión considerando un cierto número de documentos. Se reajusta la escala al intervalo (0, 0.2) para observar la

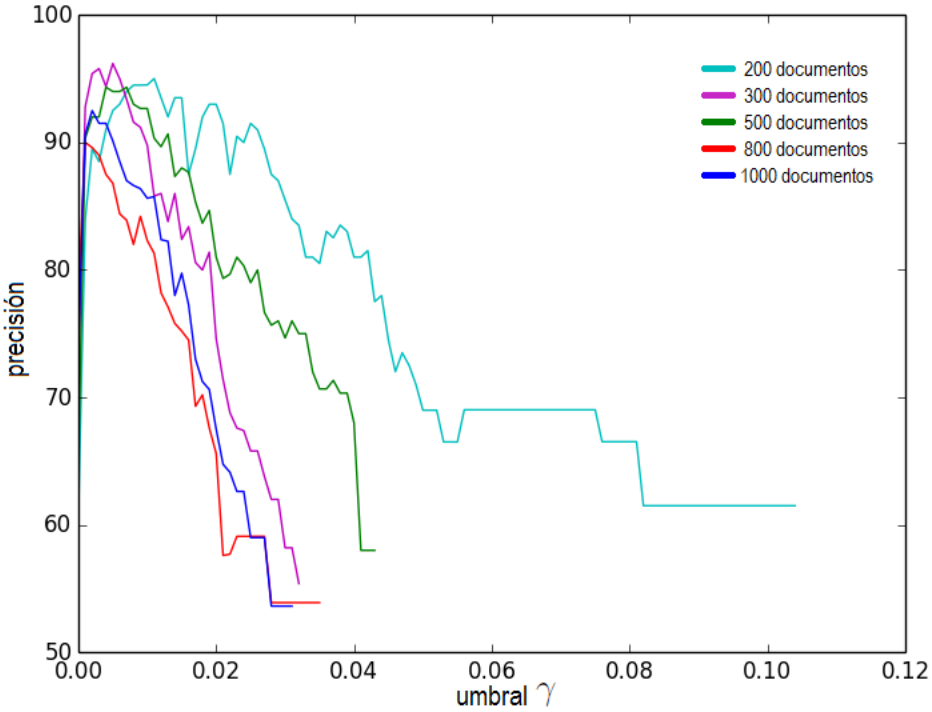


Figura 7: Umbral de ganancia de información vs. precisión de clasificación con incrementos de 0.001.

precisión en este intervalo. En la Figura 8 se observa el reajuste con incrementos de 0.00025 para observar mejor el comportamiento de la precisión.

Para ver los valores precisos en ciertos puntos se muestra la Tabla 4 donde se indica en negritas la mejor precisión comparado con los umbrales 0.00025, 0.0025 y 0.025. Estos valores se obtuvieron utilizando validación cruzada de cinco pliegues.

Tabla 4: Precisiones de clasificación obtenidas de distintos documentos en diferentes umbrales de ganancia de información.

Documentos / Umbral	200	300	500	800	1000	2000
0.00025	84.5	85	87	87.5	87	**89.75**
0.0025	**92**	**94.6**	**96.2**	**93.125**	**90.2**	84.9
0.025	86.5	80.3	64.6	58.875	58.9	-

Del conjunto de 2000 documentos, se tomó aleatoriamente un conjunto de 200 documentos para repetir la prueba 30 veces para tener significancia estadística. Se calculó la precisión con las características obtenidas variando el umbral desde 0 hasta 0.02 con incrementos de 0.00025. Se hizo lo mismo para 300, 500, 800 y 1000 documentos. Se tomó el promedio de

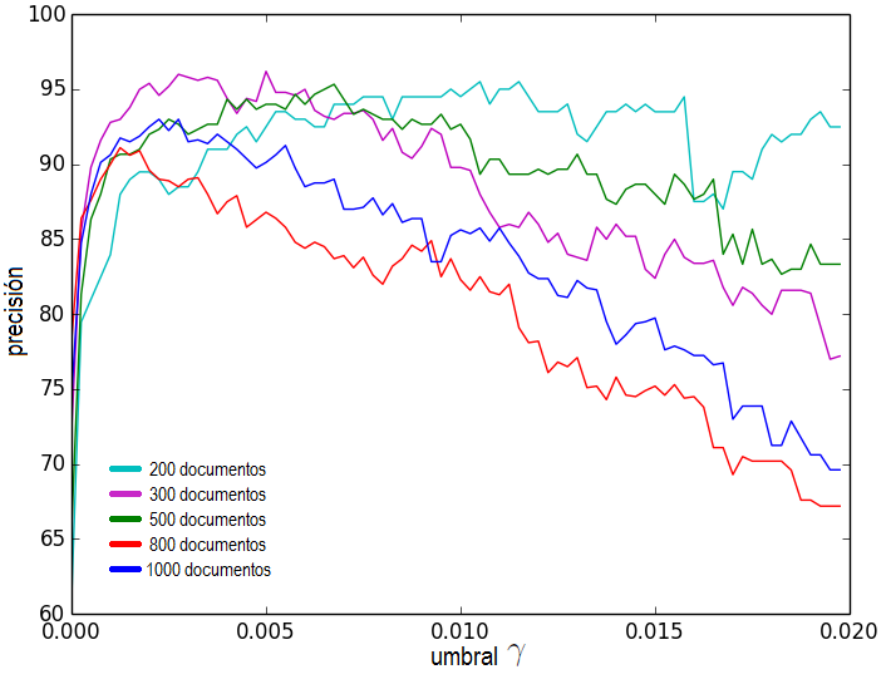


Figura 8: Umbral de ganancia de información vs. precisión de clasificación con incrementos de 0.00025.

las precisiones al evaluar los documentos en una SVM con validación cruzada de 3 pliegues en cada corrida para cada umbral, se disminuye el número de pliegues para reducir el costo computacional, y se grafica en la Figura 9. Además en la Tabla 5, podemos ver los valores de las mejores precisiones según el conjunto de documentos elegido, junto con el valor de umbral de GI donde se presenta esa precisión. El promedio de estos umbrales es 0.00385. Este valor está cercano a 0.0025, el cual es uno de los valores que se proponen para ser utilizado como el umbral de GI, por lo que se decide trabajar con este valor de 0.0025.

Tabla 5: Mejor umbral con distintos conjuntos de documentos.

Documentos	**Umbral**	**Precisión**	**No. características**
200	0.0075	95.78 %	3913
300	0.003	92.84 %	6249
500	0.00475	94.27 %	2509
800	0.00175	90.93 %	9398
1000	0.00225	91.65 %	4982
Promedio	0.00385	93.09 %	4330

Estas pruebas con el umbral de GI se realizan para partir de una buena solución en la inicialización de la población del AG. Una vez que se ha elegido el valor de 0.0025 de umbral

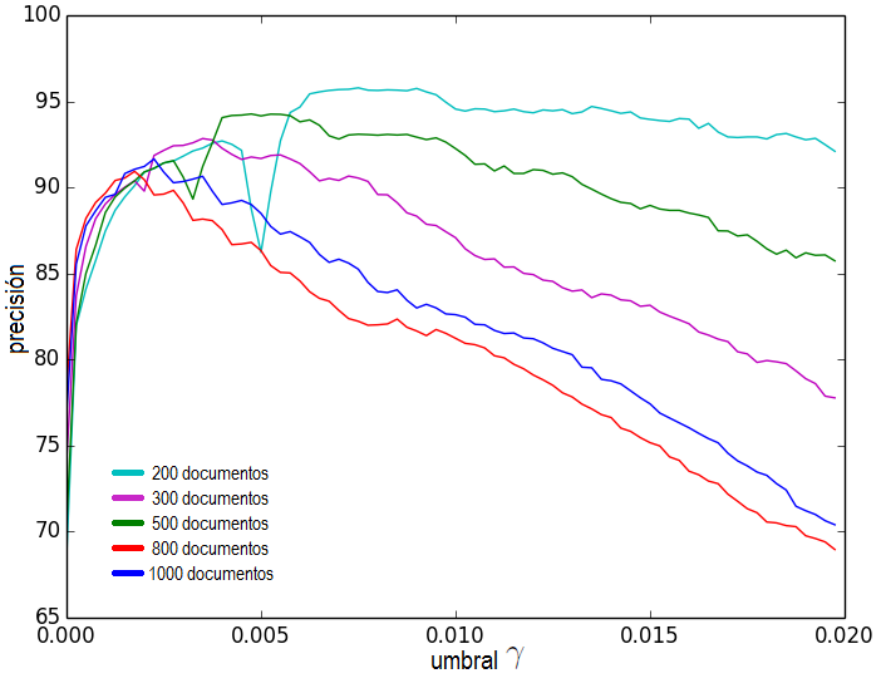


Figura 9: Umbral de ganancia de información vs precisión de clasificación con incrementos de .00025 promediado de 30 corridas.

de GI, tendremos un individuo que tome las características que tengan una GI mayor a 0.0025, y este individuo podrá evolucionar con el resto de la población generada aleatoriamente en el AG.

4.1.3. Parámetros del algoritmo genético

Se consideraron probabilidades de que ocurra cruzamiento o mutación de 0.7 y 0.3, respectivamente. Tomando estas probabilidades se observó la convergencia de la función objetivo, ver Figura 10, en esta la curva verde muestra el promedio del valor de la función objetivo de toda la población y la azul el valor del mejor individuo. Con estos resultados se ha elegido manejar estos valores ya que se han obtenido buenos resultados, así como utilizar una población de 80 individuos por 200 generaciones.

Se utilizó un kernel lineal para el SVM con los parámetros en sus valores por omisión[3]. Se consideró hacer validación cruzada de 10 pliegues, sin embargo, esto aumentaba el tiempo de ejecución. Con validación cruzada, el algoritmo tarda aproximadamente 14 horas (en

[3]C=1.0, degree=3, gamma=0.0, coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, random_state=None

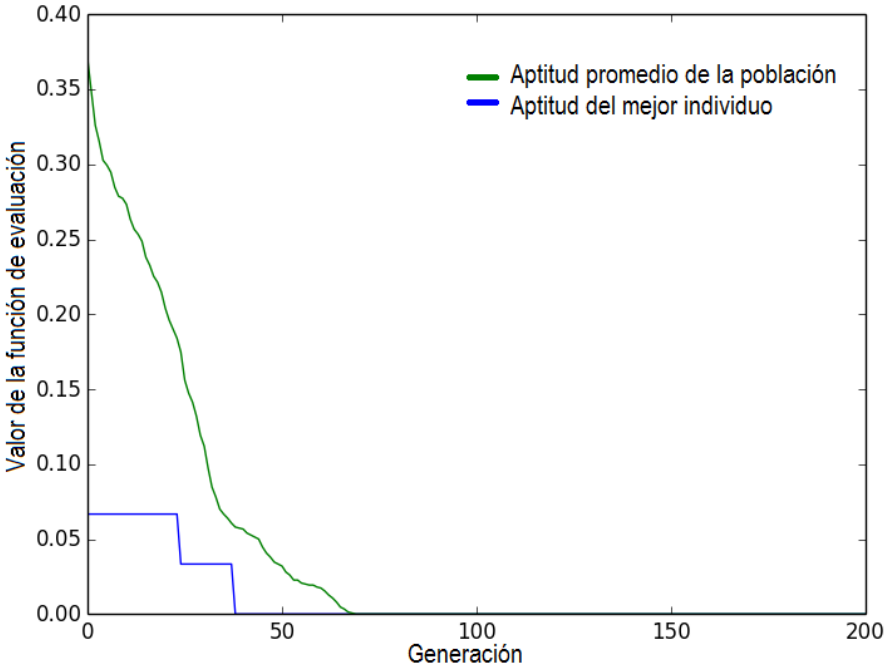


Figura 10: Gráfica de convergencia de la función objetivo con una población de 80 individuos por 200 generaciones.

un servidor Mac descrito en la siguiente sección) en generar la población inicial, ya que para cada individuo calcula la precisión de clasificación en la función objetivo. Además, por cada generación tarda aproximadamente dos horas, lo que resultaría en 17 días para obtener resultados de una sola corrida con 2000 documentos. Esto para una sola corrida completa del AG con una población de 80 individuos por 200 generaciones. Debido a esto se decidió cambiar el esquema de validación cruzada de 10 pliegues por uno donde el 90 % de los documentos se usan para el entrenamiento del SVM y el 10 % para prueba, esto se realiza una sola vez por individuo, lo cual hace que el algoritmo tarde entre cinco y ocho horas en total (en el mismo servidor Mac).

4.2. Evaluación de algoritmo genético

Debido al gran costo computacional que se requiere para evaluar cada individuo de la población, se utilizaron dos computadoras para hacer distintas pruebas. Para pruebas con menos de 1000 documentos se utilizó una computadora personal ASUS con 3 GB de RAM y un procesador Intel Pentium con una frecuencia de reloj de 2 GHz. Para pruebas con los 2000 documentos del corpus se utilizó un servidor Mac Pro con 64 GB de RAM y un procesador

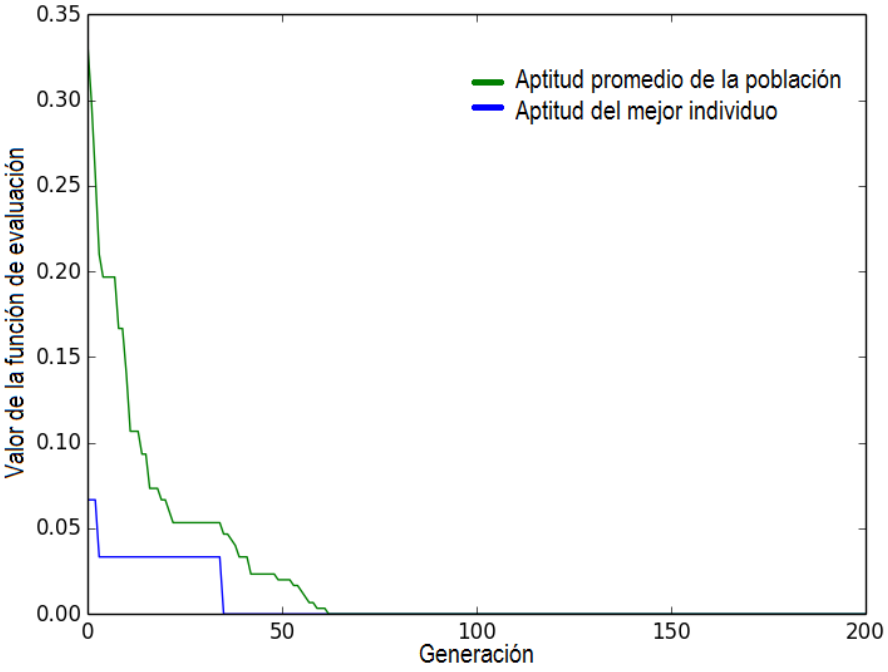


Figura 11: Convergencia de la función objetivo utilizando 10 individuos por 200 generaciones con una semilla de cero en la función *train_test_split*.

6-Core Intel Xeon de 2x2.66 GHz.

A continuación se presenta la descripción de los experimentos realizados, así como los datos de entrada y resultados.

Se usó una población de 10 individuos por 200 generaciones para 300 documentos. Un 90 % de documentos fueron utilizados para entrenar el clasificador y un 10 % para probarlo. Estos conjuntos de documentos fueron elegidos aleatoriamente con una función llamada *train_test_split* de la biblioteca *sklearn.cross_validation* que utiliza un parámetro semilla para elegir la forma en que se tomarán los documentos. Esta semilla se toma con el valor igual a cero. En la primer generación, el mejor individuo obtenido fue el de ganancia de información utilizando un total de 20868 características y con un error de precisión igual a 0.0667. El error de precisión lo definimos como $1 - precisión\_de\_clasificación$. Después de las 200 generaciones el error cae a 0 con 11255 características obteniendo 14 documentos negativos y 16 positivos clasificados correctamente. Revisando en la Figura 11 la gráfica de convergencia de la función objetivo se aprecia que el mejor individuo con error de 0, se obtiene desde la generación 34. La duración de la corrida fue de nueve minutos en la computadora personal.

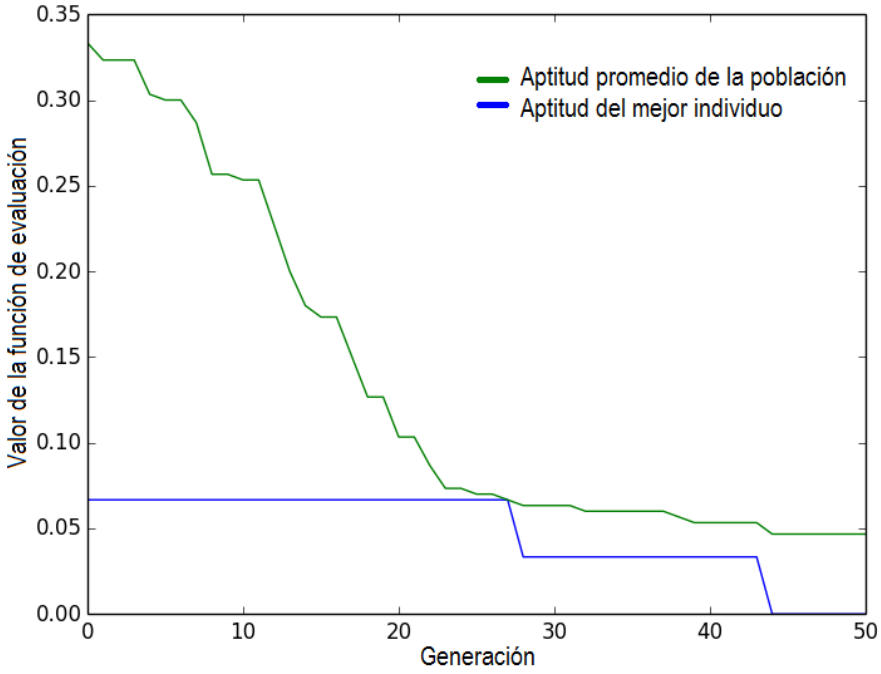


Figura 12: Convergencia de la función objetivo utilizando 10 individuos por 50 generaciones con una semilla de uno en la función train_test_split.

Se repitió el experimento cambiando la semilla en la función *train_test_split* a uno para que tome diferentes documentos de entrenamiento y prueba. Debido a que en el experimento anterior se llegó a un individuo con error de 0 desde la generación 20, disminuimos el número de generaciones a 50 esperando que el algoritmo converja en pocas generaciones. Después de las 50 generaciones el mejor individuo tiene un error de 0 con 11197 características obteniendo 9 documentos negativos y 21 positivos clasificados correctamente. Revisando en la Figura 12 la gráfica de convergencia de la función objetivo se aprecia que el error cae a 0 en la generación 46. La duración de la corrida fue de cinco minutos en la computadora personal.

Se regresa la semilla a cero para que se tomen los mismos documentos de entrenamiento y prueba en los futuros experimentos. Después de 50 generaciones con una población de 10 individos, se obtiene un individuo con un error de 0.0333 y 14540 características con 14 documentos negativos y 15 documentos positivos clasificados correctamente y un documento positivo clasificado como negativo. Por lo que efectivamente se tomaron los mismos documentos para el entrenamiento y prueba. La duración de la corrida fue de cinco minutos en la computadora personal.

Para la siguiente prueba, se cambia el porcentaje de documentos para entrenamiento y prueba en el SVM, utilizando el 75 % de documentos para el entrenamiento y 25 % para la prueba, tomados aleatoriamente con la función train_test_split. Así, el mejor individuo en la primer generación es el individuo creado con ganancia de información con un error de precisión de 0.08 y 20868 características. Después de 50 generaciones el mejor individuo presenta un error de 0.04 y 6883 características donde clasificó correctamente 34 negativos y 38 positivos, y 3 documentos positivos clasificados como negativos. La duración de la corrida fue de tres minutos en la computadora personal.

Se cambió el tamaño de población a 50 individuos y se evaluó por 200 generaciones. En la población final, todos los individuos llegaron a un error de 0, pero como no se están tomando en cuenta el número de características, el mejor individuo tiene 3669 características y el peor 3466. La duración de la corrida fue de 33 minutos en la computadora personal.

En este punto sólo se tomaba la precisión de clasificación de los documentos en la función objetivo, sin tomar en cuenta el número de características. Es por esto que, en ocasiones cuando toda la población tenía el mismo valor de error, el primer individuo podía tener más características que otros individuos por lo que no necesariamente era el mejor, tomando en cuenta que la población siempre está ordenada de menor a mayor en cuanto al valor de la función de aptitud. A partir de esto, se decidió tomar en cuenta el número de características en la función objetivo normalizando el número de características de cada individuo.

Con el objetivo de tomar en cuenta el número de características que se toman para hacer la clasificación en el SVM para reducir costos de almacenamiento, se probó multiplicar el error obtenido con el número de características normalizado. Con esto, en la primer generación con 300 documentos se obtuvo un individuo mejor al de ganancia de información con 1711 características y su valor en la función objetivo de 0.0167, pero con precisión de 56 %. Después de 200 generaciones el mejor individuo tiene una sola característica con un valor en la función objetivo de 9.79×10^{-6} y la precisión de este individuo es de 56 %, donde clasificó correctamente 16 positivos, 1 negativo, y 13 negativos se clasifican como positivos. Debido a estos malos resultados, se decide sumar el error de precisión con el número de características normalizado en lugar de multiplicarlos.

Tabla 6: Promedio de resultados del mejor individuo tras 200 generaciones con distintos pesos para 300 documentos, con $f_e = error\_de\_clasificación + p \times \bar{nc}$.

p	promedio de no. características	promedio de error
0	8373	0
0.25	1277	0
0.4	1164.6	0
0.5	1368.6	0
0.6	1078.8	0
0.75	1399.6	0
1	1463.2	0.033134677

En la Tabla 6 se muestran los resultados en promedio de cinco corridas de los mejores individuos al terminar 200 generaciones y 50 individuos con distintos pesos para 300 documentos con la función de evaluación $f_e = error\_de\_clasificación + p \times \bar{nc}$. Por cada peso $p \in (0, 0.25, 0.4, 0.5, 0.6, 0.75, 1)$ se corrió el algoritmo genético cinco veces, se guardaron los datos del mejor individuo y se obtuvo un promedio de los resultados. Se observa que para todos los pesos, excepto con el peso 1, se llega a un error de cero, lo cual significa que clasificó el 10 % de los documentos correctamente en cada caso. Sin embargo, se decide ponderar el error de precisión en la función objetivo en relación al peso que multiplica al número de características. Entonces, tenemos que el número de características normalizado se multiplica por el peso p y el error se multiplica por la diferencia de $1 - p$.

Tabla 7: Promedio de resultados del mejor individuo tras 200 generaciones con distintos pesos para 300 documentos, con $f_e = (1 - p) \times error\_de\_clasificación + p \times \bar{nc}$.

p	promedio de no. características	promedio de error
0	8647	0
0.25	1157.4	0
0.4	644.4	0
0.5	754.2	0.006
0.6	895.8	0.06
0.75	510.4	0.106
1	1	0.49

Se agrega este peso $(1 - p)$ al valor del error, por el hecho de que el algoritmo se enfoca más en disminuir el número de características solamente, sin tomar en cuenta la precisión de clasificación que es lo que más nos interesa. En la Tabla 7 se muestran del mismo modo los resultados en promedio de 5 corridas de los mejores individuos al terminar 200 generaciones y 50 individuos con distintos pesos para 300 documentos, con la función de evaluación $f_e =$

$(1-p) \times error\_de\_clasificación + p \times \bar{nc}$. En esta tabla, el peso afecta tanto al número de características como al error de precisión. Se realizó la misma prueba para 500 documentos (Tabla 8) y para 800 documentos (Tabla 9).

Tabla 8: Promedio de resultados del mejor individuo tras 200 generaciones con distintos pesos para 500 documentos, con $f_e = (1-p) \times error\_de\_clasificación + p \times \bar{nc}$.

p	promedio de no. características	promedio de error
0	22368.6	0.016
0.25	1801	0.008
0.4	1866	0.02
0.5	1400	0.02
0.6	1353.4	0.048
0.75	740.6	0.124
1	1	0.42

Tabla 9: Promedio de resultados del mejor individuo tras 200 generaciones con distintos pesos para 800 documentos, con $f_e = (1-p) \times error\_de\_clasificación + p \times \bar{nc}$.

p	promedio de no. características	promedio de error
0	14148.8	0.0075
0.25	5057.6	0.025
0.4	2215.6	0.0155
0.5	2156.2	0.01
0.6	2125.6	0.0255
0.75	1042.4	0.0575
1	1	0.4375

Para este momento, el 10 % de los documentos que se utilizaban para prueba, se tomaron aleatoriamente con una semilla de 0 en la función train_test_split. Sin embargo, aunque fuera aleatorio, siempre era el mismo 10 % en cada prueba, por lo que se decide eliminar la semilla de modo que el 10 % de documentos de prueba se elija de forma aleatoria para cada individuo. Esto es porque el algoritmo genético se concentraba más en clasificar bien ese pequeño conjunto de documentos que en clasificar bien cualquier conjunto de documentos. En otras palabras, se generaba un sobre-aprendizaje.

En las Tablas 10 y 11, podemos ver los resultados del mejor individuo de la última generación para 300 y 800 documentos, respectivamente. Del mismo modo, se obtiene el promedio de 5 corridas para los distintos pesos usando el 90 % de los documentos como entrenamiento y el 10 % restante para prueba. En la Tabla 12 se muestran los resultados de elegir 50 % de los documentos para entrenamiento, dejando el otro 50 % de documentos

para prueba. Ya que en promedio el error es mayor que en los experimentos anteriores, se decide seguir probando la clasificación con 90 % de documentos para entrenamiento y 10 % de prueba, y se decide probar con $p \in [0.25, 0.5]$ con 2000 documentos.

Tabla 10: Promedio de resultados del mejor individuo tras 200 generaciones con distintos pesos para 300 documentos, con $f_e = (1-p) \times error\_de\_clasificación + p \times \bar{nc}$.

p	promedio de no. características	promedio de error
0	10109	0
0.25	1497	0
0.4	868.8	0
0.5	1570.4	0
0.6	773.6	0.0734
0.75	586	0.1
1	1	0.6

Tabla 11: Promedio de resultados del mejor individuo tras 200 generaciones con distintos pesos para 800 documentos, con $f_e = (1-p) \times error\_de\_clasificación + p \times \bar{nc}$.

p	promedio de no. características	promedio de error
0	16127.4	0.005
0.25	6984.6	0.0075
0.4	3420.4	0.02
0.5	2661.6	0.0275
0.6	1850.4	0.045
0.75	1484.8	0.0575
1	1	0.4825

Tabla 12: Promedio de resultados del mejor individuo tras 200 generaciones con distintos pesos para 300 documentos, usando 50 % de los datos para entrenamiento y 50 % para prueba.

p	promedio de no. características	promedio de error
0	16248	0.0254
0.25	5292.6	0.0628
0.4	2936	0.072
0.5	1371.8	0.2162
0.6	1107	0.192
0.75	346.4	0.2682
1	1	0.5234

Además, podemos ver en la Figura 13 la gráfica de la convergencia de la población y el mejor individuo de una corrida con $p = 0.25$, donde se ve que efectivamente disminuye el valor de la función objetivo conforme pasan las generaciones. Esto quiere decir que el error de precisión disminuye al mismo tiempo que se reduce el número de características.

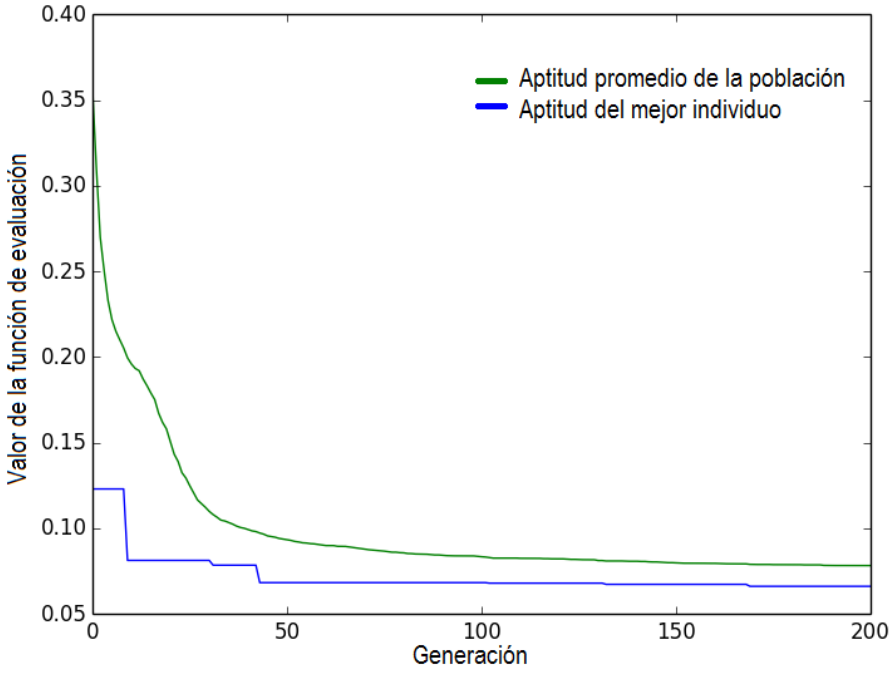


Figura 13: Convergencia de la función objetivo con $p = 0.25$.

En resumen podemos ver en la Tabla 13 los parámetros del AG que se tienen para pruebas con 2000 documentos y proceder a realizar pruebas con documentos nuevos, los cuales no se tomaron en cuenta para la extracción de características y llamamos documentos desconocidos.

Tabla 13: Parámetros del AG para el problema de selección de características en el análisis de sentimientos en español.

Parámetros	Valor
Tamaño de la población	80
Número de generaciones	200
Probabilidad de cruzamiento	0.7
Probabilidad de mutación	0.3
Selección	torneo y ruleta
Tipo de cruzamiento	unión, intersección, diferencia
Tipo de mutación	agregar, eliminar

Para hacer pruebas con documentos desconocidos (los 624 documentos que se han mantenido aparte) es necesario guardar los resultados del algoritmo genético para reutilizarlos. Para estas pruebas es necesario conocer las características del mejor individuo así como el modelo utilizado para hacer la clasificación. Este modelo es un SVM que ya está entrenado con el 90 % de los documentos desde el algoritmo genético.

Python maneja un módulo llamado Pickle[4], que sirve para guardar objetos, y ya que en Python todo es un objeto, nos sirve tanto para guardar las características y el SVM del mejor individuo. El módulo pickle implementa un algoritmo fundamental, pero de gran alcance para serializar con la función pickle.dump(obj, file[, protocol]) y de-serializar con la función pickle.load(file) una estructura de objetos de Python.

4.3. Pruebas con documentos desconocidos

Se cuenta con 231004 características para 2000 documentos, y en la Tabla 14 se puede ver que las características predominantes son los bigramas y trigramas. Recordemos que estas características se obtienen únicamente de los 2000 documentos y se eliminan las que no aparecen en más de un documento. Se realizaron varias corridas del algoritmo genético con distintos pesos y se guardaron los resultados, tal como se hizo para 300, 500 y 800 documentos. Sin embargo, al trabajar con los documentos desconocidos la precisión en todos los casos bajó a 50 % o menos. Por ejemplo, con 2000 documentos y peso de 0.25 tenemos que la precisión en los documentos desconocidos es de 50.16 %, donde en la matriz de confusión tenemos los valores de 206 verdaderos negativos; 107 verdaderos positivos; 235 falsos negativos; y 76 falsos positivos. La matriz de confusión está compuesta por dos renglones y dos columnas, donde en el primer renglón tenemos que la primer columna son los verdaderos negativos y la segunda columna los falsos positivos, y en el segundo renglón tenemos que la primer columna son los falsos negativos y la segunda columna verdaderos positivos, esta se puede apreciar en la Figura 14. Debido a que un algoritmo aleatorio podría dar estos mismos resultados de 50 % de precisión, se decide eliminar los bigramas y trigramas, que aunque aparezcen en más de una ocasión en los documentos conocidos, no aparecen tanto en los documentos desconocidos.

	Negativo	**Positivo**
Negativo	Verdadero Negativo (TN)	Falso positivo (FP)
Positivo	Falso Negativo (FN)	Verdadero Positivo (TP)

Figura 14: Matriz de confusión para dos clases.

Con 2000 documentos, eliminando bigramas y trigramas, se cuenta con un total de 44121 características. Utilizando ganancia de información con un umbral de 0.0025 se reduce a 1809 características, el individuo que contiene estas características obtiene una precisión de 81.5 %.

[4]https://docs.python.org/2/library/pickle.html

Tabla 14: Número de características de cada tipo.

Tipo de característica	**Número de características**
unigramas	27643
bigramas	93666
trigramas	93217
lemas	6879
POS-tag	317
POS-bigramas	9282
total	231004

Utilizando el modelo generado con estas características en los documentos desconocidos, se obtiene una precisión de 79.64 %, donde en la matriz de confusión tenemos los valores de 238 verdaderos negativos; 259 verdaderos positivos; 83 falsos negativos; y 44 falsos positivos.

Tabla 15: Resultados con distintos pesos de documentos desconocidos comparado con los conocidos.

p	Número de características	Precisión en AG	Precisión en desconocidos	Matriz de confusión de desconocidos
0.25	2623	91.5	77.88	$\begin{bmatrix} 228 & 54 \\ 89 & 253 \end{bmatrix}$
0.4	818	90.5	73.87	$\begin{bmatrix} 288 & 74 \\ 94 & 248 \end{bmatrix}$
0.5	555	90.5	74.67	$\begin{bmatrix} 232 & 50 \\ 108 & 234 \end{bmatrix}$

Se obtuvieron los mejores individuos en diferentes corridas del algoritmo genético con pesos de 0.25, 0.4 y 0.5. En la Tabla 15 podemos observar el número de características que se obtuvieron y la precisión que da la SVM generada con ese conjunto de características. Además se observa la precisión que se obtiene utilizando estas características en los documentos desconocidos, así como la matriz de confusión correspondiente.

Se corrió 10 veces más el algoritmo genético con $p = 0.25$ y $p = 0.5$, se guardaron los resultados y se presentan en las tablas 16 y 17, respectivamente. Allí se observan el número de características, la precisión en los documentos conocidos y desconocidos y la matriz de confusión de los desconocidos, para el mejor individuo al terminar las 200 generaciones del algoritmo genético.

Se guardó la población completa para algunas corridas y se observa en el resto de los individuos cuál clasificó mejor a los desconocidos. En la Tabla 18 se muestra la posición del

Tabla 16: Resultados con pesos 0.25 de documentos desconocidos comparado con los conocidos.

No. corrida	Número de características	Precisión en AG	Precisión en desconocidos	Matriz de confusión de desconocidos
1	710	90	72.6	$\begin{bmatrix} 212 & 70 \\ 101 & 241 \end{bmatrix}$
2	2623	91.5	77.88	$\begin{bmatrix} 228 & 54 \\ 89 & 253 \end{bmatrix}$
3	1261	90	75.96	$\begin{bmatrix} 230 & 52 \\ 98 & 244 \end{bmatrix}$
4	535	91	72.75	$\begin{bmatrix} 207 & 75 \\ 95 & 247 \end{bmatrix}$
5	895	90	75.32	$\begin{bmatrix} 217 & 65 \\ 89 & 253 \end{bmatrix}$
6	1296	89	79.48	$\begin{bmatrix} 227 & 55 \\ 73 & 269 \end{bmatrix}$
7	1079	89.5	76.44	$\begin{bmatrix} 212 & 70 \\ 77 & 265 \end{bmatrix}$
8	705	90	74.67	$\begin{bmatrix} 233 & 49 \\ 109 & 233 \end{bmatrix}$
9	1057	90	76.28	$\begin{bmatrix} 225 & 57 \\ 91 & 251 \end{bmatrix}$
10	663	89.5	71.31	$\begin{bmatrix} 218 & 64 \\ 115 & 227 \end{bmatrix}$

individuo que se desempeña mejor con los documentos desconocidos, así como el número de características con las que cuenta, la precisión que obtuvo con los documentos conocidos dentro del algoritmo genético, la precisión que tiene con los documentos desconocidos y su matriz de confusión. Estos cinco individuos forman parte de una población de 80 individuos, donde el primer individuo es aquel que tiene un mejor valor en la función objetivo según la ecuación 5, con $p = 0.25$. En la Tabla 19, del mismo modo, se muestra el mejor individuo de cinco corridas distintas con $p = 0.5$.

Así mismo podemos observar las precisiones de la población entera en las figuras 15 y 16 con pesos $p = 0.25$ y $p = 0.5$, respectivamente. Por cada individuo, según el número de características que contiene, podemos ver la precisión sobre los documentos conocidos y sobre los documentos desconocidos. En estas figuras, se grafica toda la población obtenida al final de una corrida con $p = 0.25$ y una corrida con $p = 0.5$, respectivamente. De color rojo se muestra la precisión sobre los documentos conocidos y de azul la precisión sobre los documentos desconocidos correspondientes al individuo que contiene el número de características marcado

Tabla 17: Resultados con pesos 0.5 de documentos desconocidos comparado con los conocidos.

No. corrida	Número de características	Precisión en AG	Precisión en desconocidos	Matriz de confusión de desconocidos
1	495	88.5	70.51	$\begin{bmatrix} 207 & 76 \\ 109 & 232 \end{bmatrix}$
2	555	90.5	74.67	$\begin{bmatrix} 232 & 50 \\ 108 & 234 \end{bmatrix}$
3	715	88.5	75.48	$\begin{bmatrix} 222 & 60 \\ 93 & 249 \end{bmatrix}$
4	826	89	73.39	$\begin{bmatrix} 229 & 53 \\ 113 & 229 \end{bmatrix}$
5	832	89.5	71.79	$\begin{bmatrix} 220 & 62 \\ 114 & 228 \end{bmatrix}$
6	1134	90	79.16	$\begin{bmatrix} 238 & 44 \\ 86 & 256 \end{bmatrix}$
7	477	90.5	71.47	$\begin{bmatrix} 217 & 65 \\ 113 & 229 \end{bmatrix}$
8	584	88.5	72.11	$\begin{bmatrix} 210 & 72 \\ 102 & 240 \end{bmatrix}$
9	980	90.5	73.71	$\begin{bmatrix} 214 & 68 \\ 96 & 246 \end{bmatrix}$
10	614	88	73.87	$\begin{bmatrix} 214 & 68 \\ 95 & 247 \end{bmatrix}$

en el eje x. En la Figura 15 se ve la mayoría de la población en un valor cercano a las 1500 características y en la Figura 16 cercano a las 600 características. En ambas gráficas se aprecia que la precisión, tanto de los documentos conocidos como desconocidos, aumenta entre más características se tengan.

Se creó otra población con $p = 0.25$ corriendo el algoritmo genético para tener 11 individuos, el mejor de cada corrida, y construir con los modelos obtenidos un consenso. Por cada documento desconocido, se clasificó con todos los individuos y se tomó la clase que eligiera la mayoría. Si seis o más individuos toman a un documento como positivo, este se clasifica como positivo y en el caso contrario como negativo. De este modo se clasificaron correctamente 82.85 % de los documentos, donde en la matriz de confusión tenemos los valores de 244 verdaderos negativos; 273 verdaderos positivos; 69 falsos negativos; y 38 falsos positivos.

Los resultados de la clasificación en los documentos desconocidos, en resumen, se pueden ver en la Tabla 20 con las 11 corridas utilizando un peso de 0.25 y el consenso. Se observa

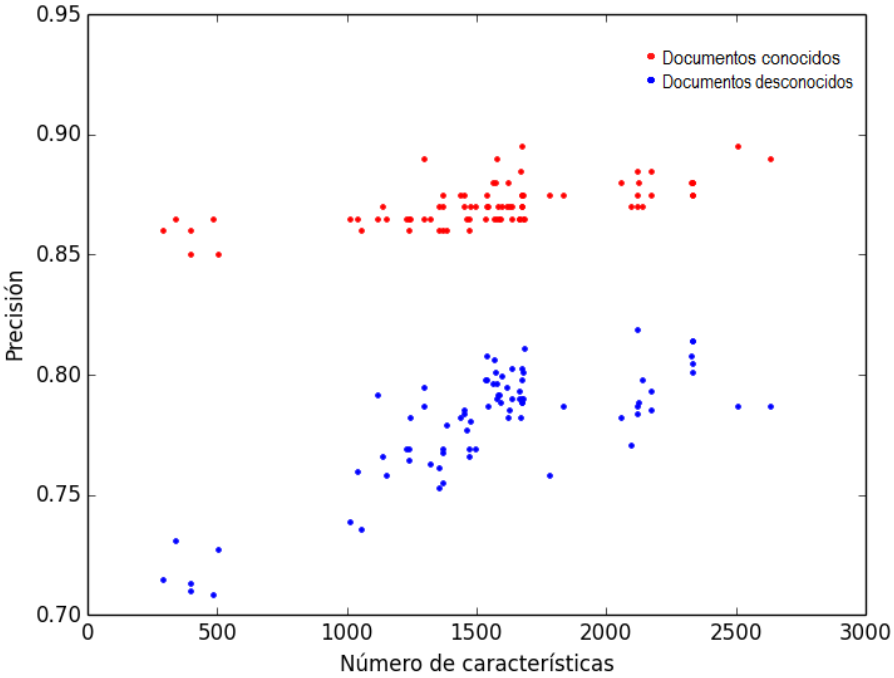


Figura 15: Población con peso 0.25.

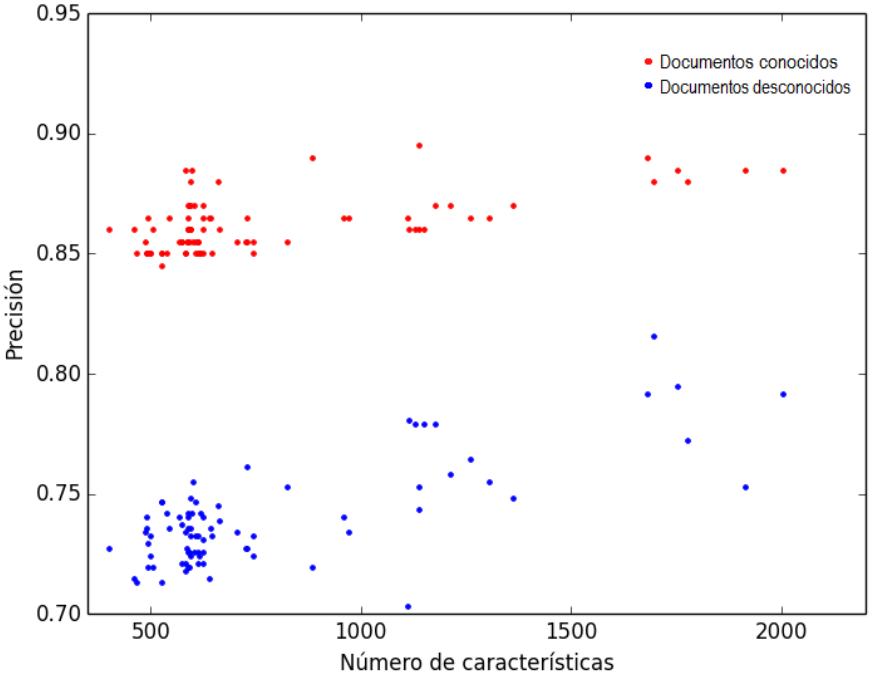


Figura 16: Población con peso 0.5.

Tabla 18: Resultados del individuo con mejor desenvolvimiento en documentos desconocidos con peso 0.25.

Posición	Número de características	Precisión en AG	Precisión en desconocidos	Matriz de confusión de desconocidos
9	1752	89	80.28	$\begin{bmatrix} 239 & 43 \\ 80 & 262 \end{bmatrix}$
55	2120	87.5	81.89	$\begin{bmatrix} 237 & 45 \\ 68 & 274 \end{bmatrix}$
70	1992	87	81.08	$\begin{bmatrix} 243 & 39 \\ 79 & 263 \end{bmatrix}$
67	1950	87.5	79	$\begin{bmatrix} 239 & 43 \\ 88 & 254 \end{bmatrix}$
11	2446	89.5	80.6	$\begin{bmatrix} 229 & 53 \\ 68 & 274 \end{bmatrix}$

Tabla 19: Resultados del individuo con mejor desenvolvimiento en documentos desconocidos con peso 0.5.

Posición	Número de características	Precisión en conocidos	Precisión en desconocidos	Matriz de confusión de desconocidos
18	1862	88.5	80.76	$\begin{bmatrix} 242 & 40 \\ 80 & 262 \end{bmatrix}$
74	1677	87.5	78.36	$\begin{bmatrix} 227 & 55 \\ 80 & 262 \end{bmatrix}$
65	1779	87.5	77.88	$\begin{bmatrix} 203 & 52 \\ 86 & 256 \end{bmatrix}$
27	1994	89	81.08	$\begin{bmatrix} 234 & 48 \\ 70 & 272 \end{bmatrix}$
43	1655	87.5	80.76	$\begin{bmatrix} 238 & 44 \\ 76 & 266 \end{bmatrix}$

por separado los falsos positivos (FP), falsos negativos (FP), verdaderos positivos (TP), verdaderos negativos (TN), la precisión de clasificación en los documentos desconocidos y el número de características que se utilizó.

4.4. Análisis de resultados

En la Tabla 21 se observan las precisiones obtenidas por el modelo desarrollado con el AG en 11 corridas utilizando un peso $p = 0.25$ en la función de evaluación $f_e = (1 - p) \times error\_de\_clasificación + p \times \bar{nc}$, esto es en los documentos conocidos. El promedio de precisión obtenido es 90.09 % con una desviación estandar 0.70. Estos datos siguen una distribución parecida a la normal como se puede ver en la gráfica de su función de densidad que tiene una

Tabla 20: Resumen de resultados con peso 0.25 en documentos desconocidos.

corrida	TP	FN	FP	TN	Precisión	#características
1	241	101	70	212	72.6	710
2	253	89	54	228	77.88	2623
3	244	98	52	230	75.96	1261
4	247	95	75	207	72.75	535
5	253	89	65	217	75.32	895
6	269	73	55	227	79.48	1296
7	265	77	70	212	76.44	1079
8	233	109	49	233	74.67	705
9	251	91	57	225	76.28	1057
10	227	115	64	218	71.31	663
11	248	94	63	219	74.83	374
consenso	273	69	38	244	82.85	

forma acampanada en la Figura 17.

Tabla 21: Resumen de precisiones con peso 0.25 con documentos conocidos.

Características	Precisión
710	90
2623	91.5
1261	90
535	91
859	90
1296	89
1079	89.5
705	90
1057	90
663	89.5
374	90.5

Los autores Cruz *et al.* (2008) y Martínez *et al.* (2011) utilizan el mismo corpus que en este trabajo. Cruz *et al.* (2008) utilizan un enfoque semántico y la configuración óptima que su sistema ofrece es de 77.5 %, mientras que Martínez *et al.* (2011) obtienen una precisión de 86.40 % utilizando SVM y reduciendo las características con el uso de *stopper* y 85.67 % utilizando también *stemmer* que es su configuración más eficiente. En este trabajo el algoritmo genético nos ha dado una precisión de 91.5 %. Martínez *et al.* (2011) no han mostrado resultados utilizando documentos desconocidos ya que utilizaron todo el corpus para el clasificador, y en nuestro caso hemos obtenido hasta un 79.48 % de precisión de clasificación, lo cual es 1.98 puntos mayor que la configuración de Cruz *et al.* (2008). Estos resultados no son

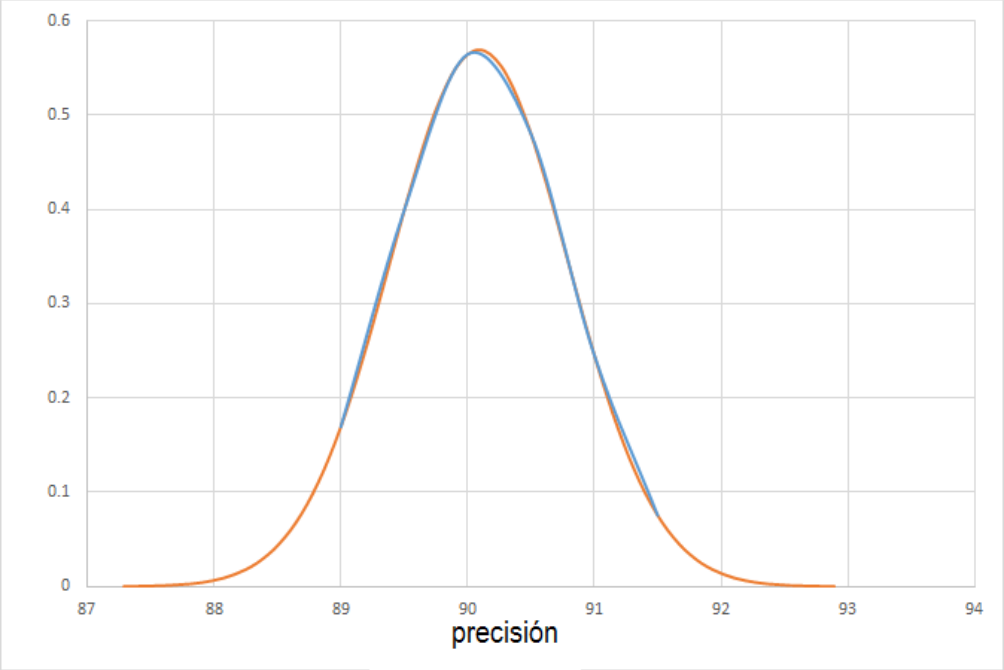


Figura 17: Precisión de los mejores individuos con peso 0.25.

estadísticamente significativos, ya que no se usaron los mismos documentos de entrenamiento y prueba.

Entre nuestros resultados, la mejor precisión utilizando 2000 documentos fue de 91.5 %, sin embargo, este individuo utiliza 2623 características que son más de las 1809 obtenidas utilizando GI donde se cuenta con una precisión de 81.5 %. Se cuenta también con resultados con menor número de características y mejor precisión que utilizando solo ganancia de información. Por ejemplo, se cuenta con un individuo con 535 características y un SVM con precisión de 91 % obtenido por el AG. Estas características se pueden consultar en el Apéndice A, así como un ejemplo de un documento del conjunto de documentos desconocidos que la SVM dio como resultado en verdadero positivo, uno verdadero negativo, uno falso positivo y uno falso negativo.

En las Tablas 22, 23, 24 y 25, se muestran las proporciones de los tipos de características tomadas por el individuo con GI y por el mejor individuo de cada corrida con $p = 0.25$. En la Tabla 22, se observa el total de unigramas elegidos, el porcentaje de unigramas con respecto a las otras características elegidas (i.e. el 59.92 % de las características seleccionadas fueron unigramas) y el porcentaje de unigramas elegidos con respecto al número total de unigramas. De la misma forma tenemos la Tabla 23 para lemas, la Tabla 24 para POS-tag y la Tabla 25

para POS-bigramas.

Tabla 22: Porcentaje de unigramas elegidos.

Individuo	unigramas	% entre características elegidas	% entre total unigramas
GI	1084	59.92	3.92
1	425	59.86	1.54
2	1609	61.34	5.82
3	741	58.76	2.68
4	342	63.93	1.24
5	521	58.21	1.88
6	775	59.80	2.80
7	642	59.50	2.32
8	429	60.85	1.55
9	629	59.51	2.28
10	387	58.37	1.40
11	213	56.95	0.77

Tabla 23: Porcentaje de lemas elegidos.

Individuo	lemas	% entre características elegidas	% entre total lemas
GI	276	15.26	4.01
1	104	14.65	1.51
2	400	15.25	5.81
3	204	16.18	2.97
4	71	13.27	1.03
5	132	14.75	1.92
6	200	15.43	2.91
7	168	15.57	2.44
8	107	15.18	1.56
9	171	16.18	2.49
10	110	16.59	1.60
11	52	13.90	0.76

Podemos ver en la Tabla 22 que tanto los individuos obtenidos con el AG como el individuo obtenido con GI están conformados principalmente por unigramas (59.75 % en promedio), en segundo lugar, podemos observar en la Tabla 25, se tienen los POS-bigramas (22.85 % en promedio), en tercer lugar, podemos ver en la Tabla 23, se tienen los lemas (15.18 % en promedio), y por último están conformados por POS-tag (2.22 % en promedio), que se observa en la Tabla 24. Son similares la proporción de características elegidas tanto por los individuos obtenidos con el AG como con las elegidas por el individuo con GI. Sin embargo, el individuo con ganancia de información utiliza más características que la gran mayoría de los individuos obtenidos por el AG y obtiene una precisión más baja con los documentos

Tabla 24: Porcentaje de POS-tag elegidos por el individuo de ganancia de información y por los mejores individuos de distintas corridas.

Individuo	POS-tag	% entre características elegidas	% entre total POS-tag
GI	43	2.38	13.56
1	20	2.82	6.31
2	48	1.83	15.14
3	25	1.98	7.89
4	9	1.68	2.84
5	19	2.12	5.99
6	27	2.08	8.52
7	20	1.85	6.31
8	13	1.84	4.10
9	19	1.80	5.99
10	18	2.71	5.68
11	13	3.48	4.10

Tabla 25: Porcentaje de POS-bigramas elegidos.

Individuo	POS-bigramas	% entre características elegidas	% entre total POS-bigramas
GI	406	22.44	4.37
1	161	22.68	1.73
2	566	21.58	6.10
3	291	23.08	3.14
4	113	21.12	1.22
5	223	24.92	2.40
6	294	22.69	3.17
7	249	23.08	2.68
8	156	22.13	1.68
9	238	22.52	2.56
10	148	22.32	1.59
11	96	25.67	1.03

conocidos.

Comparando el porcentaje que se eligió entre el mismo tipo de característica en las tablas 22–25 tenemos que en promedio se eligió 2.35 % unigramas entre el número total de unigramas, 2.42 % con los lemas, 7.2 % con los POS-tag y 2.64 % con los POS-bigramas. Siendo las características tipo POS-tag las que más se eligieron entre el total de características POS-tag.

4.4.1. Discusión

La precisión con los documentos desconocidos es sobre 624 documentos contra 200 documentos que son los documentos conocidos del AGAS. Este puede ser un factor a considerar

del porqué baja la precisión desde 90 % a 70 % (aproximadamente). En las pruebas con menos documentos la precisión que se obtuvo con el AGAS fue de 100 % en algunos casos ya que eran sobre 30, 50 y 80 documentos.

En casi todas las corridas con peso $p = 0.25$ y $p = 0.5$ se obtuvieron individuos con menos características que el individuo generado con GI, el cual contiene 1809 características y precisión de 81.5 % con documentos conocidos y 79.64 % con desconocidos. El único individuo obtenido con mayor número de características fue de 2623 características. Este individuo obtuvo mejor precisión con los documentos conocidos (91.5 %), sin embargo, con los desconocidos disminuye a 77.88 %. Este comportamiento se mantiene entre los mejores individuos que se obtuvieron de cada corrida, son mejores que el individuo generado con GI en los documentos conocidos, pero peores con los documentos desconocidos.

Se obtuvieron individuos con mejor precisión que el individuo generado con GI tanto en los documentos conocidos como en los documentos desconocidos (Tablas 18 y 19). Sin embargo, estos individuos no obtuvieron la primer posición de la corrida donde se obtuvieron, y en cada corrida el mejor tiene una posición diferente por lo que no hay forma de prever la posición en la que se encontrará este individuo.

Capítulo 5. Conclusiones

En este capítulo se presenta un sumario del trabajo de investigación realizado, así como las conclusiones a las que se llegó y algunas propuestas de ideas para trabajo a futuro.

5.1. Sumario

En este trabajo se propuso un método basado en un algoritmo genético para resolver el problema de selección de características dentro del análisis de sentimientos. Se utilizó un enfoque de envoltura para seleccionar el subconjunto de características con el AG evaluando su rendimiento con la precisión de un clasificador, que para nuestro estudio se utilizó una SVM.

Se trabajó con el corpus de *Muchocine* el cual consta de críticas de cine en español escritas por usuarios comunes. Cada documento se compone en promedio de 500 palabras resultando en un total de 44,121 características con 2,000 documentos entre los cuales hay 27,643 unigramas, 6,879 lemas, 317 POS-tag y 9,282 POS-bigramas. Se aplicó un filtro para eliminar las características que no aparecieran en más de un documento.

Se propusieron operadores evolutivos de mutación y cruzamiento, para manejar individuos que se comportan como conjuntos. Las operaciones de cruzamiento se definieron como la unión, intersección y diferencia de conjunto de características, mientras que la mutación consistió en agregar o eliminar características del individuo a mutar.

Se realizó un conteo de los tipos de características que elige el algoritmo genético. Así mismo se han guardado los individuos presentados en este trabajo, de modo que se pueden obtener fácilmente las características elegidas por el AG.

Como cada ejecución del algoritmo genético genera al menos un individuo con precisión máxima, se tomaron 11 de estos individuos, uno por corrida, para clasificar documentos según los resultados que elegía la mayoría de los modelos de clasificación pertenecientes a estos 11 individuos. Esto con el objetivo de comparar el resultado que logra el consenso de los modelos con el logrado por cada modelo por separado.

5.2. Conclusiones

Un análisis de los resultados obtenidos nos permite llegar a las siguientes conclusiones:

El algoritmo genético implementado dio resultados competitivos (91.5 %) al clasificar documentos conocidos. Bajo el contexto de documentos en español, y a lo mejor de nuestro conocimiento, se ha obtenido la mejor precisión que se ha documentado. Cabe mencionar que esta precisión se logró con el 10 % de los documentos que se tomaron de prueba mientras que el 90 % restante se utilizó para el entrenamiento.

Las operaciones propuestas de mutación y cruzamiento funcionaron adecuadamente para nuestro problema por la estructura de los individuos, ya que el AG siempre convergió a un individuo con buena precisión de clasificación, mejorando la obtenida utilizando las características generadas al fijar un umbral de GI.

En los porcentajes de características obtenidos se observa que en todos los individuos predominan los unigramas, con un promedio de 59 % de unigramas entre las características elegidas. Además se obtiene que entre las características de tipo POS-tag se elige en promedio 7.2 % del total de características de ese tipo, el cual es un mayor porcentaje con respecto a los otros tipos de características que no son mayores en promedio a 2.6 %. Esto podría significar que los unigramas y los POS-tag son características relevantes. Un argumento a favor de esta hipótesis es que los unigramas contienen más información que los lemas al tener género, conjugación, número. Por su parte, los POS-tag también contienen esta información, pero además, es común encontrarlos en los documentos por no ser una palabra específica.

Aunque la clasificación dentro del AG, que es donde se utilizan únicamente los documentos conocidos, haya brindado una precisión confiable de 90.09 % en promedio, esta precisión baja a 75 % clasificando los documentos desconocidos. Sin embargo, el sistema de consenso propuesto obtuvo 82.85 % de precisión, la cual es la mejor precisión obtenida con los documentos desconocidos empleados y podría dar mejores resultados con documentos en problemas de la vida real, ya que son $2k + 1$ clasificadores que dictarán si un documento tiene una opinión positiva o negativa en lugar de solo uno.

5.3. Trabajo a futuro

Las características que se han tomado para este trabajo tienen un contenido con alto componente semántico. En este trabajo, solamente se ha tomado en cuenta la presencia o ausencia de características, por lo que se pudieran mejorar los resultados agregando orientación semántica.

Los modelos de clasificación generados en este trabajo se han guardado, por lo que se pueden reutilizar para más pruebas con otros corpora. Debido a que los modelos de clasificación se entrenaron con un corpus de críticas de cine, lo recomendable es utilizar corpora bajo el mismo contexto. Esto es porque el estilo de escritura cambia bajo el contexto que se escribe, así como el vocabulario que se utiliza y su significado.

Existen características dentro de los documentos desconocidos que no se incluyen para seleccionarse con los documentos conocidos ya que estos no las contienen. Se propone extender el corpus y crear una bolsa de palabras al extraer todas las características, de modo que al agregar un nuevo documento, no aumente el número de características de esta bolsa de palabras.

Se considera importante explorar otro tipo de representación de los documentos, como individuos representados por árboles, y revisar su impacto dentro del AG así como en la clasificación de documentos desconocidos.

Existe un número elevado de contenidos generados por usuarios en la web que son idóneos para la creación de recursos. Se propone diseñar un sistema que pueda extraer nuevos documentos de páginas de Internet (v.g. Twitter, páginas de noticias, etc.) y con base en estos disponer de un nuevo corpus para seleccionar características con el AG, y entrenar nuevas SVMs. Así, con estas SVMs, se pudiera crear una aplicación donde un usuario pueda elegir algún objeto que puede ser sometido a crítica y ver los resultados de las opiniones que la gente ha escrito sobre este en Internet.

Lista de referencias

Abbasi, A., Chen, H., y Salem, A. (2008). Sentiment analysis in multiple languages. *ACM Transactions on Information Systems*, **26**(3): 1–34.

Ahmad, K., Glucksberg, S., Goatly, A., Hobbs, J. R., Gordon, A. S., Vogel, C., Wallington, A., Agerri, R., Branden, J., Lee, M., Rumbell, T., Dral, J., Heylen, D., y op den Akker, R. (2011). *Affective Computing and Sentiment Analysis*, Vol. 45 de *Text, Speech and Language Technology*. Springer Netherlands. Dordrecht.

Annett, M. y Kondrak, G. (2008). A comparison of sentiment analysis techniques: Polarizing movie blogs. En: *Proceedings of the Twenty-First Canadian Conference on Artificial Intelligence*, Springer, Heidelberg. pp. 25–35.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., y Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source nlp library. En: *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, May, Genoa, Italy. ELRA.

Banea, C., Mihalcea, R., Wiebe, J., y Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. En: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics, EMNLP '08, pp. 127–135.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. p. 738.

Boulis, C. y Ostendorf, M. (2005). Text Classification by Augmenting the Bag-of-Words Representation with Redundancy- Compensated Bigrams. En: *Workshop on Feature Selection in Data Mining, in conjunction with SIAM conference on Data Mining*, Seatle, USA.

Brooke, J., Tofiloski, M., y Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. En: K. Angelova, Galia andinproceedings Bontcheva, R. Mitkov, N. Nicolov, y N. Nikolov (eds.), *RANLP*. RANLP 2009 Organising Committee / ACL, pp. 50–54.

Cruz, F. L., Troyano, J. A., Enriquez, F., y Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, **41**: 73–80.

Cruz Mata, F. (2011). *Extracción de Opiniones sobre Características: Un Enfoque Práctico Adaptable al Dominio*. Tesis de doctorado, Universidad de Sevilla.

Duda, R., Hart, P., y Stork, D. (2001). *Pattern Classification*. John Wiley Sons, segunda edición. New York, NY, USA.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1): 61–74.

Duric, A. y Song, F. (2011). Feature selection for sentiment analysis based on content and syntax models. En: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011*, Portland, Oregon. pp. 96–103.

Eiben, A. E. y Smith, J. E. (2003). *Introduction to Evolutionary Computing*. SpringerVerlag.

El Ayadi, M., Kamel, M. S., y Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.*, **44**(3): 572–587.

Esuli, A. y Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. En: *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*. pp. 417–422.

García Castellano, F. J. (2009). *Modelos bayesianos para la clasificación supervisada. Aplicaciones al análisis de datos de expresión genética*. Tesis de doctorado, Universidad de Granada.

Gibbon, D., Moore, R., y Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter. Berlin.

Go, A., Bhayani, R., y Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pp. 1–6.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., primera edición. Boston, MA, USA.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press. Ann Arbor, MI.

Humphries, L. y McDonald, S. (2011). Emotion faces: The design and evaluation of a game for preschool children. En: *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA. ACM, CHI EA '11, pp. 1453–1458.

Kohavi, R. y John, G. H. (1997). Wrappers for feature subset selection. *Elsevier Science*, **97**(97): 273–324.

Koncz, P. y Paralic, J. (2011). An approach to feature selection for sentiment analysis. En: *5th IEEE International Conference on Intelligent Engineering Systems*, Junio, Poprad, Eslovaquia. IEEE, pp. 357–362.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Vol. 1. MIT Press. Cambridge, MA, USA.

Lewis, D. D. y Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. En: *Third Annual Symposium on Document Analysis and Information Retrieval*. pp. 81–93.

Li, Y., Li, S., Xu, W., y Guo, J. (2012). Analyzing semantic orientation of terms using affinity propagation. En: *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*, Dec. pp. 30–34.

Lin, Y.-C., Yang, Y.-H., y Chen, H. H. (2011). Exploiting online music tags for music emotion classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, **7S**(1): 26:1–26:16.

Liu, B. (2010). Sentiment analysis: a multi-faceted problem. *IEEE Intelligent Systems*, **25**(3): 76–80.

Liu, B., Hu, M., y Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. pp. 342–351.

Liu, G. y Lai, H. (2010). Predicting the semantic orientation of movie reviews. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, (FSKD): 2483–2488.

Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., y Ureña López, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. En: *Expert Systems with Applications*. pp. 3934–3942.

Martínez, E., Martín, T., Perea, J., y Ureña, A. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento de Lenguaje Natural*, **47**(Septiembre): 163–170.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., primera edición. New York, NY, USA.

Molina, L., Belanche, L., y Nebot, A. (2002). Feature selection algorithms: a survey and experimental evaluation. En: *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. pp. 306–313.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., y Ureña López, L. A. (2012). Detección de la polaridad en citas periodísticas: una solución no supervisada. *Procesamiento del Lenguaje Natural*, **49**(Septiembre): 149–156.

Montejo-Ráez, A., Díaz-Galiano, M. C., Perea-Ortega, J. M., y Ureña López, L. A. (2013). Spanish knowledge base generation for polarity classification from masses. En: L. Carr, A. H. F. Laender, B. F. Lóscio, I. King, M. Fontoura, D. Vrandecic, L. Aroyo, J. P. M. de Oliveira, F. Lima, y E. Wilde (eds.), *WWW (Companion Volume)*. International World Wide Web Conferences Steering Committee / ACM, pp. 571–578.

Moreno, A., Pérez, A., y Torres, S. (2010). Sentitext: Sistema de análisis de sentimiento para el español. *Procesamiento del Lenguaje Natural*, **45**: 297–298.

Mudinas, A., Zhang, D., y Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12*, pp. 1–8.

Myszkowski, P. y Bicz, A. (2010). Evolutionary algorithm in forex trade strategy generation. En: *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, Oct. pp. 81–88.

O'Keefe, T. y Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. En: *Proceedings of the fourteenth Australasian Document Computing Symposium*, Sidney, Australia. pp. 67–81.

Pak, A. y Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. En: *Proceedings of LREC 2010*, Paris, Francia. pp. 1320–1326.

Pang, B., Lee, L., y Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP*, pp. 79–86.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**: 2825–2830.

Qiao, X., Li, H., Xiang, J., y Deng, H. (2011). The study of images emotion based on fmri. En: *Proceedings of the 2011 International Conference on Web Information Systems and Mining - Volume Part I*, Berlin, Heidelberg. Springer-Verlag, WISM'11, pp. 66–72.

R-Moreno, M. D., Cuesta, A., y Barrero, D. F. (2013). Twitter stream analysis in Spanish. En: *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13*, New York, New York, USA. ACM Press, p. 1.

Rosá, A. (2011). *Identificación de Opiniones de diferentes fuentes en Textos en Español*. Tesis doctoral, Instituto de Computación de la Universidad de la República de Montevideo, Uruguay.

Rosá, A., Wonsever, D., y Minel, J.-L. (2010). Opinion identification in spanish texts. En: *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, Los Angeles, California. Association for Computational Linguistics, Association for Computational Linguistics, p. 54–61.

Smith, P. y Lee, M. (2012). Cross-discourse development of supervised sentiment analysis in the clinical domain. En: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, Stroudsburg, PA, USA. Association for Computational Linguistics, WASSA '12, pp. 79–83.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., y Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, **37**(2): 267–307.

Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. En: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, July, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics, pp. 417–424.

Vilares, D., Alonso, M. A., y Gómez-Rodríguez, C. (2013). Supervised polarity classification of spanish tweets based on linguistic knowledge. En: S. Marinai y K. Marriott (eds.), *ACM Symposium on Document Engineering*. ACM, pp. 169–172.

Waila, P., Singh, V. K., y Singh, M. K. (2012). Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews. *2012 IEEE International Conference on Computational Intelligence and Computing Research*, (c): 1–6.

Westerski, A. (2007). Sentiment analysis: Introduction and the state of the art overview. *Universidad Politecnica de Madrid, España*, pp. 211–218.

Wiebe, J. y Cardie, C. (2005). Annotating expressions of opinions and emotions in language. language resources and evaluation. En: *Language Resources and Evaluation (formerly Computers and the Humanities)*. Springer, Vol. 39, pp. 165–210.

Wiener, E., Pedersen, J. O., y Weigend, A. S. (1995). A neural network approach to topic spotting. En: *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*. pp. 317–332.

Xu, J.-M., Zhu, X., y Bellmore, A. (2012). Fast learning for sentiment analysis on bullying. En: *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12*, New York, New York, USA. ACM Press, pp. 1–6.

Yang, Y. (1995). Noise reduction in a statistical approach to text categorization. En: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA. ACM, SIGIR '95, pp. 256–263.

Yang, Y. y Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. En: *Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., ICML '97, pp. 412–420.

Yang, Y. y Wilbur, W. J. (1996). Using corpus statistics to remove redundant words in text categorization. *JASIS*, **47**(5): 357–369.

Yu, L., Ma, J., Tsuchiya, S., y Ren, F. (2008). Opinion mining: A study on semantic orientation analysis for online document. En: *Proceedings of the 7th World Congress on Intelligent Control and Automation*, Chongqing, China. IEEE, pp. 4548–4552.

Apéndice A. Apéndice

A continuación se presenta una lista de las características de un individuo obtenido por el AG, con las cuales se logró obtener una precisión de 91 % (TP=98; TN=85; FN=15; FP=2) en el SVM. En total son 535 características, siendo 342 unigramas, 71 lemas, 9 pos-tags y 113 posbigramas.

A.1. Unigramas

guapas	infumable	narrando	toro
año	nirvana	después_de	raimi
rostro	pierdas	vive	solitario
experiencia	calma	ellos	perdimos
fantástica	dicaprio	tampoco	iwo_jima
todas	casi	mundo	excelente
media	consciente	mundos	travelling
felicidad	intragable	desatado	inteligencia
olvidable	pasión_de_cristo	inocencia	cerca
honesto	borde	mediocre	máquinas
oscarizada	rollito	comunicación	ritmo
tradicional	disfruta	relatando	basta
extraordinario	disfrute	vinci	él
memorables	conciencia	manera	heroe
ambos	relaciones	zodiac	michael_caine
james_bond	seguros	formidable	emocionado
pesimismo	persona	propio	vacuo
falla	clint_eastwood	clavado	podemos
diálogos	ojos	empeña	dura
inquisición	robert_rodríguez	ivana_baquero	articula
desaprovechadas	aviador	serenidad	silencios
ayudar	greg_kinnear	duramente	excepcional
salidos	mantiene	ciudadanos	punto

gozar	intentar	tonto	cutre
altamente	destacar	a_la_perfección	hacia
molestado	mezcla	entorno	ganando
muerte	espartanos	valerie_faris	madre
última	escasamente	maestría	poco
desastrosa	gran	cuidada	realidad
sensible	obsesiones	fallido	trámite
gesto	venom	sobrado	significado
inquietudes	espartano	fallida	ratas
encima_de	apariencia	literaria	quedamos
odio	sobrenaturales	utilización	historias
lleva	manidas	explotar	necesario
lleve	argumento	siendo	pellejo
presos	sueño	llegar	hubiera
magistral	comprender	death_proof	restante
robert_graysmith	lector	amistad	bodrio
escuchar	resulta	marvel	energía
relato	legua	cara_a_cara	escarpias
samuel_l._jackson	jack_nicholson	pesadilla	aceptable
punzante	realista	poética	descanso
incuestionable	hombres	rienda	retrata
aburrí	aguanto	llena	aumento
unas	llevada_a_cabo	cultural	inverosímil
mala	gibson	impostado	desmesurado
hipnótica	demuestra	greengrass	los_ángeles
aburre	nicholson	de_por_vida	telefilm
eduardo_noriega	churros	scorsese	christopher_nolan
positivo	complicar	culpa	movimiento
acecha	perturbador	chapeau	interés
elegantes	perfectamente	bochornoso	carácter
hannibal_lecter	lanzan	personal	uso

significación
plana
humanidad
vamos
a_el_igual_que
comencé
tragedia
violencia
saco
elena_anaya
caballo
injusticia
dedo
lucidez
entretener
cotidianas
íntima
angustia
síndrome
siempre
cutrez
maquis
inteligente
simplona
viaje
estupendo
estupenda
crimen
democracia
palabras
negar
divertida
misa
perfección
colas
ajustado
saga
coger
mediocridad
penoso
oportuna
su
castaña
toque
acercan
odia
años
poético
imágenes
hacen
barato
ken_watanabe
carmen_maura
música
a_quemarropa
visión
antes
pobre
acertado
monica_bellucci
miguel_rellán
insípida
recupere
minucioso
bello
obligando
sumerge
puesta
también
otro
sosa
otorgar
estado
gracia
superproducción
entretenido
proceso
desaprovechados
universal
a_la_vez
perfectas
incomprendida
inverosimilitud
asombrosa
ay
e
ah
firme
2
redondo
sam_raimi
-
trágica
solo
brad_pitt
film
¿
cruda
acusado
spider-man
jerjes
terapeuta
literalmente
británica
suribachi
ví
enorme
preciso
hannibal
mejorcito
simpáticas
madurez
eragon
perfecta
sobretodo
personaje
bla
prescindible
josé_luis_alcaine
a_la_vez_que
llevará
pueblo
junto_a
aunque
4
prota

A.2. Lemas

bache	crudo	conmovedor	debilidad
acertar	conmocionar	sugerente	divertido
prisionero	interpretar	contener	persa
colmar	lleno	dictar	gracioso
vivir	cesar	cuidado	acompañar
impregnar	mantener	desconocer	técnico
circunstancia	surgir	directo	dudar
demostrar	alto	manejar	discusión
tedioso	artificio	susto	desvanecer
estafar	maltratar	extraordinario	aburrir
adquirir	hombre	impostar	paja
estupendo	imaginativo	controlador	apunte
sentir	salvar	reconocer	flojo
desnudar	desatar	miedo	merecido
reforzar	maqui	esconder	complicar
superar	descarnar	malo	imponer
acérrimo	pintar	tardo	contemplación
palabra	impresionante	descerebrado	

A.3. Pos-tag

I	(interjección)
NCFS00D	(nombre común femenino singular)
PP1CSO00	(pronombre personal primera persona común singular oblicuo)
AQDFS0	(adjetivo calificativo femenino singular)
PP3NS000	(pronombre personal tercera persona singular)
FD	(signo de puntuación dos puntos)
PX3MP0C0	(pronombre posesivo tercera persona masculino plural)
VMSP1S0	(verbo principal subjuntivo presente primera persona singular)
PI0MP000	(pronombre indefinido masculino plural)

A.4. Posbigramas

DI0FS0_DP3CS0	NCFS000_VAIC1S0	AQ0CP0_FP
NCMP000_DD0FS0	PP1CS000_CS	PI0MP000_RG
DI0MP0_DA0MP0	DP1MPP_NCMP000	AQDMP0_FC
NCFP000_FC	DD0FS0_NCFS000	RN_VSSP1S0
FC_VMG0000	DI0MS0_Z	DI0MS0_VMIP1S0
CC_DA0MP0	VMP00PM_SPS00	NP00000_NC00000
PP3CSD00_VMII3P0	NP00000_AQ0CP0	DD0FP0_Z
VSIP3S0_FC	AQ0CP0_FPT	NCFS000_FH
NP00000_VSIP3S0	AQ0MS0_VMP00SM	NC00000_NP00000
RG_AQ0CS0	CS_VMII2S0	FC_DI0MS0
FC_VSII3P0	FPA_VMP00SM	VMIS3S0_SPS00
FC_VMP00SF	NCMS000_DI0MS0	PR0CN000_PP1CP000
NCMP000_RG	NCCS000_NP00000	AQ0FP0_NCFP000
PR0CN000_FC	DA0FS0_SPS00	AQ0MS0_NP00000
DA0MP0_PX1MP0P0	AQ0CP0_CC	FE_Z
NCMS000_AQSMS0	NCFP000_CC	DA0FP0_AO0FP0
VMG0000_PP3CSD00	AQ0CS0_CC	VMN0000_PP1CP000
VMP00SM_PP3CSD00	DA0FS0_AQSFS0	VMP00PF_SPS00
DA0MS0_AO0MS0	NP00000_VMIS3S0	DI0MS0_NCCS000
FP_VSIP3S0	CC_AQ0MP0	AQ0MP0_SPS00
CC_PP1CP000	NCMS000_DD0MS0	FC_VMSI1S0
PI0FS000_FIT	VMP00SM_PI0CS000	PR0CN000_NP00000
PR0CN000_DI0MP0	PI0MS000_PP3NS000	VAIP3P0_CC
VMIP3S0_NCFS000	VMN0000_DD0MP0	VMIS3P0_VMG0000
FC_DI0FS0	PR0CS000_VMIP3P0	FC_VMP00SM
NCFS000_RG	VASP1S0_FC	VMIF1P0_VMN0000
PP3NS000_SPS00	DP3CP0_AQ0FP0	P0000000_VASI1S0

NP00000_SPS00	PP1CSO00_FC	VMG0000_DI0CS0
NP00000_PP3CNA00	CC_VMP00SM	FC_DD0MS0
NCFS000_VMII1S0	DP1FSP_NCFS000	DI0FS0_NCFS00D
VMP00PF_RG	AQ0FP0_CC	PT0CN000_VMSI2P0
FIT_I	FG_SPS00	CC_AQ0FS0
FG_FC	VMIP1S0_FIT	NCFP000_FP
DI0MS0_AQSMS0	PI0MS000_VMSP1S0	AQ0FS0_FIT
DP3CP0_NCMP000	PR0CN000_VAIP1P0	FC_AQ0FP0
SPS00_DA0FP0	VMIS3S0_DI0MS0	VSN0000_AQ0MS0
PP1CP000_VMIP3S0	AQ0MS0_VMIP3S0	RN_VMIP3S0
SPS00_DI0MP0	NCFS000_PR0CN000	

A continuación se presenta un documento del conjunto de los documentos desconocidos para el cual este modelo de clasificación y con estas características se obtuvo como verdadero positivo, otro documento como falso positivo, uno como verdadero negativo y por último uno falso negativo. Se remarcan los unigramas que han sido tomados en cuenta para la clasificación.

Verdadero positivo. Funny Games es un estilete, una navaja de hoja estrecha y aguda esperando la oportunidad de clavarse con saña entre tus costillas para desangrarte lentamente. **Siempre** viene bien una mirada retrospectiva para revisar la obra de uno de los directores más comprometidos política y socialmente del panorama europeo, Michael Haneke, y gracias a Dios, igualmente comprometido con traspasar las fronteras del lenguaje cinematográfico desde una perspectiva completamente autoral. La mirada es, si cabe, doblemente acertada, puesto que está a **punto** de estrenarse en Estados Unidos un remake de Funny Games (sí, Michael Haneke versionándose a sí mismo, suponemos/esperemos que lo hace para financiar transgresores proyectos venideros de los que nos tiene acostumbrados) protagonizada por la prolífica Naomi Watts y Michael Pitt. Si Caché, la **última** película de Haneke, plantea una mirada incisiva a la herida que provocó el enfrentamiento entre la comunidad argelina en Francia y los propios parisinos en los **años** 60, conflicto que tuvo como resultado la masacre de más de 400 argelinos (muchos de los restos quizás todavía yacen en el fondo del Sena),

Funny Games pone **su** acento en la **violencia** descarnada, la **violencia** por la **violencia**, desde un **punto** de vista menos político, **aunque** no pueda evitar sacar a la luz los fantasmas que aparecen en **casi todas** sus películas (la banalización de la **violencia**, la hibernación de la culpa en la sociedad occidental, la desacralización de la brutalidad por parte de los medios de **comunicación**, la lucha de claseshellip;) Funny Games es la historia de una familia de clase alta que se dispone a pasar **unas** vacaciones tranquilas en una lujosa villa costera, pero sus planes se ven truncados por la aparición de una perversa pareja de jóvenes, que como los dragos de la Naranja Mecánica (y con una estética curiosamente similar) se inmiscuyen sibilinamente en sus vidas, en su casa, para inyectar sangre y terror en sus cabezas y jugar **literalmente** con sus vidas. Cuando parecen encontrar una vía de salvación, el perverso juego traspasa la pantalla, se hace global, meta-cinematográfico, y nos damos cuenta de que no hay salida alguna. Es el triunfo del **pesimismo**, del trágico final, el triunfo del juego perverso y la disolución de las fronteras entre **realidad** y ficción. La película incluso realiza un fast rewind para anular todo atisbo de optimismo, y los verdugos dialogan con el espectador, forzándole a posicionarse de un bando u **otro**. La sociedad opulenta genera **violencia** y ésta no tiene la menor intención de diluir **su** efecto y ataca con toda **su** virulencia, y además lo hace porque sí. Haneke nos hace reflexionar sobre la inconveniencia de insensibilizarse **antes** las **imágenes** que nos muestran el dolor. Las películas de Haneke, y ésta en especial, enraízan con la corriente slasher del cine francés contemporáneo, y en cierto grado explica el éxito de crítica de una película tan **dura** y ultraviolencia como A l’interieur que, con un planteamiento parecido, se alzó con algunos premios en este Sitges 2007. Cuando veáis como un hombre mata a **otro** hombre a través de una pantalla de televisión y no sintáis prácticamente nada, no llaméis a un psicólogo, Michael Haneke y **su** Funny Games os darán el diagnóstico.

Falso positivo. **Antes** de comenzar a destripar la cuarta entrega de Saw IV conviene aclarar que desde esta tribuna duele criticar una **saga** de este calibre. Las **sagas** de terror kilométricas y carentes de ningún tipo de pretensión son, como sabéis, muy apreciadas por aquí, y las andanzas de Jigsaw (ya convertidas en una parodia de sí mismas) parece confirmado que llegarán al menos hasta **su** sexta entrega.Las películas de Saw cumplen muchos de los requisitos para convertirse en una franquicia mítica al más puro estilo Viernes 13 o **Pesadilla**

en Elm Street: Subrayadísimas características generales, como la generosidad a la hora de mostrar sangre y vísceras (rozando el gore en la tercera entrega y seriamente atenuada en esta **última**) la sucesión de entregas sin **casi** tiempo para el respiro, la explotación de los moderadamente sorprendentes aciertos de la primera parte y la ineludible falta de coherencia y pretensiones, son mostradas en cada uno de los filmes sin ningún tipo de trampa ni cartón. Saw no engaña a nadie (al menos ya no) y el espectador que acude a la sala debe estar preparado para lo que se le va a ofrecer.Sin embargo, a pesar de mi aprecio por la apuesta casquera que representan estas cintas, no puedo evitar sentirme decepcionado por **su** cuarta entrega. Si en Saw III se había relegado **casi** totalmente el espíritu más cercano al thriller que se pudiera intuir levemente en las anteriores entregas (sobre todo en la primera) situando **casi** el cien por cien de la acción en una sucesión de las más macabras **e** hilarantes torturas, Saw IV reconduce esos cauces **poco** más o menos autoparódicos **hacia** un estilo indefinible.La película de nuevo intenta hacerse fuerte en la sorpresa final, con más bríos que nunca de hecho, convirtiendo el guión en un galimatías imposible de seguir por quien no haya visto las anteriores entregas. Da la sensación de que Darren Lynn Bousman intenta recuperar la seriedad y **otorgar** a la serie de filmes un espíritu de obra geométrica, pensada y diseñada como un todo desde el principio. Esta pretensión, si es que la **hubiera**, cae por su **propio** peso con un guión que se sabe, si cabe, más disparatado que nunca: El grueso de la trama se fundamenta en un policía que sigue los tenebrosos planes de Puzzle, paso a paso y sin saltarse una coma, de forma total y absolutamente **inverosímil**, y, como en anteriores ocasiones, la narración salta entre **historias** paralelas centrando **su** atención, en segunda instancia, en el delirante **e** injustificado (al menos no correctamente) origen del psicópata.De nuevo un final sorprendente **e** inesperado adorna y acelera el cierre de la peli con dos sorpresas esta vez, una bastante interesante relacionada con la línea temporal y la otra, resobada y encima más estúpida que nunca, concerniente a los personajes. Todo ello aderezado con la ya famosa y eficaz melodía marca de la casa.En el apartado visual la peli pierde muchos enteros respecto a sus predecesoras. Las trampas ideadas por Jigsaw son soslayadas en esta ocasión; prácticamente no se le da importancia a las torturas y la mayoría de los juegos no están demasiado concretados y son muy confusos. Exceptuando una o dos bromas macabras, en el resto cuesta saber el procedimiento que debe adoptar la víctima, y encima todo rodado con una cámara loca que se interpone ante la claridad de las situaciones. Si no fuera por

la autopsia inicial, **punto** fuerte de la peli, la decepción sería irreconciliable.En definitiva, incluso viendo Saw IV con una mirada compasiva, defensor como me considero de este tipo de propuestas, no acabé de seguirle el hilo al director. Si bien no albergaba esperanzas de que la trama me enganchara, sí al menos esperaba encontrar una buena ración de sofisticadas torturas dispuestas para ser degustadas. Desgraciadamente, esta vez ni una ni otra.

Verdadero negativo. Ocasionalmente intento llevarme al cine a un buen crítico que conozco, es mi primo de nueve **años** y puede ser la mar de útil para algunas películas que me resultan inclasificables o ciertamente frustrantes. Hace muchos **años** las productoras se dieron cuenta de que ver una película en el **entorno** de un cine plagado de personas que disfrutan de la cinta puede afectar a la **visión** del crítico y hacerle sentirse un **poco** más generoso a la hora de escribir sobre ella. Por eso invitar a pases de prensa en cines con niños en el caso de películas infantiles fue una práctica común que consiguió bastante buen resultado. No hay corazón que resista decir que una película infantil es **mala** cuando cientos de críos han **estado** disfrutando y arrojándose palomitas unos a otros a tu alrededor porque a pesar de que el **lector** de tu crítica sea **su** padre no puedes obviar el hecho de que el niño se lo haya pasado bien ya que, para bien o para mal, era el público par el que se creó la película y que la consumirá en las salas, DVDs y merchandising vario. La variante que existe ahora es que tras la irrupción del cine fantástico hay un nuevo género de aventuras para niños y adolescentes que está repuntando de forma alarmante con protagonistas infantiles y aventuras épicas en **mundos** paralelos como Las Crónicas de Narnia, **Eragon**, o Harry Potter. El público de estas es más amplio y debe estudiarse con bastante cuidado ya que a pesar de que a los infantes de hasta quince **años** disfrutan con estos divertimentos de sobremesa el posible espectador adulto que ha disfrutado con cintas como Los Gonnies o El Chip Prodigioso quiere revivir la década de los ochenta y acaba topándose con un simple despliegue de efectos Especiales aderezando el último éxito de la literatura juvenil. Los Seis Signos de la Luz narra la historia de Will, un adolescente americano que se acaba de mudar con **su gran** familia a Inglaterra y que descubrirá que es el elegido para llevar a cabo una especial misión de encontrar a través del tiempo los Seis Signos de la Luz que ayudarán a luchar contra las Tinieblas que quieren apoderarse del **mundo**. En **su** camino **hacia** los signos Will contará con la ayuda de los Ancestrales, personas atemporales que ayudan a los buscadores para descubrir sus poderes

y emplearlos a tiempo de salvar el **mundo**. Soy una **gran** seguidora del cine fantástico de acción y aventuras. Crecí con aventureros con sombrero y látigo y Chicos de Oro. No puedo remediar cada vez que un adolescente se mete en un **mundo** mágico recordar lecturas infantiles como La Historia Interminable y todo este bagaje es lo que hace que cintas como esta me resulten algo aburridas y flojas. De factura impecable pero descuidada en el cuidado del público adulto consigue dejar poca huella y un recuerdo entre nostálgico y pretencioso **después de su** visionado. Totalmente **prescindible** si no vas con un menor. Eso si, garantizo que si tienes menos de 14 **años** o tu acompañante está en ese rango de edad es una **puesta** segura que disfrutará mucho.

Falso negativo. De un tiempo a esta parte me he dado cuenta de que la relación existente entre mis deseos de ver una película y la impresión que esta acaba dejando en mí es inversamente proporcional. Es decir, cuando me muero por ver algo suele resultar ser un **bodrio** de mucho cuidado. Por eso y ante las expectativas suscitadas por la promoción, crítica y acogida en Estados Unidos de Stardust me he vuelto muy cauta y he intentado ocultar mi excitación para engañar a la estadística.Parece ser que ha resultado bien ya que está fábula **fantástica** ha cumplido mis más altas esperanzas. Suele ser habitual que cuando algún cineasta decide esconder **su** propuesta tras la etiqueta de “Fábula o cuentordquo; hay dos resultados posibles: un petardo aburrido e infantiloide anclado en los FX más modernos (como Una Serie de desafortunadas desdichas o Las Crónicas de Narnia) o, al contrario una aventura mágica y entretenida (como Big Fish, Chocolat o Charley y la Fábrica de Chocolate por citar unos cuantos). Yo colocaría al objeto de esta crítica en el segundo de los grupos.Tristan (Cox) quiere conquistar el amor de Victoria (Miller) por eso le promete que le traerá como obsequio una estrella caída del cielo. Para ello se debe adentrar en un mundo mágico donde encontrará personajes entrañables como el pirata Shakespeare (DeNiro) un americano enamorado de Inglaterra y que esconde un oscuro secreto, o brujas terribles (Pfeiffer) que conscientes del poder mágico de la estrella deciden perseguirla y hacerse con ella sin contar con los hijos de un rey de un lejano planeta que están destinados a acabar los unos con los otros para poder heredar (y viajan junto con sus fantasmas a la búsqueda del medallón que les otorgará el poder). Cuando nuestro joven galán encuentra **su** estrella no es exactamente lo que esperaba ya que al caer en la zona mágica se ha convertido en una damisela llamada

Yvaine (Danes) que puede darle tantos dolores de cabeza como quebraderos de cabeza.Con un **ritmo** trepidante la historia de aventuras es capaz de enlazar a todos sus personajes en una entretenida cinta que **mezcla** comedia, acción y romanticismo muy al estilo de La Princesa Prometida. Si además se cuenta con una clasificación para todos los públicos y el mejor elenco de actores secundarios que hayamos encontrado en la **gran** pantalla en muchos **años** la película de las navidades estaba clara pero la estrenan en octubre, veremos que pasa.