UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Detección de perfiles políticos en Twitter

Trabajo Final de Máster

Máster en Big Data Analytics

Autor: José Alberto Pérez Melián

Tutor: Francisco Manuel Rangel Pardo

Curso 2015-2016

Canarias son siete islas
arrulladas por el mar,
siete corazones guanches,
siete notas de un cantar.

Una guitarra en la mano
y en el aire una folía;
no hay canto como mi canto
ni tierra como la mía.

Popular

Agradecimientos

Resumen

Resumen en lengua española

Palabras clave: Palabra clave español 1, palabra clave español 2

Abstract

Resumen en lengua inglesa

Key words: Keyword english 1, keyword english 2

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

Las redes sociales como Facebook, Twitter o Instagram se pueden definir como servicios que proveen a los usuarios la posibilidad de establecer conexiones con sus amigos a través de una aplicación y compartir información con ellos. Son, con diferencia, las aplicaciones más populares de lo que se conoce como Web 2.0, término que comprende «aquellos sitios web que facilitan el compartir información, la interoperabilidad, el diseño centrado en el usuario y la colaboración en la World Wide Web»[1]. Se cuentan por millones los usuarios que las utilizan para estar en contacto con amigos, para conocer gente nueva o para hacer vínculos profesionales.

El crecimiento de las redes sociales por su uso e impacto en la sociedad ha crecido de forma exponencial en los últimos años al permitir «evaluar» la opinión de la gente en áreas muy diversas. Hoy en día es normal leer noticias en los periódicos donde se hacen eco de la repercusión social en Twitter de un suceso, como la aprobación o derogación de una ley, un evento deportivo o musical o la opinión sobre un nuevo producto o servicio. A la vez que aumenta la cantidad de usuarios de estas redes y los servicios que estas prestan, aumenta también el interés de muchos sectores (política, comercial, cultural...) debido al incremento de los datos generados por los usuarios que, tratados y analizados, pueden ser usados para la extracción de conocimiento mediante tareas de minería de opiniones (*opinion mining*) y de análisis de sentimientos (*sentiment analysis*).

La mayoría de los mensajes enviados por estas redes contienen información personal de los usuarios. Sin embargo, su uso y finalidad está cambiando y en los últimos años se ha visto un incremento en la cantidad de mensajes que contienen algún contenido político. El auge de estos mensajes ha permitido a los partidos políticos, a los candidatos y a los ciudadanos interactuar en tiempo real sobre cuestiones de actualidad. La importancia de las redes sociales en el ámbito político pudo verse claramente en la campaña de Barack Obama para las Elecciones Presidenciales de los Estados Unidos de América del año 2008, donde la presencia en redes sociales como Twitter, Facebook o MySpace fue parte fundamental de la campaña del candidato del Partido Demócrata. Algunos analistas atribuyen su victoria a la extensa presencia en estas redes, marcando así un hito que superó con creces el avance que supuso la aparición de las primeras páginas web de los candidatos a la Presidencia en el año 1996, la aparición del correo electrónico

[1] Wikipedia. Accedido el 25 de junio de 2016. [Enlace]

(1998), la recaudación online (2000) o la aparición de los primeros blogs (2004), estableciendo así redes como Twitter como un medio legítimo de comunicación en el terreno político desde el año 2008.

De todas las redes sociales existentes en este trabajo nos centraremos en Twitter, una red social de *microblogging* lanzada en el año 2006 cuya misión es «ofrecer a todo el mundo la posibilidad de generar y compartir ideas e información al instante y sin obstáculos»[2]. Cuenta con alrededor de 310 millones de usuarios activos mensuales y en el año 2012[3] se generaban en ella un total de 340 millones de *tweets* diarios[4]. Los mensajes o *tweets* no son más que pequeños mensajes de texto limitados a 140 caracteres de longitud en los que los usuarios pueden insertar fotos, vídeos, hipervínculos y emoticonos. Las siglas `RT` hacen referencia a los retweets, que son mensajes escritos por un usuario que otro comparte y propaga a su red de seguidores, dando más difusión al tweet original. También se ofrece la posibilidad de mencionar en el tweet a otros usuarios, añadiendo el símbolo @ seguido del nombre de la cuenta del usuario, y de poner etiquetas o *hashtags*, precedidas por el símbolo #, permitiendo así clasificar el tweet en una o varias temáticas.

El contenido y la estructura de la discusión que se lleva a cabo en Twitter presenta una oportunidad única para estudiar el rol que desempeña la red social como plataforma generadora de datos sobre los que realizar análisis posteriores.

El objetivo de este trabajo es lograr determinar el perfil político de los usuarios en Twitter. Para ello se presenta un corpus compuesto por hasta 1000 tweets de los diputados y senadores electos de la XII Legislatura de las Cortes Generales Españolas que tienen presencia en Twitter etiquetados según el partido político al que pertenecen. Este corpus se evalúa mediante distintas técnicas basadas en n-gramas, utilizadas para este tipo de tareas. Se usa, además, las Support Vector Machines (SVM) como método de clasificación para detectar el partido político de las cuentas de Twitter recuperadas.

La memoria se encuentra estructurada tal como sigue. En el Capítulo 2 se presenta la metodología empleada para la construcción del dataset que será objeto de estudio. Seguidamente, en el Capítulo 3 se presentarán las herramientas y técnicas por las que se evaluará el conjunto de datos para finalmente, en el Capítulo 4, presentar los resultados experimentales del estudio. Por último, en el Capítulo 5, se presentan las conclusiones y se esbozan las líneas de trabajo futuras.

[2] About Twitter. Accedido el 25 de junio de 2016. [Enlace]
[3] No se han encontrado datos oficiales más recientes
[4] Twitter Blog, 12 de marzo de 2012. Accedido el 25 de junio de 2016. [Enlace]

CAPÍTULO 2
Metodología para la construcción del dataset

En este capítulo se detalla la metodología que se ha seguido para la construcción de un dataset con mensajes escritos en Twitter por los diputados y senadores de la XII legislatura de las Cortes Generales Españolas. El proceso de construcción del dataset puede verse en más detalle en la Figura XX. En primer lugar, para la obtención de las cuentas de Twitter de los diputados y senadores se hace uso de una araña web que rastrea sus perfiles en las páginas institucionales del Congreso de los Diputados y del Senado de España, obteniendo para cada diputado o senador su nombre de usuario de Twitter y el grupo parlamentario al que se encuentra adscrito. Una vez obtenida esta información se hace una revisión manual en la que se comprueban los diputados o senadores que no disponen de cuenta de Twitter en las páginas institucionales y se añaden sus respectivas cuentas si disponen de ellas. A continuación se realiza un refinamiento del corpus, donde se identifica a cada diputado o senador según el partido político (y no grupo parlamentario) al que pertenece. Por último se procede a recuperar los timelines de los usuarios obtenidos en las etapas anteriores, descargando los últimos 1000 tweets escritos por cada diputado o senador.

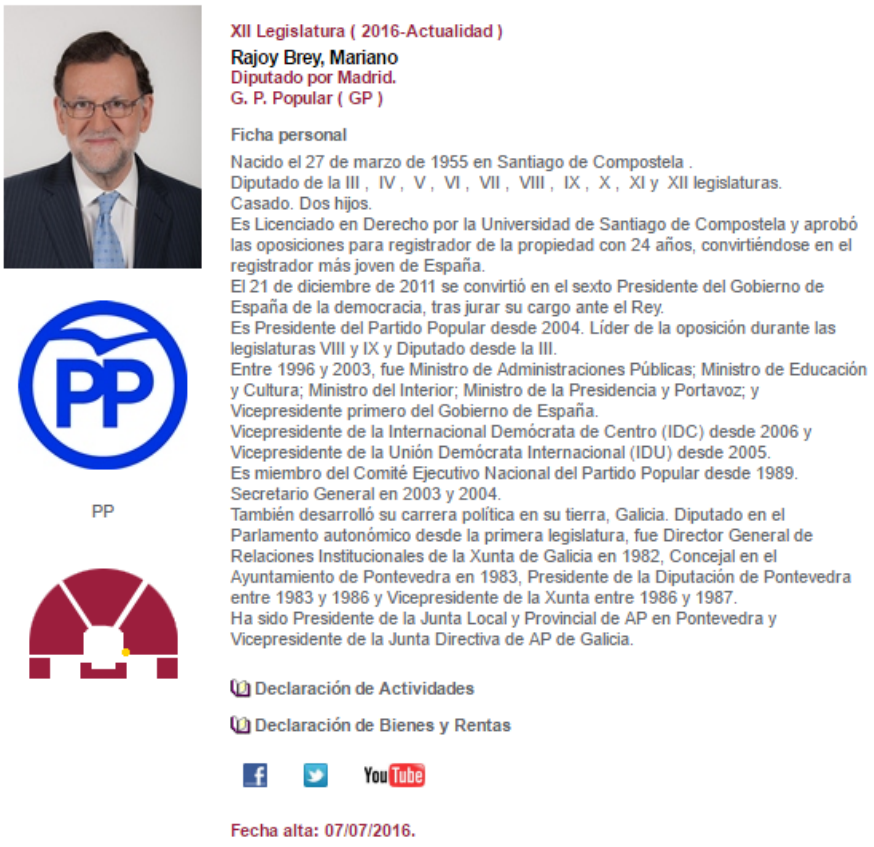
2.1 Araña web

Para obtener las cuentas de usuario de Twitter de los diputados y senadores de la XII legislatura de las Cortes Generales Españolas se ha desarrollado una araña web con la herramienta Scrapy[1], un software de código abierto implementado en Python para extraer los datos que se quieran de la web de una forma rápida y sencilla.

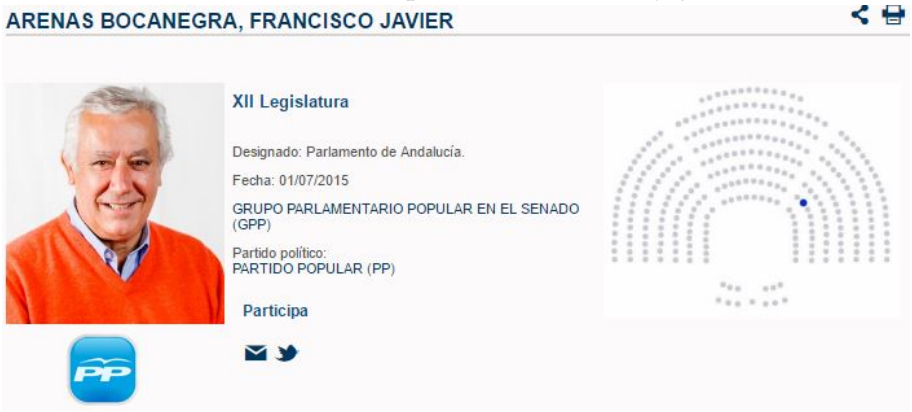
La araña web se encarga de extraer los datos de la página web del Congreso de los Diputados y del Senado de España, donde se encuentran los perfiles personales de los miembros que integran ambas cámaras junto a sus datos personales. En la Figura 2.1 puede verse el perfil del diputado Mariano Rajoy Brey (Congreso) y de Francisco Javier Arenas Bocanegra (Senado). En ellos se tiene información como el nombre completo, su foto, el grupo parlamentario al que pertenece o la

[1] https://scrapy.org

provincia por la que ha resultado electo, entre otras. En la parte inferior se ve también información relativa a las redes sociales; en este caso, Facebook, Twitter y YouTube. El objetivo de la araña web es extraer, si existe, la información relacionada con Twitter junto al nombre y apellidos del diputado o senador, el grupo parlamentario al que está adscrito y el nombre de usuario de Twitter del mismo.



(a) Perfil del diputado Mariano Rajoy



(b) Perfil del senador Javier Arenas

Figura 2.1: Ejemplo de perfiles de las páginas web del Congreso y del Senado de España

Las Tablas 2.1 y 2.2 muestran respectivamente para el Congreso y el Senado el número de diputados y senadores que disponen o no de cuenta en su perfil personal de ambas cámaras agrupados por los grupos parlamentarios a los que están adscritos. Los datos muestran que aproximadamente el 82 % de los dipu-

tados tienen cuenta en Twitter frente al 30 % de senadores que también disponen de cuenta en la red social.

Congreso de los Diputados			
Grupo	**Con Twitter**	**Sin Twitter**	**Total**
GCUP-EC-EM	62	5	67
GC	32	0	32
GER	8	1	9
GMx	14	5	19
GP	94	40	134
GS	71	13	84
GV (EAJ-PNV)	4	1	5
Total	285	65	350

Tabla 2.1: Número de diputados por grupo parlamentario que disponen o no de una cuenta de Twitter

Senado de España			
Grupo	**Con Twitter**	**Sin Twitter**	**Total**
GPER	2	10	13
GPMX	6	10	16
GPP	52	96	148
GPPOD	8	13	21
GPS	11	61	62
GPV	0	6	6
Total	79	186	265

Tabla 2.2: Número de senadores por grupo parlamentario que disponen o no de una cuenta de Twitter

2.2 Revisión manual

Los diputados o senadores no tienen la obligación de añadir a su perfil las redes sociales en las que participan, por lo que los datos obtenidos en el punto 2.1. deben ser revisados para buscar aquellos que no han proporcionado su usuario, bien porque no tienen cuenta en Twitter o porque no la han especificado. Por ello es necesario hacer una revisión manual y comprobar si hay algún caso que cumpla alguna de estas condiciones.

Tras comprobar si los diputados o senadores sin cuenta en las páginas oficiales de las cámaras disponen de cuentas en Twitter se realiza una anotación manual cuyos resultados pueden verse en las Tablas 2.3 y 2.4.

Grupo	Diputados con Twitter
GCUP-EC-EM	63 (+1)
GC	32 (=)
GER	8 (=)
GMx	18 (+4)
GP	98 (+4)
GS	73 (+2)
GV (EAJ-PNV)	4 (=)
Total	296 (+11)

Tabla 2.3: Número de diputados por grupo parlamentario que disponen o no de una cuenta de Twitter tras la revisión manual

Grupo	Senadores con Twitter
GPER	7 (+5)
PMX	10 (+4)
GPP	74 (+22)
GPPOD	13 (+5)
GPS	27 (+16)
GPV	1 (+1)
Total	132 (+53)

Tabla 2.4: Número de senadores por grupo parlamentario que disponen o no de una cuenta en Twitter tras la revisión manual

2.3 Refinamiento del corpus

Llegados a este punto se tiene, para cada diputado y senador, el grupo parlamentario al que se encuentra adscrito. Se procede ahora a sustituir el grupo parlamentario por el partido político o coalición electoral a la que pertenece. En el Congreso de los Diputados en la XII legislatura hay 7 grupos parlamentarios, mientras que en el Senado hay 6. Las Tablas 2.5 e 2.6 muestran las siglas de los grupos parlamentarios de ambas cámaras junto su nombre, además del partido político al que encuentran adscritos los diputados y senadores.

Grupo	Nombre	Partido político
GCUP-EC-EM	Grupo de Candidaturas de Unidad Popular - En Comú - En Marea	PODEMOS
GC	Grupo Ciudadanos	CIUDADANOS
GER	Grupo Esquerra Republicana	ERC
GMx	Grupo Mixto	CDC, COMPROMIS, BILDU, UPN, CC, FORO, NC
GP	Grupo Popular	PP
GS	Grupo Socialista	PSOE
GV (EAJ-PNV)	Grupo Vasco	PNV

Tabla 2.5: Correspondencia entre grupos parlamentarios en el Congreso y partidos políticos

Una vez se llega a este punto es necesario etiquetar a los diputados y senadores en el partido político al que están adscritos para predecir con mayor precisión la opción política en el modelo que se construirá en pasos sucesivos.

Realizado este cambio el estado del dataset queda como en la Tabla 2.7, donde se observa el número de diputados y senadores etiquetados por partido político.

Grupo	Nombre	Partido político
GER	Grupo Esquerra Republicana	ERC
GC	Grupo Popular	PP
GPPOD	Grupo Esquerra Republicana	PODEMOS
GPS	Grupo Socialista	PSOE
GP V	Grupo Popular	PNV
GMx	Grupo Mixto	CIUDADANOS, CDC, CC, COMPROMIS, NC, ASG, BILDU, FORO, UPN

Tabla 2.6: Correspondencia entre grupos parlamentarios en el Senado y partidos políticos

Partido	Diputados	Senadores	Total
PP	98	74	172
PSOE	73	27	100
PODEMOS	63	13	76
CIUDADANOS	32	2	34
ERC	8	7	15
PNV	4	1	5
COMPROMIS	3	2	5
BILDU	2	0	2
CDC	8	3	11
UPN	2	0	2
CC	1	2	3
FORO	1	0	1
NC	1	1	2
Total	296	132	428

Tabla 2.7: Número de diputados y senadores etiquetados por partido político que disponen de cuentas de Twitter

2.4 Recuperación de timelines

Una vez etiquetados todos los diputados y senadores con cuenta en Twitter el siguiente paso consiste en recuperar los últimos 1.000 tweets de los *timelines* de los diputados y senadores del punto anterior. En la Tabla 2.8 se muestran, para cada partido político, el número de tweets que han sido recuperados para ambas cámaras, así como el número de tweets por usuario y la longitud media de los tweets en palabras y en caracteres.

				Longitud media	
Partido	**Dip./Sen.**	**Tweets**	**Tweets/usuario**	**Palabras**	**Caracteres**
PP	172	135.458	787,55	17,06	126,30
PSOE	100	81.870	818,70	16,88	124,08
PODEMOS	76	68.784	905,05	17,54	129,94
CIUDADANOS	34	33.747	992,56	17,29	130,28
ERC	15	14.416	943,07	17,05	124,67
PNV	5	3.398	677,80	15,78	119,43
COMPROMIS	5	3.796	759,20	16,21	122,04
BILDU	2	1.995	997,50	16,72	129,08
CDC	11	10.992	999,27	17,60	127,13
UPN	2	1.996	998,00	16,30	119,13
CC	3	2.439	812,33	16,81	125,88
FORO	1	99	99,00	14,06	110,77
NC	2	1.672	836,00	16,75	127,46
Total	428	360,662	-	-	-
Media	32,92	27.743,23	817,39	16,62	124,32
SDev	52,32	42.117,95	240,23	0,93	5,43

Tabla 2.8: Número de diputados y senadores, tweets, proporción de tweets por persona y longitud media (en palabras y en caracteres) de los tweets tras la descarga de los timelines.

CAPÍTULO 3
Evaluación del dataset

En este capítulo se abordan las técnicas aplicadas sobre el dataset para construir un sistema que sea capaz de discernir la orientación política de un perfil en Twitter. Para ello se experimenta con distintas representaciones de documentos basadas en *n*-gramas de palabras empleando los métodos TF y TF-IDF y usando las Máquinas de Vectores de Soporte como método de clasificación para la identificación de la orientación política.

3.1 Representación de los documentos

En esta sección se detallan las técnicas empleadas para la representación de los documentos o *tweets*. Estos documentos se representan mediante vectores de características de *n*-gramas en los que se proyectan la frecuencia de los términos en los documentos o *Term Frequency* (TF) o *Term Frequency-Inverse Document Frequency* (TF-IDF).

3.1.1. Bolsas de *n*-gramas de palabras

Los *n*-gramas son muy utilizados en tareas de minería de texto o `Text Mining` y en procesamiento del lenguaje natural o Natural Language Processing (NLP). Un *n*-grama es una subsecuencia de *n* elementos de una secuencia dada. La Tabla 3.1 muestra la descomposición del texto «somos una gran nación» en unigramas ($n = 1$) bigramas ($n = 2$) y trigramas ($n = 3$) de palabras.

Texto: «somos una gran nación»

n	***n*-gramas de palabras**
1	(somos), (una), (gran), (nación)
2	(somos, una), (una, gran), (gran, nación)
3	(somos una gran), (una, gran, nación)

Tabla 3.1: Ejemplo de la extracción de *n*-gramas de palabras para $n = 1, 2, 3$.

3.1.2. Métodos de representación de palabras: TF y TF-IDF

Una vez se tiene para cada documento los *n*-gramas calculados es necesario obtener el vector de características bien con la frecuencia con la que cada término aparece en el documento o la frecuencia ponderada de aparición en el total de documentos. En el primer caso se hace referencia a la técnica *Term Frequency* (TF), en la que se contabiliza el número de apariciones de un término en un documento. En el otro caso se trata del método *Term Frequency-Inverse Document Frequency* (TF-IDF).

3.2 Algoritmos de clasificación

En esta sección de describen las Máquinas de Vectores de Soporte (SVM) como el método de clasificación implementado para abordar el problema de la identificación de perfiles políticos en Twitter.

Para implementar todos los algoritmos se hace uso de la librería `scikit-learn`[1].

3.2.1. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (en inglés, *Support Vector Machines*) son un método de clasificación que se basa en la construcción de hiperplanos para separar las muestras de las clases a predecir, de forma que el margen entre la distancia entre las muestras más próximas de cualquier clase y los hiperplanos sea la máxima, haciendo uso de métodos basados en *kernels* para operar en espacios de alta dimensionalidad.

3.2.2. Naïve Bayes

Los métodos de Bayes naïve son un conjunto de algoritmos de aprendizaje supervisados basados en la aplicación del teorema de Bayes con la suposición «ingenua» de independencia entre cada par de características.

3.2.3. Árboles de clasificación

Los árboles de clasificación sirven para resolver problemas de clasificación, con el objetivo de asignar un caso u observación a una de las categorías o clases especificadas. Un árbol consiste en un conjunto secuencial de condiciones y acciones que relacionan unos determinados factores con un resultado o decisión. Es un método de clasificación supervisada que analiza los datos proporcionados en busca de patrones y usa los resultados de este análisis para definir la secuencia y condiciones para la creación del modelo de clasificación. Un árbol de clasificación busca en cada nodo (split) maximizar la variabilidad explicada por ese nodo.

[1] scikit-learn.org

3.3 Configuración experimental

A continuación se procede a explicar el dataset empleado, el marco experimental y las medidas de evaluación para medir la calidad del modelo.

3.3.1. Corpus de evaluación

El proceso de evaluación se realiza con una muestra significativa del dataset construído en el Capítulo 2, que contiene hasta 1.000 tweets de los miembros del Congreso de los Diputados y del Senado de España de la XII Legislatura de las Cortes Generales Españolas. La Tabla 2.8 muestra más en detalle la composición del corpus.

Con el objetivo de estudiar el rendimiento de los distintos clasificadores se generan tres escenarios para los que se construyen varios modelos modelos:

- **Escenario 1**. Se tienen en cuenta los partidos PP y PSOE.
- **Escenario 2**. Se tienen en cuenta los partidos PP, PSOE, PODEMOS y CIUDADANOS.
- **Escenario 3**. Se tienen en cuenta los partidos PP, PSOE, PODEMOS, CIUDADANOS y OTROS, en los que se engloban el resto de partidos presentes del dataset.
- **Escenario 4**. Se tienen en cuenta todos los partidos presentes en el dataset.

3.3.2. Marco experimental

A continuación se explican más en detalle los pasos seguidos para la realización de los experimentos.

Paso 1 Se extrae el texto de los tweets descargados y se aplica un proceso de tokenización para la división del texto en palabras, pasándolas a minúscula.

Paso 2 Con las secuencias de palabras obtenidas se generan bolsas de n-gramas de palabras para distintos valores de n (de 1 a 5) variando el tamaño del vocabulario (500, 1.000 y 5.000 términos). Para cada configuración se generan las características aplicando los métodos TF y TF-IDF de representación de documentos.

Paso 3 Generación de los modelos mediante las técnicas de clasificación vistas en el punto 3.2.

Paso 4 Realización de un estudio ANOVA multifactorial para determinar el modelo más significativo para los escenarios propuestos.

Los experimentos se han llevado a cabo bajo un esquema de validación cruzada de k iteraciones o *K-Fold cross-validation* con $k = 10$.

3.3.3. Medidas de evaluación

Para la evaluación de los modelos generados para los distintos escenarios se emplean las métricas aquí descritas.

Accuracy Relación entre el número de verdaderos positivos (TP) y verdaderos negativos (TN) frente a la suma de positivos (P) y negativos (N). Indica la relación de muestras bien clasificadas frente al total.

$$accuracy = \frac{TP + TN}{P + N}$$

Precision Relación entre los verdaderos positivos (TP) frente a la suma de verdaderos positivos (TP) y falsos positivos (FP).

$$precision = \frac{TP}{TP + FP}$$

Recall Relación entre el número de verdaderos positivos (TP) frente a la suma de verdaderos positivos (TP) y falsos negativos (FN).

$$recall = \frac{TP}{TP + FN}$$

F1-score o F-score Media armónica entre la *precision* y el *recall*.

$$Fscore = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

3.4 Análisis de la varianza (ANOVA)

El análisis de la varianza o ANOVA (*Analysis of variance*) estudia el efecto de una o varias variables independientes, denominadas factores, sobre una o varias variables dependientes. Se trata, por tanto, de una generalización del contraste de medias para dos muestras con datos independientes y se aplica en situaciones en las que se quieran comparar tres o más grupos de datos. Los grupos vienen definidos por los factores a estudio.

Para este trabajo se tienen las siguientes variables o factores a estudio:

- **CLASIFICADOR**, con los niveles NAIVE-BAYES, SVM y TREE.
- **NGRAMAS**, con los niveles 1, 2, 3, 4 y 5.
- **VOCABULARIO**, con los niveles 500, 1.000 y 5.000.
- **CARACTERÍSTICAS**, con los niveles TF y TF-IDF.

Debido a esto se aplica un modelo factorial de análisis de varianza con los factores antes mencionados para evaluar sus efectos individuales y en conjunto sobre una variable dependiente cuantitativa, que en este caso es el ACCURACY de los modelos. En un análisis de varianza factorial existe una hipótesis nula por cada factor y por cada posible combinación de factores. La hipótesis nula referida a un factor afirma que las medias de las poblaciones definidas por los niveles del factor son iguales. La hipótesis referida al efecto de una interacción afirma que tal efecto es nulo. Para contrastar estas hipótesis el ANOVA facorial se sirve de estadísticos F.

En un modelo de cuatro factores, los efectos de interés a estudiar son quince: los cuatro factores principales (uno por cada factor), el efecto de la interacción de los factores dos a dos, el efecto de la interacción de los factores tres a tres y el efecto de la interacción entre todos los factores. A medida que aumenta el número de factores en un análisis ANOVA este se complica, por lo que se ha optado por realizar un análisis con dos factores, siendo éstos una combinación de los originales. Más adelante se argumentará y comprobará que el análisis no varía en sus resultados y es equivalente. A continuación se especifican los nuevos dos factores y la combinación escogida entre los iniciales:

- **CLASNGRAM**, combinación de los factores CLASIFICADOR y NGRAMAS.
- **VOCABCARACT**, combinación de los factores VOCABULARIO y CARACTERÍSTICAS.

CAPÍTULO 4
Resultados experimentales

En este capítulo se describen los resultados obtenidos durante el proceso de evaluación, aplicando las técnicas y representaciones ya mencionadas en el Capítulo 3 para los cuatro escenarios propuestos.

4.1 Modelos de *n*-gramas de palabras

Las Tablas A.1, A.2 y A.3 muestra los resultados para el Escenario 1, las Tablas A.4, A.5 y A.6 para el Escenario 2, las Tablas A.7, A.8 y A.9 para el Escenario 3 y las Tablas A.10, A.11 y A.12 para el Escenario 4.

Las tablas anteriores muestran para cada combinación la media de 10 modelos. Con el objetivo de estudiar el impacto de cada variable en la precisión de los modelos se realiza un análisis de la varianza o ANOVA. Tras el mismo se presentan los resultados y la combinación de características que mejor explican la precisión global.

4.2 Análisis de varianza (ANOVA)

4.2.1. Consideraciones previas

Para poder aplicar el análisis ANOVA se deben cumplir tres hipótesis, aunque se aceptan ligeras desviaciones de las condiciones ideales:

- **Independencia de las observaciones**. Cada conjunto de datos debe ser independiente del resto.
- **Normalidad de las observaciones**. Los resultados obtenidos para cada conjunto deben seguir una distribución normal.
- **Homogeneidad de las varianzas**. Las varianzas de cada conjunto no deben diferir de forma significativa.

4.2.1.1. Independencia

En los modelos generados en los cuatro escenarios propuestos se cumple la condición de independencia de las observaciones. Los tweets con los que se han generado dichos modelos han sido seleccionados de forma totalmente aleatoria e independiente para cada uno de ellos.

4.2.1.2. Normalidad

En la Figura B.1 se observan las pruebas de normalidad relativas al modelo ANOVA de 4 factores para el Escenario 1. Según el test de normalidad de Kolmogorov-Smirnov, de los 90 grupos todos resultan asemejarse a una distribución normal ($p > 0{,}05$) salvo 4 de ellos. Si estudiamos ahora la normalidad para el mismo escenario en el modelo ANOVA de 2 factores (ver Figura B.2), también se cumple el test de normalidad para los grupos salvo para 4 casos. De igual forma se cumple la normalidad para el Escenario 2 (ver Figuras B.3 y B.3), para el Escenario 3 (ver Figuras B.5 y B.5) y para el Escenario 4 (ver Figuras B.7 y B.7).

4.2.1.3. Homogeneidad de varianzas

La prueba de igualdad de Levene de varianzas de error prueba la hipótesis nula que la varianza de error de la variable dependiente (ACCURACY) es igual entre grupos.

En las Figuras B.9a y B.9b se observa el resultado de la prueba para el Escenario 1 con el modelo ANOVA de 4 y 2 factores respectivamente. En ambos casos se puede afirmar que no existen diferencias significativas entre las varianzas de error de los grupos para ambos casos y se ve además que coinciden, probando la equivalencia de los dos modelos. De igual forma se cumple este supuesto para el Escenario 2 (ver Figuras B.10a y B.10b), para el Escenario 3 (ver Figuras B.11a y B.11b) y para el Escenario 4 (ver Figuras B.12a y B.12b).

Tras presentar los resultados de las consideraciones previas necesarias para realizar los análisis ANOVA se pasan a estudiar los demolos de 2 variables.

4.2.2. Escenario 1 - PP y PSOE

En primer lugar se procede a estudiar los resultados obtenidos para los dos partidos mayoritarios de ambas cámaras, el Partido Popular (PP) y el Partido Socialista Obrero Español (PSOE). En la Figura B.13 se observa la tabla de los factores inter-sujetos. El *modelo corregido* se refiere a todos los efectos del modelo tomados juntos (el efecto de los dos factores, el de la interacción y de la constante o intersección). El p-valor asociado es menor a 0.05, lo que indica que el modelo explica una parte significativa de la variación observada en la variable dependiente. El valor $R^2 = 0{,}538$ indica que los factores incluídos en el modelo están explicando el 53.8 % de la varianza de la variable dependiente ACCURACY. La *intersección* se refiere a la constante del modelo. Esta constante forma parte del

modelo y permite contrastar, el en caso de que esto tenga sentido, la hipótesis de que la media total de la variable dependiente vale cero en la población.

El resto de items recogen los efectos principales, es decir, los efectos individuales de los dos factores incluídos en el modelo: CLASNGRAM y VOCABCARACT. Los p-valores menores a 0.05 indican que los grupos definidos por los factores combinados CLASIFICADOR-NGRAMAS y VOCABULARIO-CARACTERÍSTICAS producen que las precisiones de los modelos sean significativamente diferentes para cada combinación de ellos. Por último vemos el efecto de la interacción entre CLASNGRAM y VOCABCARACT. Como el p-valor es mayor a 0.05 el efecto de esta interacción no es significativo, lo que indica que en base a las distintas combinaciones de estos nuevos factores no se producen diferencias significativas en la precisión de los modelos; esto es, ni el método de clasificación, ni la combinación de *n*-gramas, ni el tamaño de vocabulario ni el método de representación de características hacen que la precisión de los modelos sea distinta o, dicho de otra forma, da igual la combinación de factores a escoger, pues la precisión obtenida será significativamente igual en todos los casos.

Como en todo análisis ANOVA se hace necesario realizar comparaciones post hoc de los factores. En la Figura B.14 se muestra el efecto de la variable CLASNGRAM sobre la precisión de los modelos. De todas las combinaciones son solamente dos (NAIVE-BAYES-1 y SVM-1) las que presentan un mayor accuracy, siendo significativamente diferentes del resto. En la Figura B.15 se evalúa, por otro lado, el factor VOCABCARACT. La mejor combinación en este caso la presentan las combinaciones 5000-TF-IDF y 5000-TF.

4.2.3. Escenario 2 - PP, PSOE, PODEMOS y CIUDADANOS

Si se observa ahora el escenario 2, formado por los 4 partidos mayoritarios (PP, PSOE, PODEMOS y CIUDADANOS), se ve en la Tabla B.16 los factores intersujetos. El p-valor asociado al *modelo corregido* es menor a 0.05, signo de que el modelo explica la variación observada en la variable dependiente. El valor $R^2 = 0{,}824$ indica que los factores incluídos en el modelo están explicando el 82.4 % de la varianza de la variable dependiente ACCURACY.

Los p-valores menores a 0.05 indican que los grupos definidos por los factores combinados CLASIFICADOR-NGRAMAS y VOCABULARIO-CARACTERÍSTICAS producen que las precisiones de los modelos sean significativamente diferentes para cada combinación de ellos. El efecto de la interacción entre CLASNGRAM y VOCABCARACT arroja un p-valor de 0.640 (mayor, por tanto, a 0.05) indicando así que no hay diferencias significativas en la interacción de ambos factores.

Las comparaciones post-hoc de la Figura B.17 indica que la mejor combinación es SVM-1. En la Figura B.18, por otra parte, las mejores combinaciones son 5000-TF-IDF y 5000-TF.

4.2.4. Escenario 3 - PP, PSOE, PODEMOS, CIUDADANOS y OTROS

En el escenario 3 se realiza el análisis para los cuatro partidos mayoritarios pero se incluye una clase, OTROS, donde se aglutinan el resto de partidos de las

cámaras legislativas españolas. En la Tabla B.19 se muestran los factores inter-sujetos. Con un p-valor inferior a 0.05, el *modelo corregido* explica la variación en la variable independiente y su valor $R^2 = 0{,}896$ indica que explica el 89.6 % de la varianza de la variable dependiente ACCURACY.

Los p-valores menores a 0.05 indican que los grupos definidos por los factores combinados CLASIFICADOR-NGRAMAS y VOCABULARIO-CARACTERÍSTICAS producen que las precisiones de los modelos sean significativamente diferentes para cada combinación de ellos. El efecto de la interacción entre CLASNGRAM y VOCABCARACT arroja un p-valor menor a 0.05) indicando que sí que hay diferencias significativas en la interacción de ambos factores.

En lo que respecta a las pruebas post-hoc, la Figura B.20 muestra que la mejor combinación de la variable CLASNGRAM es aquella que tiene las SVM como método de clasficación y 1 n-gramas. La Figura B.21 muestra que la mejor combinación de la variable VOCABCARACT es aquella donde se tienen en cuenta solamente 5000 palabras de vocabulario y se escoge el método TF como método de representación de características.

4.2.5. Escenario 4 - Todos los partidos

El escenario 4, donde se tienen en cuenta todos los partidos presentes en las Cortes, arroja los siguientes resultados. En la Tabla B.22 se muestran los factores inter-sujetos. Con un p-valor inferior a 0.05, el *modelo corregido* explica la variación en la variable independiente y su valor $R^2 = 0{,}985$ indica que explica el 98.5 % de la varianza de la variable dependiente ACCURACY.

Los p-valores menores a 0.05 indican que los grupos definidos por los factores combinados CLASIFICADOR-NGRAMAS y VOCABULARIO-CARACTERÍSTICAS producen que las precisiones de los modelos sean significativamente diferentes para cada combinación de ellos. El efecto de la interacción entre CLASNGRAM y VOCABCARACT arroja un p-valor menor a 0.05) indicando que sí que hay diferencias significativas en la interacción de ambos factores.

Las pruebas post-hoc (ver Figuras B.23 y B.24) sostienen que la mejor combinación para la variable VOCABCARACT es aquella en la que se tiene un vocabulario de tamaño 5000 usando el método de representación TF. Para la variable CLASNGRAM, por su parte, la mejor combinación la produce el clasificador SVM escogiendo 1 n-gramas de palabras.

CAPÍTULO 5
Conclusiones y líneas de trabajo futuras

En esta memoria se presenta el trabajo llevado a cabo para la tarea de identificación de perfiles políticos de usuarios en Twitter. Para ello se ha construido un dataset con hasta 1000 tweets de cada diputado y senador de la XII Legislatura de las Cortes Españoles con presencia en dicha red social. Para obtener las cuentas de sus señorías ha sido necesario programar una araña web para extraer la información que se encuentra en las páginas oficiales tanto del Congreso de los Diputados como del Senado de España, y una vez descargados los tweets se ha prodecido a etiquetarlos según la adscripción al partido político de su autor.

Para la tarea de identificación de perfiles políticos se han aplicado distintas técnicas de representación de documentos: se han utilizado *n*-gramas de palabras (con $n = 1, 2, 3, 4, 5$), distintos métodos de representación como el TF y el TF-IDF teniendo en cuenta distintos tamaños de vocabulario (500, 1000 ó 5000 palabras). Con el objetivo de evaluar el rendimiento de los clasificadores y las características antes mencionadas se diseñaron 4 escenarios de pruebas. El primer escenario, formado por cuentas del PP y del PSOE; el segundo, con tweets pertenecientes a los 4 partidos mayoritarios de las Cámaras: PP, PSOE, PODEMOS y CIUDADANOS; un tercer modelo similar al segundo pero añadiendo una categoría OTROS (donde se incluían el resto de partidos) y, finalmente, un último escenario con todos los partidos presentes en las Cortes.

Para cada combinación de factores y para cada escenario se realizó un *k*-fold cross validation con $k = 10$ y se midió el rendimiento de los modelos, para luego realizar un análisis de la varianza o ANOVA para estudiar y determinar el comportamiento del *accuracy* de los mismos. Si nos fijamos en el rendimiento de los clasificadores en el escenario 1, las máquinas de vectores de soporte o SVM y el clasificador bayesiano presentan los mejores resultados, escogiendo uni-gramas de palabras, 5000 palabras de vocabulario y con los métodos TF y TF-IDF de representación de documentos, obteniendo *accuracies* cercanos al 60 %. Fijándonos en el escenario 2 se cumplen los supuestos vistos en el primer escenario. Las SVM, junto con los uni-gramas de palabras, las 5000 palabras de vocabulario y los métodos TF-IDF presentan los mejores resultados, logrando *accuracies* del 40 %. Para los dos escenarios restantes se repite la misma configuración: los mejores resultados se obtienen, nuevamente, con las SVM y escogiendo 5000 palabras de vocabulario, uni-gramas de palabras y TF como método de representación de

documentos, logrando *accuracies* del 40 % para el escenario 3 y del 36 % para el escenario 4.

Uno de los trabajos de investigación futuros consiste en la reducción de las variables consideradas a estudio en este trabajo. La recolección de tweets es muy costosa en tiempo de descarga debido también a la cantidad de usuarios que se han tenido en cuenta en el estudio y el número de tweets a obtener para cada uno de ellos.

Otro futuro trabajo consistiría en realizar el estudio para distintos usuarios, ya que en este caso se han tenido en cuenta perfiles de políticos de primer nivel. Resultaría de mucho interés analizar los resultados obtenidos si se cambia el corpus de usuarios; por ejemplo, teniendo en cuenta perfiles de analistas políticos o periodistas e incluso usuarios a nivel general de Twitter. Esto añadiría cierta complejidad al trabajo, ya que sería difícil encontrar un dataset de personas de distintos perfiles de los cuales se conozca su afiliación política para entrenar los modelos correspondientes. En este caso se debería estudiar también el posible impacto legal del estudio, ya que se tendrían almacenados en uno o varios ficheros una lista de usuarios o personas con su correspondiente afiliación política.

Bibliografía

APÉNDICE A
Resultados del análisis experimental

Escenario 1 - PP y PSOE Naïve Bayes						
	500 términos		**1.000 términos**		**5.000 términos**	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,5781	0,5868	0,6038	0,6113	0,6193	0,6206
2	0,5588	0,5631	0,5738	0,5731	0,5718	0,5825
3	0,5043	0,5193	0,5019	0,5087	0,5250	0,5181
4	0,5043	0,5193	0,5019	0,5087	0,5250	0,5181
5	0,5019	0,5137	0,5075	0,5125	0,5425	0,5206

Tabla A.1: *Accuracies* obtenidos para el Escenario 1 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Naïve Bayes como clasificador para distintos tamaños de vocabulario.

Escenario 1 - PP y PSOE SVM						
	500 términos		**1.000 términos**		**5.000 términos**	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,6000	0,5981	0,6175	0,5875	0,6188	0,6106
2	0,5725	0,5668	0,5706	0,5525	0,5775	0,5569
3	0,5300	0,5275	0,5306	0,5431	0,5581	0,5493
4	0,4887	0,4956	0,5062	0,5075	0,5106	0,5138
5	0,4706	0,4931	0,4775	0,4768	0,4906	0,5050

Tabla A.2: *Accuracies* obtenidos para el Escenario 1 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con SVM como clasificador para distintos tamaños de vocabulario.

Escenario 1 - PP y PSOE Árboles de clasificación						
	500 términos		1.000 términos		5.000 términos	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,5700	0,5613	0,5800	0,5781	0,5750	0,5675
2	0,5425	0,5406	0,5419	0,5438	0,5706	0,5488
3	0,5363	0,5356	0,5175	0,5262	0,5393	0,5300
4	0,4731	0,4918	0,4981	0,5000	0,5088	0,4988
5	0,4868	0,4863	0,4793	0,4868	0,5038	0,5031

Tabla A.3: *Accuracies* obtenidos para el Escenario 1 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Árboles de clasificación como clasificador para distintos tamaños de vocabulario.

Escenario 1 - PP, PSOE, PODEMOS y CIUDADANOS Naïve Bayes						
	500 términos		1.000 términos		5.000 términos	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,3800	0,3775	0,3921	0,3734	0,4134	0,4003
2	0,3478	0,3440	0,3400	0,3415	0,3640	0,3603
3	0,3062	0,2990	0,3146	0,3100	0,3306	0,3231
4	0,2662	0,2706	0,2709	0,2625	0,2828	0,2875
5	0,2518	0,2593	0,2587	0,2603	0,2821	0,2759

Tabla A.4: *Accuracies* obtenidos para el Escenario 2 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Naïve Bayes como clasificador para distintos tamaños de vocabulario.

Escenario 1 - PP, PSOE, PODEMOS y CIUDADANOS SVM						
	500 términos		1.000 términos		5.000 términos	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,4156	0,4053	0,4121	0,3931	0,4287	0,3943
2	0,3490	0,3425	0,3659	0,3456	0,3762	0,3609
3	0,3081	0,3006	0,3253	0,3084	0,3321	0,3190
4	0,2481	0,2631	0,2437	0,2556	0,2803	0,2581
5	0,2462	0,2521	0,2484	0,2521	0,2503	0,2571

Tabla A.5: *Accuracies* obtenidos para el Escenario 2 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con SVM como clasificador para distintos tamaños de vocabulario.

Escenario 2 - PP, PSOE, PODEMOS y CIUDADANOS Árboles de clasificación						
	500 términos		**1.000 términos**		**5.000 términos**	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,3584	0,3618	0,3562	0,3665	0,3737	0,3721
2	0,3412	0,3459	0,3562	0,3625	0,3653	0,3725
3	0,3078	0,2938	0,3128	0,3075	0,3171	0,3238
4	0,2612	0,2618	0,2646	0,2496	0,2606	0,2588
5	0,2556	0,2496	0,2521	0,2531	0,2481	0,2544

Tabla A.6: *Accuracies* obtenidos para el Escenario 2 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Árboles de clasificación como clasificador para distintos tamaños de vocabulario.

Escenario 3 - PP, PSOE, PODEMOS, CIUDADANOS y OTROS Naïve Bayes						
	500 términos		**1.000 términos**		**5.000 términos**	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,3318	0,3267	0,3533	0,3475	0,4223	0,4018
2	0,2948	0,2910	0,3210	0,3210	0,3645	0,3620
3	0,2553	0,2580	0,2708	0,3675	0,2958	0,2898
4	0,2203	0,2251	0,2303	0,2293	0,2368	0,2353
5	0,2140	0,1985	0,2070	0,2110	0,2193	0,2170

Tabla A.7: *Accuracies* obtenidos para el Escenario 3 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Naïve Bayes como clasificador para distintos tamaños de vocabulario.

Escenario 3 - PP, PSOE, PODEMOS, CIUDADANOS y OTROS SVM						
	500 términos		**1.000 términos**		**5.000 términos**	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,3793	0,3700	0,3198	0,3823	0,4250	0,3738
2	0,3393	0,3268	0,3498	0,3380	0,3703	0,3415
3	0,2733	0,2693	0,2773	0,2768	0,2955	0,2945
4	0,2218	0,2195	0,2283	0,2290	0,2338	0,2350
5	0,1928	0,2033	0,2005	0,2043	0,2190	0,2188

Tabla A.8: *Accuracies* obtenidos para el Escenario 3 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con SVM como clasificador para distintos tamaños de vocabulario.

Escenario 3 - PP, PSOE, PODEMOS, CIUDADANOS y OTROS Árboles de clasificación						
	500 términos		1.000 términos		5.000 términos	
n	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF
1	0,3225	0,3098	0,3148	0,3125	0,3383	0,3228
2	0,3110	0,3185	0,3228	0,3455	0,3363	0,3308
3	0,2648	0,2698	0,2753	0,2740	0,2788	0,2788
4	0,2225	0,2175	0,2225	0,2230	0,2268	0,2265
5	0,2078	0,2070	0,1958	0,1963	0,2145	0,2135

Tabla A.9: *Accuracies* obtenidos para el Escenario 3 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Árboles de clasificación como clasificador para distintos tamaños de vocabulario.

Escenario 4 - Todos los partidos Naïve Bayes						
	500 términos		1.000 términos		5.000 términos	
n	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF
1	0,2590	0,2532	0,2931	0,2898	0,3670	0,3610
2	0,1775	0,1756	0,2129	0,2074	0,3008	0,2901
3	0,1413	0,1410	0,1636	0,1597	0,2108	0,2148
4	0,1114	0,1109	0,1164	0,1169	0,1331	0,1386
5	0,0955	0,1016	0,0979	0,0968	0,1115	0,1085

Tabla A.10: *Accuracies* obtenidos para el Escenario 4 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Naïve Bayes como clasificador para distintos tamaños de vocabulario.

Escenario 4 - Todos los partidos SVM						
	500 términos		1.000 términos		5.000 términos	
n	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF
1	0,3584	0,3576	0,3896	0,3774	0,4292	0,3747
2	0,2561	0,2519	0,2923	0,2793	0,3402	0,2981
3	0,1654	0,1647	0,1846	0,1831	0,2284	0,2173
4	0,2218	0,2195	0,2283	0,2290	0,2338	0,2350
5	0,1139	0,1122	0,1193	0,1164	0,1381	0,1389

Tabla A.11: *Accuracies* obtenidos para el Escenario 4 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con SVM como clasificador para distintos tamaños de vocabulario.

Escenario 4 - Todos los partidos Árboles de clasificación						
	500 términos		**1.000 términos**		**5.000 términos**	
n	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
1	0,2858	0,2876	0,3011	0,3020	0,3169	0,3150
2	0,2226	0,2219	0,2497	0,2514	0,2809	0,2830
3	0,1530	0,1580	0,1781	0,1729	0,2153	0,2151
4	0,2225	0,2175	0,2225	0,2230	0,2268	0,2265
5	0,1129	0,1143	0,1160	0,1149	0,1329	0,1335

Tabla A.12: *Accuracies* obtenidos para el Escenario 4 utilizando representaciones TF y TF-IDF para *n*-gramas de palabras con Árboles de clasificación como clasificador para distintos tamaños de vocabulario.

APÉNDICE B
Resultados del estudio del análisis de la varianza (ANOVA)

Pruebas de normalidad

					Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASIFICADOR	NGRAMAS	VOCABULARIO	CARACTERÍSTICAS		Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES	1	500	TF	ACCURACY	,137	10	,200*	,974	10	,921
			TF-IDF	ACCURACY	,209	10	,200*	,905	10	,249
		1000	TF	ACCURACY	,192	10	,200*	,938	10	,533
			TF-IDF	ACCURACY	,137	10	,200*	,962	10	,804
		5000	TF	ACCURACY	,151	10	,200*	,945	10	,608
			TF-IDF	ACCURACY	,277	10	,028	,831	10	,035
	2	500	TF	ACCURACY	,165	10	,200*	,944	10	,597
			TF-IDF	ACCURACY	,154	10	,200*	,938	10	,527
		1000	TF	ACCURACY	,243	10	,095	,921	10	,367
			TF-IDF	ACCURACY	,166	10	,200*	,899	10	,214
		5000	TF	ACCURACY	,229	10	,145	,823	10	,028
			TF-IDF	ACCURACY	,176	10	,200*	,962	10	,805
	3	500	TF	ACCURACY	,144	10	,200*	,968	10	,874
			TF-IDF	ACCURACY	,240	10	,106	,831	10	,035
		1000	TF	ACCURACY	,175	10	,200*	,947	10	,632
			TF-IDF	ACCURACY	,239	10	,112	,879	10	,129
		5000	TF	ACCURACY	,276	10	,030	,796	10	,013
			TF-IDF	ACCURACY	,256	10	,062	,847	10	,054
	4	500	TF	ACCURACY	,215	10	,200*	,878	10	,125
			TF-IDF	ACCURACY	,141	10	,200*	,979	10	,960
		1000	TF	ACCURACY	,268	10	,041	,907	10	,259
			TF-IDF	ACCURACY	,151	10	,200*	,934	10	,486
		5000	TF	ACCURACY	,135	10	,200*	,965	10	,845
			TF-IDF	ACCURACY	,130	10	,200*	,980	10	,968
	5	500	TF	ACCURACY	,205	10	,200*	,951	10	,678
			TF-IDF	ACCURACY	,195	10	,200*	,898	10	,208
		1000	TF	ACCURACY	,175	10	,200*	,951	10	,682
			TF-IDF	ACCURACY	,168	10	,200*	,960	10	,780
		5000	TF	ACCURACY	,255	10	,065	,870	10	,100
			TF-IDF	ACCURACY	,170	10	,200*	,957	10	,746
SVM	1	500	TF	ACCURACY	,152	10	,200*	,951	10	,684
			TF-IDF	ACCURACY	,214	10	,200*	,896	10	,197
		1000	TF	ACCURACY	,173	10	,200*	,940	10	,557
			TF-IDF	ACCURACY	,224	10	,170	,929	10	,441
		5000	TF	ACCURACY	,250	10	,076	,861	10	,078
			TF-IDF	ACCURACY	,182	10	,200*	,953	10	,709
	2	500	TF	ACCURACY	,217	10	,200*	,914	10	,308
			TF-IDF	ACCURACY	,172	10	,200*	,943	10	,588
		1000	TF	ACCURACY	,182	10	,200*	,950	10	,674
			TF-IDF	ACCURACY	,158	10	,200*	,916	10	,323
		5000	TF	ACCURACY	,252	10	,071	,853	10	,062
			TF-IDF	ACCURACY	,121	10	,200*	,979	10	,959
	3	500	TF	ACCURACY	,255	10	,064	,918	10	,338
			TF-IDF	ACCURACY	,131	10	,200*	,981	10	,971
		1000	TF	ACCURACY	,175	10	,200*	,953	10	,702
			TF-IDF	ACCURACY	,193	10	,200*	,916	10	,327
		5000	TF	ACCURACY	,167	10	,200*	,927	10	,417
			TF-IDF	ACCURACY	,139	10	,200*	,957	10	,749
	4	500	TF	ACCURACY	,154	10	,200*	,962	10	,810
			TF-IDF	ACCURACY	,142	10	,200*	,967	10	,864
		1000	TF	ACCURACY	,238	10	,115	,870	10	,101
			TF-IDF	ACCURACY	,180	10	,200*	,921	10	,364
		5000	TF	ACCURACY	,158	10	,200*	,956	10	,742
			TF-IDF	ACCURACY	,189	10	,200*	,896	10	,197
	5	500	TF	ACCURACY	,162	10	,200*	,937	10	,524
			TF-IDF	ACCURACY	,238	10	,116	,862	10	,080
		1000	TF	ACCURACY	,249	10	,079	,883	10	,142
			TF-IDF	ACCURACY	,178	10	,200*	,942	10	,575
		5000	TF	ACCURACY	,148	10	,200*	,936	10	,509
			TF-IDF	ACCURACY	,148	10	,200*	,940	10	,554
TREE	1	500	TF	ACCURACY	,158	10	,200*	,956	10	,739
			TF-IDF	ACCURACY	,156	10	,200*	,963	10	,816
		1000	TF	ACCURACY	,166	10	,200*	,909	10	,277
			TF-IDF	ACCURACY	,169	10	,200*	,965	10	,840
		5000	TF	ACCURACY	,207	10	,200*	,894	10	,187
			TF-IDF	ACCURACY	,314	10	,006	,770	10	,006
	2	500	TF	ACCURACY	,190	10	,200*	,965	10	,839
			TF-IDF	ACCURACY	,175	10	,200*	,865	10	,087
		1000	TF	ACCURACY	,255	10	,065	,918	10	,341
			TF-IDF	ACCURACY	,152	10	,200*	,957	10	,748
		5000	TF	ACCURACY	,129	10	,200*	,983	10	,980
			TF-IDF	ACCURACY	,196	10	,200*	,913	10	,303
	3	500	TF	ACCURACY	,199	10	,200*	,935	10	,504
			TF-IDF	ACCURACY	,150	10	,200*	,964	10	,829
		1000	TF	ACCURACY	,189	10	,200*	,947	10	,634
			TF-IDF	ACCURACY	,348	10	,001	,795	10	,012
		5000	TF	ACCURACY	,137	10	,200*	,942	10	,578
			TF-IDF	ACCURACY	,146	10	,200*	,942	10	,581
	4	500	TF	ACCURACY	,120	10	,200*	,975	10	,935
			TF-IDF	ACCURACY	,202	10	,200*	,875	10	,114
		1000	TF	ACCURACY	,100	10	,200*	,990	10	,997
			TF-IDF	ACCURACY	,142	10	,200*	,942	10	,571
		5000	TF	ACCURACY	,178	10	,200*	,941	10	,569
			TF-IDF	ACCURACY	,152	10	,200*	,970	10	,894
	5	500	TF	ACCURACY	,202	10	,200*	,895	10	,191
			TF-IDF	ACCURACY	,217	10	,200	,956	10	,738
		1000	TF	ACCURACY	,182	10	,200*	,932	10	,473
			TF-IDF	ACCURACY	,161	10	,200*	,956	10	,744
		5000	TF	ACCURACY	,171	10	,200*	,910	10	,280
			TF-IDF	ACCURACY	,204	10	,200*	,915	10	,317

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.1: Prueba de normalidad para el Escenario 1 con un ANOVA de 4 factores

Pruebas de normalidad

			Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASNGRAM	VOCABNCARACT		Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES-1	1000-TF	ACCURACY	,192	10	,200*	,938	10	,533
	1000-TF-IDF	ACCURACY	,137	10	,200*	,962	10	,804
	500-TF	ACCURACY	,137	10	,200*	,974	10	,921
	500-TF-IDF	ACCURACY	,209	10	,200*	,905	10	,249
	5000-TF	ACCURACY	,151	10	,200*	,945	10	,608
	5000-TF-IDF	ACCURACY	,277	10	,028	,831	10	,035
NAIVE-BAYES-2	1000-TF	ACCURACY	,243	10	,095	,921	10	,367
	1000-TF-IDF	ACCURACY	,166	10	,200*	,899	10	,214
	500-TF	ACCURACY	,165	10	,200*	,944	10	,597
	500-TF-IDF	ACCURACY	,154	10	,200*	,938	10	,527
	5000-TF	ACCURACY	,229	10	,145	,823	10	,028
	5000-TF-IDF	ACCURACY	,176	10	,200*	,962	10	,805
NAIVE-BAYES-3	1000-TF	ACCURACY	,175	10	,200*	,947	10	,632
	1000-TF-IDF	ACCURACY	,239	10	,112	,879	10	,129
	500-TF	ACCURACY	,144	10	,200*	,968	10	,874
	500-TF-IDF	ACCURACY	,240	10	,106	,831	10	,035
	5000-TF	ACCURACY	,276	10	,030	,796	10	,013
	5000-TF-IDF	ACCURACY	,256	10	,062	,847	10	,054
NAIVE-BAYES-4	1000-TF	ACCURACY	,268	10	,041	,907	10	,259
	1000-TF-IDF	ACCURACY	,151	10	,200*	,934	10	,486
	500-TF	ACCURACY	,215	10	,200*	,878	10	,125
	500-TF-IDF	ACCURACY	,141	10	,200*	,979	10	,960
	5000-TF	ACCURACY	,135	10	,200*	,965	10	,845
	5000-TF-IDF	ACCURACY	,130	10	,200*	,980	10	,968
NAIVE-BAYES-5	1000-TF	ACCURACY	,175	10	,200*	,951	10	,682
	1000-TF-IDF	ACCURACY	,168	10	,200*	,960	10	,780
	500-TF	ACCURACY	,205	10	,200*	,951	10	,678
	500-TF-IDF	ACCURACY	,195	10	,200*	,898	10	,208
	5000-TF	ACCURACY	,255	10	,065	,870	10	,100
	5000-TF-IDF	ACCURACY	,170	10	,200*	,957	10	,746
SVM-1	1000-TF	ACCURACY	,173	10	,200*	,940	10	,557
	1000-TF-IDF	ACCURACY	,224	10	,170	,929	10	,441
	500-TF	ACCURACY	,152	10	,200*	,951	10	,684
	500-TF-IDF	ACCURACY	,214	10	,200*	,896	10	,197
	5000-TF	ACCURACY	,250	10	,076	,861	10	,078
	5000-TF-IDF	ACCURACY	,182	10	,200*	,953	10	,709
SVM-2	1000-TF	ACCURACY	,182	10	,200*	,950	10	,674
	1000-TF-IDF	ACCURACY	,158	10	,200*	,916	10	,323
	500-TF	ACCURACY	,217	10	,200*	,914	10	,308
	500-TF-IDF	ACCURACY	,172	10	,200*	,943	10	,588
	5000-TF	ACCURACY	,252	10	,071	,853	10	,062
	5000-TF-IDF	ACCURACY	,121	10	,200*	,979	10	,959
SVM-3	1000-TF	ACCURACY	,175	10	,200*	,953	10	,702
	1000-TF-IDF	ACCURACY	,193	10	,200*	,916	10	,327
	500-TF	ACCURACY	,255	10	,064	,918	10	,338
	500-TF-IDF	ACCURACY	,131	10	,200*	,981	10	,971
	5000-TF	ACCURACY	,167	10	,200*	,927	10	,417
	5000-TF-IDF	ACCURACY	,139	10	,200*	,957	10	,749
SVM-4	1000-TF	ACCURACY	,238	10	,115	,870	10	,101
	1000-TF-IDF	ACCURACY	,180	10	,200*	,921	10	,364
	500-TF	ACCURACY	,154	10	,200*	,962	10	,810
	500-TF-IDF	ACCURACY	,142	10	,200*	,967	10	,864
	5000-TF	ACCURACY	,158	10	,200*	,956	10	,742
	5000-TF-IDF	ACCURACY	,189	10	,200*	,896	10	,197
SVM-5	1000-TF	ACCURACY	,249	10	,079	,883	10	,142
	1000-TF-IDF	ACCURACY	,178	10	,200*	,942	10	,575
	500-TF	ACCURACY	,162	10	,200*	,937	10	,524
	500-TF-IDF	ACCURACY	,238	10	,116	,862	10	,080
	5000-TF	ACCURACY	,148	10	,200*	,936	10	,509
	5000-TF-IDF	ACCURACY	,148	10	,200*	,940	10	,554
TREE-1	1000-TF	ACCURACY	,166	10	,200*	,909	10	,277
	1000-TF-IDF	ACCURACY	,169	10	,200*	,965	10	,840
	500-TF	ACCURACY	,158	10	,200*	,956	10	,739
	500-TF-IDF	ACCURACY	,156	10	,200*	,963	10	,816
	5000-TF	ACCURACY	,207	10	,200*	,894	10	,187
	5000-TF-IDF	ACCURACY	,314	10	,006	,770	10	,006
TREE-2	1000-TF	ACCURACY	,255	10	,065	,918	10	,341
	1000-TF-IDF	ACCURACY	,152	10	,200*	,957	10	,748
	500-TF	ACCURACY	,190	10	,200*	,965	10	,839
	500-TF-IDF	ACCURACY	,175	10	,200*	,865	10	,087
	5000-TF	ACCURACY	,129	10	,200*	,983	10	,980
	5000-TF-IDF	ACCURACY	,196	10	,200*	,913	10	,303
TREE-3	1000-TF	ACCURACY	,189	10	,200*	,947	10	,634
	1000-TF-IDF	ACCURACY	,348	10	,001	,795	10	,012
	500-TF	ACCURACY	,199	10	,200*	,935	10	,504
	500-TF-IDF	ACCURACY	,150	10	,200*	,964	10	,829
	5000-TF	ACCURACY	,137	10	,200*	,942	10	,578
	5000-TF-IDF	ACCURACY	,146	10	,200*	,942	10	,581
TREE-4	1000-TF	ACCURACY	,100	10	,200*	,990	10	,997
	1000-TF-IDF	ACCURACY	,142	10	,200*	,942	10	,571
	500-TF	ACCURACY	,120	10	,200*	,975	10	,935
	500-TF-IDF	ACCURACY	,202	10	,200*	,875	10	,114
	5000-TF	ACCURACY	,178	10	,200*	,941	10	,569
	5000-TF-IDF	ACCURACY	,152	10	,200*	,970	10	,894
TREE-5	1000-TF	ACCURACY	,182	10	,200*	,932	10	,473
	1000-TF-IDF	ACCURACY	,161	10	,200*	,956	10	,744
	500-TF	ACCURACY	,202	10	,200*	,895	10	,191
	500-TF-IDF	ACCURACY	,217	10	,200	,956	10	,738
	5000-TF	ACCURACY	,171	10	,200*	,910	10	,280
	5000-TF-IDF	ACCURACY	,204	10	,200*	,915	10	,317

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.2: Prueba de normalidad para el Escenario 1 con un ANOVA de 2 factores

Pruebas de normalidad

					Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASIFICADOR	NGRAMAS	VOCABULARIO	CARACTERISTICAS		Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES	1	500	TF	ACCURACY	,230	10	,144	,868	10	,095
			TF-IDF	ACCURACY	,224	10	,168	,863	10	,083
		1000	TF	ACCURACY	,180	10	,200*	,922	10	,370
			TF-IDF	ACCURACY	,139	10	,200*	,926	10	,408
		5000	TF	ACCURACY	,166	10	,200*	,948	10	,641
			TF-IDF	ACCURACY	,301	10	,011	,807	10	,018
	2	500	TF	ACCURACY	,135	10	,200*	,964	10	,835
			TF-IDF	ACCURACY	,124	10	,200*	,983	10	,980
		1000	TF	ACCURACY	,147	10	,200*	,951	10	,684
			TF-IDF	ACCURACY	,165	10	,200*	,927	10	,418
		5000	TF	ACCURACY	,162	10	,200*	,977	10	,948
			TF-IDF	ACCURACY	,172	10	,200*	,912	10	,295
	3	500	TF	ACCURACY	,182	10	,200*	,924	10	,394
			TF-IDF	ACCURACY	,186	10	,200*	,948	10	,649
		1000	TF	ACCURACY	,154	10	,200*	,945	10	,613
			TF-IDF	ACCURACY	,142	10	,200*	,950	10	,667
		5000	TF	ACCURACY	,169	10	,200*	,923	10	,383
			TF-IDF	ACCURACY	,155	10	,200*	,928	10	,431
	4	500	TF	ACCURACY	,136	10	,200*	,965	10	,845
			TF-IDF	ACCURACY	,128	10	,200*	,978	10	,954
		1000	TF	ACCURACY	,185	10	,200*	,939	10	,537
			TF-IDF	ACCURACY	,230	10	,142	,898	10	,208
		5000	TF	ACCURACY	,303	10	,010	,824	10	,028
			TF-IDF	ACCURACY	,139	10	,200*	,931	10	,460
	5	500	TF	ACCURACY	,176	10	,200*	,913	10	,300
			TF-IDF	ACCURACY	,248	10	,083	,936	10	,511
		1000	TF	ACCURACY	,135	10	,200*	,952	10	,698
			TF-IDF	ACCURACY	,248	10	,083	,885	10	,148
		5000	TF	ACCURACY	,215	10	,200*	,894	10	,186
			TF-IDF	ACCURACY	,234	10	,130	,867	10	,093
SVM	1	500	TF	ACCURACY	,234	10	,129	,902	10	,231
			TF-IDF	ACCURACY	,193	10	,200*	,941	10	,568
		1000	TF	ACCURACY	,178	10	,200*	,921	10	,368
			TF-IDF	ACCURACY	,139	10	,200*	,967	10	,859
		5000	TF	ACCURACY	,150	10	,200*	,956	10	,741
			TF-IDF	ACCURACY	,289	10	,018	,861	10	,079
	2	500	TF	ACCURACY	,301	10	,010	,795	10	,013
			TF-IDF	ACCURACY	,157	10	,200*	,966	10	,857
		1000	TF	ACCURACY	,190	10	,200*	,932	10	,469
			TF-IDF	ACCURACY	,190	10	,200*	,941	10	,562
		5000	TF	ACCURACY	,160	10	,200*	,958	10	,764
			TF-IDF	ACCURACY	,273	10	,033	,846	10	,052
	3	500	TF	ACCURACY	,117	10	,200*	,985	10	,987
			TF-IDF	ACCURACY	,155	10	,200*	,959	10	,778
		1000	TF	ACCURACY	,206	10	,200*	,943	10	,591
			TF-IDF	ACCURACY	,237	10	,118	,913	10	,301
		5000	TF	ACCURACY	,098	10	,200*	,988	10	,994
			TF-IDF	ACCURACY	,219	10	,190	,917	10	,329
	4	500	TF	ACCURACY	,272	10	,035	,823	10	,027
			TF-IDF	ACCURACY	,207	10	,200*	,888	10	,160
		1000	TF	ACCURACY	,158	10	,200*	,943	10	,589
			TF-IDF	ACCURACY	,150	10	,200*	,953	10	,700
		5000	TF	ACCURACY	,141	10	,200*	,976	10	,940
			TF-IDF	ACCURACY	,230	10	,144	,897	10	,201
	5	500	TF	ACCURACY	,218	10	,196	,878	10	,122
			TF-IDF	ACCURACY	,189	10	,200*	,928	10	,424
		1000	TF	ACCURACY	,164	10	,200*	,948	10	,648
			TF-IDF	ACCURACY	,205	10	,200*	,876	10	,116
		5000	TF	ACCURACY	,157	10	,200*	,963	10	,822
			TF-IDF	ACCURACY	,143	10	,200*	,977	10	,944
TREE	1	500	TF	ACCURACY	,156	10	,200*	,959	10	,776
			TF-IDF	ACCURACY	,140	10	,200*	,959	10	,771
		1000	TF	ACCURACY	,196	10	,200*	,883	10	,142
			TF-IDF	ACCURACY	,294	10	,014	,772	10	,007
		5000	TF	ACCURACY	,182	10	,200*	,869	10	,097
			TF-IDF	ACCURACY	,171	10	,200*	,900	10	,222
	2	500	TF	ACCURACY	,127	10	,200*	,986	10	,989
			TF-IDF	ACCURACY	,204	10	,200*	,929	10	,443
		1000	TF	ACCURACY	,153	10	,200*	,948	10	,642
			TF-IDF	ACCURACY	,106	10	,200*	,949	10	,660
		5000	TF	ACCURACY	,225	10	,166	,898	10	,210
			TF-IDF	ACCURACY	,205	10	,200*	,902	10	,228
	3	500	TF	ACCURACY	,182	10	,200*	,937	10	,523
			TF-IDF	ACCURACY	,155	10	,200*	,958	10	,765
		1000	TF	ACCURACY	,147	10	,200*	,983	10	,979
			TF-IDF	ACCURACY	,134	10	,200*	,972	10	,912
		5000	TF	ACCURACY	,199	10	,200*	,924	10	,394
			TF-IDF	ACCURACY	,186	10	,200*	,896	10	,198
	4	500	TF	ACCURACY	,133	10	,200*	,932	10	,472
			TF-IDF	ACCURACY	,156	10	,200*	,936	10	,514
		1000	TF	ACCURACY	,240	10	,106	,889	10	,164
			TF-IDF	ACCURACY	,177	10	,200*	,973	10	,917
		5000	TF	ACCURACY	,133	10	,200*	,974	10	,926
			TF-IDF	ACCURACY	,148	10	,200*	,980	10	,966
	5	500	TF	ACCURACY	,170	10	,200*	,966	10	,852
			TF-IDF	ACCURACY	,167	10	,200*	,932	10	,468
		1000	TF	ACCURACY	,173	10	,200*	,962	10	,806
			TF-IDF	ACCURACY	,200	10	,200*	,923	10	,387
		5000	TF	ACCURACY	,120	10	,200*	,978	10	,953
			TF-IDF	ACCURACY	,179	10	,200*	,929	10	,439

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.3: Prueba de normalidad para el Escenario 2 con un ANOVA de 4 factores

Pruebas de normalidad

			Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASNGRAM	VOCABCARACT		Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES-1	1000-TF	ACCURACY	,180	10	,200*	,922	10	,370
	1000-TF-IDF	ACCURACY	,139	10	,200*	,926	10	,408
	500-TF	ACCURACY	,230	10	,144	,868	10	,095
	500-TF-IDF	ACCURACY	,224	10	,168	,863	10	,083
	5000-TF	ACCURACY	,166	10	,200*	,948	10	,641
	5000-TF-IDF	ACCURACY	,301	10	,011	,807	10	,018
NAIVE-BAYES-2	1000-TF	ACCURACY	,147	10	,200*	,951	10	,684
	1000-TF-IDF	ACCURACY	,165	10	,200*	,927	10	,418
	500-TF	ACCURACY	,135	10	,200*	,964	10	,835
	500-TF-IDF	ACCURACY	,124	10	,200*	,983	10	,980
	5000-TF	ACCURACY	,162	10	,200*	,977	10	,948
	5000-TF-IDF	ACCURACY	,172	10	,200*	,912	10	,295
NAIVE-BAYES-3	1000-TF	ACCURACY	,154	10	,200*	,945	10	,613
	1000-TF-IDF	ACCURACY	,142	10	,200*	,950	10	,667
	500-TF	ACCURACY	,182	10	,200*	,924	10	,394
	500-TF-IDF	ACCURACY	,186	10	,200*	,948	10	,649
	5000-TF	ACCURACY	,169	10	,200*	,923	10	,383
	5000-TF-IDF	ACCURACY	,155	10	,200*	,928	10	,431
NAIVE-BAYES-4	1000-TF	ACCURACY	,185	10	,200*	,939	10	,537
	1000-TF-IDF	ACCURACY	,230	10	,142	,898	10	,208
	500-TF	ACCURACY	,136	10	,200*	,965	10	,845
	500-TF-IDF	ACCURACY	,128	10	,200*	,978	10	,954
	5000-TF	ACCURACY	,303	10	,010	,824	10	,028
	5000-TF-IDF	ACCURACY	,139	10	,200*	,931	10	,460
NAIVE-BAYES-5	1000-TF	ACCURACY	,135	10	,200*	,952	10	,698
	1000-TF-IDF	ACCURACY	,248	10	,083	,885	10	,148
	500-TF	ACCURACY	,176	10	,200*	,913	10	,300
	500-TF-IDF	ACCURACY	,248	10	,083	,936	10	,511
	5000-TF	ACCURACY	,215	10	,200*	,894	10	,186
	5000-TF-IDF	ACCURACY	,234	10	,130	,867	10	,093
SVM-1	1000-TF	ACCURACY	,178	10	,200*	,921	10	,368
	1000-TF-IDF	ACCURACY	,139	10	,200*	,967	10	,859
	500-TF	ACCURACY	,234	10	,129	,902	10	,231
	500-TF-IDF	ACCURACY	,193	10	,200*	,941	10	,568
	5000-TF	ACCURACY	,150	10	,200*	,956	10	,741
	5000-TF-IDF	ACCURACY	,289	10	,018	,861	10	,079
SVM-2	1000-TF	ACCURACY	,190	10	,200*	,932	10	,469
	1000-TF-IDF	ACCURACY	,190	10	,200*	,941	10	,562
	500-TF	ACCURACY	,301	10	,010	,795	10	,013
	500-TF-IDF	ACCURACY	,157	10	,200*	,966	10	,857
	5000-TF	ACCURACY	,160	10	,200*	,958	10	,764
	5000-TF-IDF	ACCURACY	,273	10	,033	,846	10	,052
SVM-3	1000-TF	ACCURACY	,206	10	,200*	,943	10	,591
	1000-TF-IDF	ACCURACY	,237	10	,118	,913	10	,301
	500-TF	ACCURACY	,117	10	,200*	,985	10	,987
	500-TF-IDF	ACCURACY	,155	10	,200*	,959	10	,778
	5000-TF	ACCURACY	,098	10	,200*	,988	10	,994
	5000-TF-IDF	ACCURACY	,219	10	,190	,917	10	,329
SVM-4	1000-TF	ACCURACY	,158	10	,200*	,943	10	,589
	1000-TF-IDF	ACCURACY	,150	10	,200*	,953	10	,700
	500-TF	ACCURACY	,272	10	,035	,823	10	,027
	500-TF-IDF	ACCURACY	,207	10	,200*	,888	10	,160
	5000-TF	ACCURACY	,141	10	,200*	,976	10	,940
	5000-TF-IDF	ACCURACY	,230	10	,144	,897	10	,201
SVM-5	1000-TF	ACCURACY	,164	10	,200*	,948	10	,648
	1000-TF-IDF	ACCURACY	,205	10	,200*	,876	10	,116
	500-TF	ACCURACY	,218	10	,196	,878	10	,122
	500-TF-IDF	ACCURACY	,189	10	,200*	,928	10	,424
	5000-TF	ACCURACY	,157	10	,200*	,963	10	,822
	5000-TF-IDF	ACCURACY	,143	10	,200*	,977	10	,944
TREE-1	1000-TF	ACCURACY	,196	10	,200*	,883	10	,142
	1000-TF-IDF	ACCURACY	,294	10	,014	,772	10	,007
	500-TF	ACCURACY	,156	10	,200*	,959	10	,776
	500-TF-IDF	ACCURACY	,140	10	,200*	,959	10	,771
	5000-TF	ACCURACY	,182	10	,200*	,869	10	,097
	5000-TF-IDF	ACCURACY	,171	10	,200*	,900	10	,222
TREE-2	1000-TF	ACCURACY	,153	10	,200*	,948	10	,642
	1000-TF-IDF	ACCURACY	,106	10	,200*	,949	10	,660
	500-TF	ACCURACY	,127	10	,200*	,986	10	,989
	500-TF-IDF	ACCURACY	,204	10	,200*	,929	10	,443
	5000-TF	ACCURACY	,225	10	,166	,898	10	,210
	5000-TF-IDF	ACCURACY	,205	10	,200*	,902	10	,228
TREE-3	1000-TF	ACCURACY	,147	10	,200*	,983	10	,979
	1000-TF-IDF	ACCURACY	,134	10	,200*	,972	10	,912
	500-TF	ACCURACY	,182	10	,200*	,937	10	,523
	500-TF-IDF	ACCURACY	,155	10	,200*	,958	10	,765
	5000-TF	ACCURACY	,199	10	,200*	,924	10	,394
	5000-TF-IDF	ACCURACY	,186	10	,200*	,896	10	,198
TREE-4	1000-TF	ACCURACY	,240	10	,106	,889	10	,164
	1000-TF-IDF	ACCURACY	,177	10	,200*	,973	10	,917
	500-TF	ACCURACY	,133	10	,200*	,932	10	,472
	500-TF-IDF	ACCURACY	,156	10	,200*	,936	10	,514
	5000-TF	ACCURACY	,133	10	,200*	,974	10	,926
	5000-TF-IDF	ACCURACY	,148	10	,200*	,980	10	,966
TREE-5	1000-TF	ACCURACY	,173	10	,200*	,962	10	,806
	1000-TF-IDF	ACCURACY	,200	10	,200*	,923	10	,387
	500-TF	ACCURACY	,170	10	,200*	,966	10	,852
	500-TF-IDF	ACCURACY	,167	10	,200*	,932	10	,468
	5000-TF	ACCURACY	,120	10	,200*	,978	10	,953
	5000-TF-IDF	ACCURACY	,179	10	,200*	,929	10	,439

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.4: Prueba de normalidad para el Escenario 2 con un ANOVA de 2 factores

Pruebas de normalidad

					Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASIFICADOR	NGRAMAS	VOCABULARIO	CARACTERISTICAS		Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES	1	500	TF	ACCURACY	,175	10	,200*	,953	10	,701
			TF-IDF	ACCURACY	,136	10	,200*	,971	10	,901
		1000	TF	ACCURACY	,217	10	,198	,914	10	,311
			TF-IDF	ACCURACY	,176	10	,200*	,945	10	,615
		5000	TF	ACCURACY	,153	10	,200*	,974	10	,925
			TF-IDF	ACCURACY	,104	10	,200*	,984	10	,984
	2	500	TF	ACCURACY	,236	10	,123	,893	10	,182
			TF-IDF	ACCURACY	,285	10	,020	,822	10	,027
		1000	TF	ACCURACY	,229	10	,148	,901	10	,224
			TF-IDF	ACCURACY	,120	10	,200*	,955	10	,728
		5000	TF	ACCURACY	,209	10	,200*	,915	10	,315
			TF-IDF	ACCURACY	,228	10	,149	,923	10	,378
	3	500	TF	ACCURACY	,206	10	,200*	,882	10	,136
			TF-IDF	ACCURACY	,214	10	,200*	,868	10	,094
		1000	TF	ACCURACY	,198	10	,200*	,877	10	,119
			TF-IDF	ACCURACY	,182	10	,200*	,975	10	,932
		5000	TF	ACCURACY	,188	10	,200*	,937	10	,523
			TF-IDF	ACCURACY	,188	10	,200*	,915	10	,314
	4	500	TF	ACCURACY	,195	10	,200*	,962	10	,804
			TF-IDF	ACCURACY	,158	10	,200*	,961	10	,797
		1000	TF	ACCURACY	,215	10	,200*	,838	10	,041
			TF-IDF	ACCURACY	,187	10	,200*	,960	10	,787
		5000	TF	ACCURACY	,234	10	,129	,876	10	,119
			TF-IDF	ACCURACY	,159	10	,200*	,912	10	,295
	5	500	TF	ACCURACY	,123	10	,200*	,966	10	,848
			TF-IDF	ACCURACY	,223	10	,171	,812	10	,020
		1000	TF	ACCURACY	,141	10	,200*	,966	10	,849
			TF-IDF	ACCURACY	,220	10	,188	,941	10	,561
		5000	TF	ACCURACY	,191	10	,200*	,940	10	,555
			TF-IDF	ACCURACY	,169	10	,200*	,947	10	,632
SVM	1	500	TF	ACCURACY	,180	10	,200*	,940	10	,554
			TF-IDF	ACCURACY	,141	10	,200*	,975	10	,935
		1000	TF	ACCURACY	,174	10	,200*	,890	10	,169
			TF-IDF	ACCURACY	,122	10	,200*	,949	10	,653
		5000	TF	ACCURACY	,194	10	,200*	,917	10	,334
			TF-IDF	ACCURACY	,187	10	,200*	,931	10	,454
	2	500	TF	ACCURACY	,257	10	,060	,805	10	,016
			TF-IDF	ACCURACY	,176	10	,200*	,904	10	,241
		1000	TF	ACCURACY	,200	10	,200*	,936	10	,506
			TF-IDF	ACCURACY	,180	10	,200*	,947	10	,637
		5000	TF	ACCURACY	,136	10	,200*	,953	10	,698
			TF-IDF	ACCURACY	,225	10	,162	,883	10	,141
	3	500	TF	ACCURACY	,170	10	,200*	,921	10	,362
			TF-IDF	ACCURACY	,172	10	,200*	,920	10	,356
		1000	TF	ACCURACY	,156	10	,200*	,947	10	,631
			TF-IDF	ACCURACY	,113	10	,200*	,955	10	,724
		5000	TF	ACCURACY	,254	10	,066	,909	10	,276
			TF-IDF	ACCURACY	,170	10	,200*	,895	10	,195
	4	500	TF	ACCURACY	,213	10	,200*	,872	10	,106
			TF-IDF	ACCURACY	,110	10	,200*	,977	10	,950
		1000	TF	ACCURACY	,151	10	,200*	,944	10	,596
			TF-IDF	ACCURACY	,153	10	,200*	,932	10	,470
		5000	TF	ACCURACY	,138	10	,200*	,981	10	,972
			TF-IDF	ACCURACY	,170	10	,200*	,947	10	,634
	5	500	TF	ACCURACY	,247	10	,084	,911	10	,287
			TF-IDF	ACCURACY	,185	10	,200*	,940	10	,557
		1000	TF	ACCURACY	,222	10	,176	,896	10	,200
			TF-IDF	ACCURACY	,200	10	,200*	,892	10	,180
		5000	TF	ACCURACY	,195	10	,200*	,915	10	,314
			TF-IDF	ACCURACY	,146	10	,200*	,971	10	,904
TREE	1	500	TF	ACCURACY	,120	10	,200*	,950	10	,668
			TF-IDF	ACCURACY	,156	10	,200*	,952	10	,687
		1000	TF	ACCURACY	,166	10	,200*	,940	10	,549
			TF-IDF	ACCURACY	,225	10	,165	,910	10	,282
		5000	TF	ACCURACY	,140	10	,200*	,948	10	,647
			TF-IDF	ACCURACY	,279	10	,027	,825	10	,029
	2	500	TF	ACCURACY	,241	10	,102	,904	10	,245
			TF-IDF	ACCURACY	,152	10	,200*	,937	10	,523
		1000	TF	ACCURACY	,173	10	,200*	,904	10	,240
			TF-IDF	ACCURACY	,208	10	,200*	,880	10	,132
		5000	TF	ACCURACY	,204	10	,200*	,889	10	,167
			TF-IDF	ACCURACY	,176	10	,200*	,885	10	,148
	3	500	TF	ACCURACY	,198	10	,200*	,954	10	,711
			TF-IDF	ACCURACY	,179	10	,200*	,945	10	,614
		1000	TF	ACCURACY	,151	10	,200*	,984	10	,982
			TF-IDF	ACCURACY	,202	10	,200*	,940	10	,551
		5000	TF	ACCURACY	,202	10	,200*	,927	10	,415
			TF-IDF	ACCURACY	,147	10	,200*	,958	10	,764
	4	500	TF	ACCURACY	,166	10	,200*	,972	10	,909
			TF-IDF	ACCURACY	,176	10	,200*	,966	10	,855
		1000	TF	ACCURACY	,233	10	,134	,891	10	,174
			TF-IDF	ACCURACY	,156	10	,200*	,957	10	,756
		5000	TF	ACCURACY	,148	10	,200*	,952	10	,696
			TF-IDF	ACCURACY	,221	10	,182	,919	10	,347
	5	500	TF	ACCURACY	,163	10	,200*	,926	10	,408
			TF-IDF	ACCURACY	,134	10	,200*	,949	10	,660
		1000	TF	ACCURACY	,145	10	,200*	,984	10	,983
			TF-IDF	ACCURACY	,244	10	,094	,900	10	,219
		5000	TF	ACCURACY	,203	10	,200*	,944	10	,601
			TF-IDF	ACCURACY	,177	10	,200*	,940	10	,557

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.5: Prueba de normalidad para el Escenario 3 con un ANOVA de 4 factores

Pruebas de normalidad

			Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASSNGRAM	VOCABCARACT		Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES-1	1000-TF	ACCURACY	,217	10	,198	,914	10	,311
	1000-TF-IDF	ACCURACY	,176	10	,200*	,945	10	,615
	500-TF	ACCURACY	,175	10	,200*	,953	10	,701
	500-TF-IDF	ACCURACY	,136	10	,200*	,971	10	,901
	5000-TF	ACCURACY	,153	10	,200*	,974	10	,925
	5000-TF-IDF	ACCURACY	,104	10	,200*	,984	10	,984
NAIVE-BAYES-2	1000-TF	ACCURACY	,229	10	,148	,901	10	,224
	1000-TF-IDF	ACCURACY	,120	10	,200*	,955	10	,728
	500-TF	ACCURACY	,236	10	,123	,893	10	,182
	500-TF-IDF	ACCURACY	,285	10	,020	,822	10	,027
	5000-TF	ACCURACY	,209	10	,200*	,915	10	,315
	5000-TF-IDF	ACCURACY	,228	10	,149	,923	10	,378
NAIVE-BAYES-3	1000-TF	ACCURACY	,198	10	,200*	,877	10	,119
	1000-TF-IDF	ACCURACY	,182	10	,200*	,975	10	,932
	500-TF	ACCURACY	,206	10	,200*	,882	10	,136
	500-TF-IDF	ACCURACY	,214	10	,200*	,868	10	,094
	5000-TF	ACCURACY	,188	10	,200*	,937	10	,523
	5000-TF-IDF	ACCURACY	,188	10	,200*	,915	10	,314
NAIVE-BAYES-4	1000-TF	ACCURACY	,215	10	,200*	,838	10	,041
	1000-TF-IDF	ACCURACY	,187	10	,200*	,960	10	,787
	500-TF	ACCURACY	,195	10	,200*	,962	10	,804
	500-TF-IDF	ACCURACY	,158	10	,200*	,961	10	,797
	5000-TF	ACCURACY	,234	10	,129	,876	10	,119
	5000-TF-IDF	ACCURACY	,159	10	,200*	,912	10	,295
NAIVE-BAYES-5	1000-TF	ACCURACY	,141	10	,200*	,966	10	,849
	1000-TF-IDF	ACCURACY	,220	10	,188	,941	10	,561
	500-TF	ACCURACY	,123	10	,200*	,966	10	,848
	500-TF-IDF	ACCURACY	,223	10	,171	,812	10	,020
	5000-TF	ACCURACY	,191	10	,200*	,940	10	,555
	5000-TF-IDF	ACCURACY	,169	10	,200*	,947	10	,632
SVM-1	1000-TF	ACCURACY	,174	10	,200*	,890	10	,169
	1000-TF-IDF	ACCURACY	,122	10	,200*	,949	10	,653
	500-TF	ACCURACY	,180	10	,200*	,940	10	,554
	500-TF-IDF	ACCURACY	,141	10	,200*	,975	10	,935
	5000-TF	ACCURACY	,194	10	,200*	,917	10	,334
	5000-TF-IDF	ACCURACY	,187	10	,200*	,931	10	,454
SVM-2	1000-TF	ACCURACY	,200	10	,200*	,936	10	,506
	1000-TF-IDF	ACCURACY	,180	10	,200*	,947	10	,637
	500-TF	ACCURACY	,257	10	,060	,805	10	,016
	500-TF-IDF	ACCURACY	,176	10	,200*	,904	10	,241
	5000-TF	ACCURACY	,136	10	,200*	,953	10	,698
	5000-TF-IDF	ACCURACY	,225	10	,162	,883	10	,141
SVM-3	1000-TF	ACCURACY	,156	10	,200*	,947	10	,631
	1000-TF-IDF	ACCURACY	,113	10	,200*	,955	10	,724
	500-TF	ACCURACY	,170	10	,200*	,921	10	,362
	500-TF-IDF	ACCURACY	,172	10	,200*	,920	10	,356
	5000-TF	ACCURACY	,254	10	,066	,909	10	,276
	5000-TF-IDF	ACCURACY	,170	10	,200*	,895	10	,195
SVM-4	1000-TF	ACCURACY	,151	10	,200*	,944	10	,596
	1000-TF-IDF	ACCURACY	,153	10	,200*	,932	10	,470
	500-TF	ACCURACY	,213	10	,200*	,872	10	,106
	500-TF-IDF	ACCURACY	,110	10	,200*	,977	10	,950
	5000-TF	ACCURACY	,138	10	,200*	,981	10	,972
	5000-TF-IDF	ACCURACY	,170	10	,200*	,947	10	,634
SVM-5	1000-TF	ACCURACY	,222	10	,176	,896	10	,200
	1000-TF-IDF	ACCURACY	,200	10	,200*	,892	10	,180
	500-TF	ACCURACY	,247	10	,084	,911	10	,287
	500-TF-IDF	ACCURACY	,185	10	,200*	,940	10	,557
	5000-TF	ACCURACY	,195	10	,200*	,915	10	,314
	5000-TF-IDF	ACCURACY	,146	10	,200*	,971	10	,904
TREE-1	1000-TF	ACCURACY	,166	10	,200*	,940	10	,549
	1000-TF-IDF	ACCURACY	,225	10	,165	,910	10	,282
	500-TF	ACCURACY	,120	10	,200*	,950	10	,668
	500-TF-IDF	ACCURACY	,156	10	,200*	,952	10	,687
	5000-TF	ACCURACY	,140	10	,200*	,948	10	,647
	5000-TF-IDF	ACCURACY	,279	10	,027	,825	10	,029
TREE-2	1000-TF	ACCURACY	,173	10	,200*	,904	10	,240
	1000-TF-IDF	ACCURACY	,208	10	,200*	,880	10	,132
	500-TF	ACCURACY	,241	10	,102	,904	10	,245
	500-TF-IDF	ACCURACY	,152	10	,200*	,937	10	,523
	5000-TF	ACCURACY	,204	10	,200*	,889	10	,167
	5000-TF-IDF	ACCURACY	,176	10	,200*	,885	10	,148
TREE-3	1000-TF	ACCURACY	,151	10	,200*	,984	10	,982
	1000-TF-IDF	ACCURACY	,202	10	,200*	,940	10	,551
	500-TF	ACCURACY	,198	10	,200*	,954	10	,711
	500-TF-IDF	ACCURACY	,179	10	,200*	,945	10	,614
	5000-TF	ACCURACY	,202	10	,200*	,927	10	,415
	5000-TF-IDF	ACCURACY	,147	10	,200*	,958	10	,764
TREE-4	1000-TF	ACCURACY	,233	10	,134	,891	10	,174
	1000-TF-IDF	ACCURACY	,156	10	,200*	,957	10	,756
	500-TF	ACCURACY	,166	10	,200*	,972	10	,909
	500-TF-IDF	ACCURACY	,176	10	,200*	,966	10	,855
	5000-TF	ACCURACY	,148	10	,200*	,952	10	,696
	5000-TF-IDF	ACCURACY	,221	10	,182	,919	10	,347
TREE-5	1000-TF	ACCURACY	,145	10	,200*	,984	10	,983
	1000-TF-IDF	ACCURACY	,244	10	,094	,900	10	,219
	500-TF	ACCURACY	,163	10	,200*	,926	10	,408
	500-TF-IDF	ACCURACY	,134	10	,200*	,949	10	,660
	5000-TF	ACCURACY	,203	10	,200*	,944	10	,601
	5000-TF-IDF	ACCURACY	,177	10	,200*	,940	10	,557

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.6: Prueba de normalidad para el Escenario 3 con un ANOVA de 2 factores

Pruebas de normalidad

					Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASIFICADOR					Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES	1	500	TF	ACCURACY	,149	10	,200*	,946	10	,624
			TF-IDF	ACCURACY	,196	10	,200*	,959	10	,775
		1000	TF	ACCURACY	,188	10	,200*	,964	10	,828
			TF-IDF	ACCURACY	,253	10	,069	,904	10	,244
		5000	TF	ACCURACY	,144	10	,200*	,908	10	,268
			TF-IDF	ACCURACY	,196	10	,200*	,919	10	,350
	2	500	TF	ACCURACY	,186	10	,200*	,975	10	,932
			TF-IDF	ACCURACY	,079	10	,200*	,984	10	,984
		1000	TF	ACCURACY	,156	10	,200*	,967	10	,862
			TF-IDF	ACCURACY	,156	10	,200*	,967	10	,858
		5000	TF	ACCURACY	,152	10	,200*	,944	10	,594
			TF-IDF	ACCURACY	,186	10	,200*	,891	10	,173
	3	500	TF	ACCURACY	,255	10	,064	,921	10	,369
			TF-IDF	ACCURACY	,250	10	,076	,892	10	,180
		1000	TF	ACCURACY	,171	10	,200*	,918	10	,343
			TF-IDF	ACCURACY	,256	10	,062	,801	10	,015
		5000	TF	ACCURACY	,191	10	,200*	,961	10	,794
			TF-IDF	ACCURACY	,126	10	,200*	,960	10	,785
	4	500	TF	ACCURACY	,135	10	,200*	,949	10	,657
			TF-IDF	ACCURACY	,265	10	,045	,888	10	,162
		1000	TF	ACCURACY	,187	10	,200*	,953	10	,701
			TF-IDF	ACCURACY	,263	10	,049	,800	10	,015
		5000	TF	ACCURACY	,167	10	,200*	,961	10	,803
			TF-IDF	ACCURACY	,128	10	,200*	,938	10	,534
	5	500	TF	ACCURACY	,271	10	,036	,884	10	,143
			TF-IDF	ACCURACY	,144	10	,200*	,961	10	,800
		1000	TF	ACCURACY	,156	10	,200*	,934	10	,489
			TF-IDF	ACCURACY	,160	10	,200*	,934	10	,488
		5000	TF	ACCURACY	,250	10	,077	,862	10	,081
			TF-IDF	ACCURACY	,144	10	,200*	,977	10	,946
SVM	1	500	TF	ACCURACY	,263	10	,048	,806	10	,017
			TF-IDF	ACCURACY	,148	10	,200*	,958	10	,766
		1000	TF	ACCURACY	,232	10	,137	,945	10	,605
			TF-IDF	ACCURACY	,139	10	,200*	,972	10	,911
		5000	TF	ACCURACY	,159	10	,200*	,938	10	,531
			TF-IDF	ACCURACY	,199	10	,200*	,923	10	,381
	2	500	TF	ACCURACY	,231	10	,139	,904	10	,240
			TF-IDF	ACCURACY	,158	10	,200*	,930	10	,449
		1000	TF	ACCURACY	,199	10	,200*	,962	10	,813
			TF-IDF	ACCURACY	,159	10	,200*	,959	10	,777
		5000	TF	ACCURACY	,226	10	,161	,962	10	,814
			TF-IDF	ACCURACY	,174	10	,200*	,917	10	,330
	3	500	TF	ACCURACY	,182	10	,200*	,914	10	,308
			TF-IDF	ACCURACY	,229	10	,145	,941	10	,563
		1000	TF	ACCURACY	,125	10	,200*	,946	10	,618
			TF-IDF	ACCURACY	,213	10	,200*	,909	10	,272
		5000	TF	ACCURACY	,214	10	,200*	,855	10	,066
			TF-IDF	ACCURACY	,147	10	,200*	,974	10	,925
	4	500	TF	ACCURACY	,213	10	,200*	,921	10	,361
			TF-IDF	ACCURACY	,176	10	,200*	,916	10	,326
		1000	TF	ACCURACY	,143	10	,200*	,941	10	,563
			TF-IDF	ACCURACY	,172	10	,200*	,945	10	,608
		5000	TF	ACCURACY	,201	10	,200*	,947	10	,638
			TF-IDF	ACCURACY	,167	10	,200*	,947	10	,635
	5	500	TF	ACCURACY	,135	10	,200*	,985	10	,985
			TF-IDF	ACCURACY	,175	10	,200*	,913	10	,305
		1000	TF	ACCURACY	,209	10	,200*	,937	10	,518
			TF-IDF	ACCURACY	,139	10	,200*	,938	10	,533
		5000	TF	ACCURACY	,230	10	,142	,910	10	,278
			TF-IDF	ACCURACY	,216	10	,200*	,879	10	,128
TREE	1	500	TF	ACCURACY	,153	10	,200*	,947	10	,639
			TF-IDF	ACCURACY	,154	10	,200*	,936	10	,511
		1000	TF	ACCURACY	,246	10	,088	,905	10	,247
			TF-IDF	ACCURACY	,317	10	,005	,845	10	,051
		5000	TF	ACCURACY	,168	10	,200*	,942	10	,579
			TF-IDF	ACCURACY	,275	10	,031	,842	10	,047
	2	500	TF	ACCURACY	,157	10	,200*	,936	10	,510
			TF-IDF	ACCURACY	,144	10	,200*	,982	10	,976
		1000	TF	ACCURACY	,233	10	,133	,875	10	,114
			TF-IDF	ACCURACY	,263	10	,049	,912	10	,292
		5000	TF	ACCURACY	,246	10	,086	,895	10	,190
			TF-IDF	ACCURACY	,182	10	,200*	,922	10	,378
	3	500	TF	ACCURACY	,183	10	,200*	,948	10	,641
			TF-IDF	ACCURACY	,182	10	,200*	,880	10	,130
		1000	TF	ACCURACY	,200	10	,200*	,897	10	,203
			TF-IDF	ACCURACY	,135	10	,200*	,976	10	,938
		5000	TF	ACCURACY	,168	10	,200*	,899	10	,214
			TF-IDF	ACCURACY	,184	10	,200*	,897	10	,205
	4	500	TF	ACCURACY	,180	10	,200*	,931	10	,462
			TF-IDF	ACCURACY	,217	10	,198	,901	10	,225
		1000	TF	ACCURACY	,262	10	,051	,833	10	,037
			TF-IDF	ACCURACY	,240	10	,109	,923	10	,385
		5000	TF	ACCURACY	,148	10	,200*	,928	10	,428
			TF-IDF	ACCURACY	,173	10	,200*	,965	10	,840
	5	500	TF	ACCURACY	,195	10	,200*	,909	10	,277
			TF-IDF	ACCURACY	,225	10	,162	,843	10	,048
		1000	TF	ACCURACY	,184	10	,200*	,919	10	,345
			TF-IDF	ACCURACY	,186	10	,200*	,902	10	,230
		5000	TF	ACCURACY	,248	10	,082	,835	10	,038
			TF-IDF	ACCURACY	,181	10	,200*	,928	10	,433

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.7: Prueba de normalidad para el Escenario 4 con un ANOVA de 4 factores

Pruebas de normalidad

			Kolmogorov-Smirnov[a]			Shapiro-Wilk		
CLASSNGRAM	VOCABCARACT		Estadístico	gl	Sig.	Estadístico	gl	Sig.
NAIVE-BAYES-1	1000-TF	ACCURACY	,188	10	,200*	,964	10	,828
	1000-TF-IDF	ACCURACY	,253	10	,069	,904	10	,244
	500-TF	ACCURACY	,149	10	,200*	,946	10	,624
	500-TF-IDF	ACCURACY	,196	10	,200*	,959	10	,775
	5000-TF	ACCURACY	,144	10	,200*	,908	10	,268
	5000-TF-IDF	ACCURACY	,196	10	,200*	,919	10	,350
NAIVE-BAYES-2	1000-TF	ACCURACY	,156	10	,200*	,967	10	,862
	1000-TF-IDF	ACCURACY	,156	10	,200*	,967	10	,858
	500-TF	ACCURACY	,186	10	,200*	,975	10	,932
	500-TF-IDF	ACCURACY	,079	10	,200*	,984	10	,984
	5000-TF	ACCURACY	,152	10	,200*	,944	10	,594
	5000-TF-IDF	ACCURACY	,186	10	,200*	,891	10	,173
NAIVE-BAYES-3	1000-TF	ACCURACY	,171	10	,200*	,918	10	,343
	1000-TF-IDF	ACCURACY	,256	10	,062	,801	10	,015
	500-TF	ACCURACY	,255	10	,064	,921	10	,369
	500-TF-IDF	ACCURACY	,250	10	,076	,892	10	,180
	5000-TF	ACCURACY	,191	10	,200*	,961	10	,794
	5000-TF-IDF	ACCURACY	,126	10	,200*	,960	10	,785
NAIVE-BAYES-4	1000-TF	ACCURACY	,187	10	,200*	,953	10	,701
	1000-TF-IDF	ACCURACY	,263	10	,049	,800	10	,015
	500-TF	ACCURACY	,135	10	,200*	,949	10	,657
	500-TF-IDF	ACCURACY	,265	10	,045	,888	10	,162
	5000-TF	ACCURACY	,167	10	,200*	,961	10	,803
	5000-TF-IDF	ACCURACY	,128	10	,200*	,938	10	,534
NAIVE-BAYES-5	1000-TF	ACCURACY	,156	10	,200*	,934	10	,489
	1000-TF-IDF	ACCURACY	,160	10	,200*	,934	10	,488
	500-TF	ACCURACY	,271	10	,036	,884	10	,143
	500-TF-IDF	ACCURACY	,144	10	,200*	,961	10	,800
	5000-TF	ACCURACY	,250	10	,077	,862	10	,081
	5000-TF-IDF	ACCURACY	,144	10	,200*	,977	10	,946
SVM-1	1000-TF	ACCURACY	,232	10	,137	,945	10	,605
	1000-TF-IDF	ACCURACY	,139	10	,200*	,972	10	,911
	500-TF	ACCURACY	,263	10	,048	,806	10	,017
	500-TF-IDF	ACCURACY	,148	10	,200*	,958	10	,766
	5000-TF	ACCURACY	,159	10	,200*	,938	10	,531
	5000-TF-IDF	ACCURACY	,199	10	,200*	,923	10	,381
SVM-2	1000-TF	ACCURACY	,199	10	,200*	,962	10	,813
	1000-TF-IDF	ACCURACY	,159	10	,200*	,959	10	,777
	500-TF	ACCURACY	,231	10	,139	,904	10	,240
	500-TF-IDF	ACCURACY	,158	10	,200*	,930	10	,449
	5000-TF	ACCURACY	,226	10	,161	,962	10	,814
	5000-TF-IDF	ACCURACY	,174	10	,200*	,917	10	,330
SVM-3	1000-TF	ACCURACY	,125	10	,200*	,946	10	,618
	1000-TF-IDF	ACCURACY	,213	10	,200*	,909	10	,272
	500-TF	ACCURACY	,182	10	,200*	,914	10	,308
	500-TF-IDF	ACCURACY	,229	10	,145	,941	10	,563
	5000-TF	ACCURACY	,214	10	,200*	,855	10	,066
	5000-TF-IDF	ACCURACY	,147	10	,200*	,974	10	,925
SVM-4	1000-TF	ACCURACY	,143	10	,200*	,941	10	,563
	1000-TF-IDF	ACCURACY	,172	10	,200*	,945	10	,608
	500-TF	ACCURACY	,213	10	,200*	,921	10	,361
	500-TF-IDF	ACCURACY	,176	10	,200*	,916	10	,326
	5000-TF	ACCURACY	,201	10	,200*	,947	10	,638
	5000-TF-IDF	ACCURACY	,167	10	,200*	,947	10	,635
SVM-5	1000-TF	ACCURACY	,209	10	,200*	,937	10	,518
	1000-TF-IDF	ACCURACY	,139	10	,200*	,938	10	,533
	500-TF	ACCURACY	,135	10	,200*	,985	10	,985
	500-TF-IDF	ACCURACY	,175	10	,200*	,913	10	,305
	5000-TF	ACCURACY	,230	10	,142	,910	10	,278
	5000-TF-IDF	ACCURACY	,216	10	,200*	,879	10	,128
TREE-1	1000-TF	ACCURACY	,246	10	,088	,905	10	,247
	1000-TF-IDF	ACCURACY	,317	10	,005	,845	10	,051
	500-TF	ACCURACY	,153	10	,200*	,947	10	,639
	500-TF-IDF	ACCURACY	,154	10	,200*	,936	10	,511
	5000-TF	ACCURACY	,168	10	,200*	,942	10	,579
	5000-TF-IDF	ACCURACY	,275	10	,031	,842	10	,047
TREE-2	1000-TF	ACCURACY	,233	10	,133	,875	10	,114
	1000-TF-IDF	ACCURACY	,263	10	,049	,912	10	,292
	500-TF	ACCURACY	,157	10	,200*	,936	10	,510
	500-TF-IDF	ACCURACY	,144	10	,200*	,982	10	,976
	5000-TF	ACCURACY	,246	10	,086	,895	10	,190
	5000-TF-IDF	ACCURACY	,182	10	,200*	,922	10	,378
TREE-3	1000-TF	ACCURACY	,200	10	,200*	,897	10	,203
	1000-TF-IDF	ACCURACY	,135	10	,200*	,976	10	,938
	500-TF	ACCURACY	,183	10	,200*	,948	10	,641
	500-TF-IDF	ACCURACY	,182	10	,200*	,880	10	,130
	5000-TF	ACCURACY	,168	10	,200*	,899	10	,214
	5000-TF-IDF	ACCURACY	,184	10	,200*	,897	10	,205
TREE-4	1000-TF	ACCURACY	,262	10	,051	,833	10	,037
	1000-TF-IDF	ACCURACY	,240	10	,109	,923	10	,385
	500-TF	ACCURACY	,180	10	,200*	,931	10	,462
	500-TF-IDF	ACCURACY	,217	10	,198	,901	10	,225
	5000-TF	ACCURACY	,148	10	,200*	,928	10	,428
	5000-TF-IDF	ACCURACY	,173	10	,200*	,965	10	,840
TREE-5	1000-TF	ACCURACY	,184	10	,200*	,919	10	,345
	1000-TF-IDF	ACCURACY	,186	10	,200*	,902	10	,230
	500-TF	ACCURACY	,195	10	,200*	,909	10	,277
	500-TF-IDF	ACCURACY	,225	10	,162	,843	10	,048
	5000-TF	ACCURACY	,248	10	,082	,835	10	,038
	5000-TF-IDF	ACCURACY	,181	10	,200*	,928	10	,433

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Figura B.8: Prueba de normalidad para el Escenario 4 con un ANOVA de 2 factores

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: ACCURACY

F	df1	df2	Sig.
1,119	89	810	,223

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASIFICADOR + NGRAMAS + VOCABULARIO + CARACTERISTICAS + CLASIFICADOR * NGRAMAS + CLASIFICADOR * VOCABULARIO + CLASIFICADOR * CARACTERISTICAS + NGRAMAS * VOCABULARIO + NGRAMAS * CARACTERISTICAS + VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO + CLASIFICADOR * NGRAMAS * CARACTERISTICAS + CLASIFICADOR * VOCABULARIO * CARACTERISTICAS + NGRAMAS * VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO * CARACTERISTICAS

(a) ANOVA de 4 factores

Prueba de igualdad de Levene de varianzas de error[a]

ACCURACY

F	df1	df2	Sig.
1,119	89	810	,223

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASNGRAM + VOCABCARACT + CLASNGRAM * VOCABCARACT

(b) ANOVA de 2 factores

Figura B.9: Prueba de homogeneidad para el Escenario 1

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: ACCURACY

F	df1	df2	Sig.
1,307	89	810	,036

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASIFICADOR + NGRAMAS + VOCABULARIO + CARACTERISTICAS + CLASIFICADOR * NGRAMAS + CLASIFICADOR * VOCABULARIO + CLASIFICADOR * CARACTERISTICAS + NGRAMAS * VOCABULARIO + NGRAMAS * CARACTERISTICAS + VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO + CLASIFICADOR * NGRAMAS * CARACTERISTICAS + CLASIFICADOR * VOCABULARIO * CARACTERISTICAS + NGRAMAS * VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO * CARACTERISTICAS

(a) ANOVA de 4 factores

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: ACCURACY

F	df1	df2	Sig.
1,307	89	810	,036

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASNGRAM + VOCABCARACT + CLASNGRAM * VOCABCARACT

(b) ANOVA de 2 factores

Figura B.10: Prueba de homogeneidad para el Escenario 2

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: ACCURACY

F	df1	df2	Sig.
1,366	89	810	,018

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASIFICADOR + NGRAMAS + VOCABULARIO + CARACTERISTICAS + CLASIFICADOR * NGRAMAS + CLASIFICADOR * VOCABULARIO + CLASIFICADOR * CARACTERISTICAS + NGRAMAS * VOCABULARIO + NGRAMAS * CARACTERISTICAS + VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO + CLASIFICADOR * NGRAMAS * CARACTERISTICAS + CLASIFICADOR * VOCABULARIO * CARACTERISTICAS + NGRAMAS * VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO * CARACTERISTICAS

(a) ANOVA de 4 factores

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: ACCURACY

F	df1	df2	Sig.
1,366	89	810	,018

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASSNGRAM + VOCABCARACT + CLASSNGRAM * VOCABCARACT

(b) ANOVA de 2 factores

Figura B.11: Prueba de homogeneidad para el Escenario 3

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: ACCURACY

F	df1	df2	Sig.
1,417	89	810	,009

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASIFICADOR + NGRAMAS + VOCABULARIO + CARACTERISTICAS + CLASIFICADOR * NGRAMAS + CLASIFICADOR * VOCABULARIO + CLASIFICADOR * CARACTERISTICAS + NGRAMAS * VOCABULARIO + NGRAMAS * CARACTERISTICAS + VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO + CLASIFICADOR * NGRAMAS * CARACTERISTICAS + CLASIFICADOR * VOCABULARIO * CARACTERISTICAS + NGRAMAS * VOCABULARIO * CARACTERISTICAS + CLASIFICADOR * NGRAMAS * VOCABULARIO * CARACTERISTICAS

(a) ANOVA de 4 factores

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: ACCURACY

F	df1	df2	Sig.
1,417	89	810	,009

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Interceptación + CLASSNGRAM + VOCABCARACT + CLASSNGRAM * VOCABCARACT

(b) ANOVA de 2 factores

Figura B.12: Prueba de homogeneidad para el Escenario 4

Pruebas de efectos inter-sujetos

ACCURACY

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.
Modelo corregido	1,363[a]	89	,015	10,612	,000
Interceptación	260,829	1	260,829	180702,586	,000
CLASNGRAM	1,262	14	,090	62,472	,000
VOCABCARACT	,048	5	,010	6,588	,000
CLASNGRAM * VOCABCARACT	,053	70	,001	,527	,999
Error	1,169	810	,001		
Total	263,362	900			
Total corregido	2,532	899			

a. R al cuadrado = ,538 (R al cuadrado ajustada = ,488)

Figura B.13: Modelo ANOVA para el Escenario 1 con 2 factores

ACCURACY

HSD Tukey[a,b]

CLASNGRAM	N	Subconjunto 1	2	3	4	5	6	7	8	9
SVM-5	60	,4856250								
TREE-5	60	,4910417	,4910417							
TREE-4	60	,4951042	,4951042	,4951042						
SVM-4	60	,5037500	,5037500	,5037500						
NAIVE-BAYES-4	60		,5129167	,5129167	,5129167					
NAIVE-BAYES-5	60			,5164583	,5164583	,5164583				
TREE-3	60				,5308333	,5308333	,5308333			
NAIVE-BAYES-3	60				,5341667	,5341667	,5341667			
SVM-3	60					,5397917	,5397917			
TREE-2	60						,5480208	,5480208		
SVM-2	60							,5661458	,5661458	
NAIVE-BAYES-2	60							,5705208	,5705208	
TREE-1	60								,5719792	
NAIVE-BAYES-1	60									,6033333
SVM-1	60									,6054167
Sig.		,363	,104	,127	,132	,056	,458	,080	1,000	1,000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,001.
a. Utiliza el tamaño de la muestra de la media armónica = 60,000.
b. Alfa = ,05.

Figura B.14: Pruebas post-hoc para el Escenario 1 con el factor CLASNGRAM

ACCURACY

HSD Tukey[a,b]

VOCABCARACT	N	Subconjunto 1	2	3
500-TF	150	,5295417		
500-TF-IDF	150	,5330000	,5330000	
1000-TF	150	,5352500	,5352500	
1000-TF-IDF	150	,5364583	,5364583	
5000-TF-IDF	150		,5451250	,5451250
5000-TF	150			,5506667
Sig.		,614	,064	,805

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,001.

a. Utiliza el tamaño de la muestra de la media armónica = 150,000.

b. Alfa = ,05.

Figura B.15: Pruebas post-hoc para el Escenario 1 con el factor VOCABCARACT

Pruebas de efectos inter-sujetos

Variable dependiente: ACCURACY

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.
Modelo corregido	2,490[a]	89	,028	42,553	,000
Interceptación	89,316	1	89,316	135850,673	,000
CLASNGRAM	2,404	14	,172	261,189	,000
VOCABCARACT	,043	5	,009	13,090	,000
CLASNGRAM * VOCABCARACT	,043	70	,001	,930	,640
Error	,533	810	,001		
Total	92,339	900			
Total corregido	3,022	899			

a. R al cuadrado = ,824 (R al cuadrado ajustada = ,804)

Figura B.16: Modelo ANOVA para el Escenario 2 con 2 factores

ACCURACY

HSD Tukey[a,b]

CLASNGRAM	N	Subconjunto 1	2	3	4	5	6
SVM-5	60	,25109375					
TREE-5	60	,25218750					
SVM-4	60	,25817708	,25817708				
TREE-4	60	,25947917	,25947917				
NAIVE-BAYES-5	60	,26473958	,26473958				
NAIVE-BAYES-4	60		,27343750				
TREE-3	60			,31046875			
NAIVE-BAYES-3	60			,31395833			
SVM-3	60			,31562500			
NAIVE-BAYES-2	60				,34963542		
SVM-2	60				,35671875		
TREE-2	60				,35729167		
TREE-1	60				,36484375		
NAIVE-BAYES-1	60					,38947917	
SVM-1	60						,40822917
Sig.		,192	,077	,999	,079	1,000	1,000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,001.
a. Utiliza el tamaño de la muestra de la media armónica = 60,000.

Figura B.17: Pruebas post-hoc para el Escenario 2 con el factor CLASNGRAM

ACCURACY

HSD Tukey[a,b]

VOCABCARACT	N	Subconjunto 1	2	3
500-TF-IDF	150	,30850000		
1000-TF-IDF	150	,30947917		
500-TF	150	,30958333		
1000-TF	150	,31429167	,31429167	
5000-TF-IDF	150		,32122917	,32122917
5000-TF	150			,32706250
Sig.		,369	,178	,360

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,001.
a. Utiliza el tamaño de la muestra de la media armónica = 150,000.
b. Alfa = ,05.

Figura B.18: Pruebas post-hoc para el Escenario 2 con el factor VOCABCARACT

Pruebas de efectos inter-sujetos

Variable dependiente: ACCURACY

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.
Modelo corregido	3,388[a]	89	,038	78,369	,000
Interceptación	70,577	1	70,577	145281,572	,000
CLASSNGRAM	3,180	14	,227	467,609	,000
VOCABCARACT	,106	5	,021	43,547	,000
CLASSNGRAM * VOCABCARACT	,102	70	,001	3,009	,000
Error	,393	810	,000		
Total	74,359	900			
Total corregido	3,782	899			

a. R al cuadrado = ,896 (R al cuadrado ajustada = ,885)

Figura B.19: Modelo ANOVA para el Escenario 3 con 2 factores

ACCURACY

HSD Tukey[a,b]

CLASSNGRAM	N	Subconjunto 1	2	3	4	5	6	7	8
TREE-5	60	,205792							
SVM-5	60	,206417							
NAIVE-BAYES-5	60	,211125	,211125						
TREE-4	60		,223125	,223125					
SVM-4	60			,227875					
NAIVE-BAYES-4	60			,228875					
NAIVE-BAYES-3	60				,272833				
TREE-3	60				,273542				
SVM-3	60				,281083				
TREE-1	60					,321542			
NAIVE-BAYES-2	60					,325708			
TREE-2	60					,327458			
SVM-2	60						,344250		
NAIVE-BAYES-1	60							,363875	
SVM-1	60								,387000
Sig.		,992	,163	,985	,767	,980	1,000	1,000	1,000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,000.
a. Utiliza el tamaño de la muestra de la media armónica = 60,000.

Figura B.20: Pruebas post-hoc para el Escenario 3 con el factor CLASNGRAM

ACCURACY

HSD Tukey[a,b]

VOCABCARACT	N	Subconjunto 1	2	3	4
500-TF-IDF	150	,267133			
500-TF	150	,270050			
1000-TF	150		,277383		
1000-TF-IDF	150		,277767		
5000-TF-IDF	150			,289433	
5000-TF	150				,298433
Sig.		,862	1,000	1,000	1,000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,000.
a. Utiliza el tamaño de la muestra de la media armónica = 150,000.
b. Alfa = ,05.

Figura B.21: Pruebas post-hoc para el Escenario 3 con el factor VOCABCARACT

Pruebas de efectos inter-sujetos

Variable dependiente: ACCURACY

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.
Modelo corregido	7,565[a]	89	,085	586,746	,000
Interceptación	34,791	1	34,791	240171,863	,000
CLASSNGRAM	7,003	14	,500	3453,173	,000
VOCABCARACT	,377	5	,075	520,652	,000
CLASSNGRAM * VOCABCARACT	,184	70	,003	18,182	,000
Error	,117	810	,000		
Total	42,473	900			
Total corregido	7,682	899			

a. R al cuadrado = ,985 (R al cuadrado ajustada = ,983)

Figura B.22: Modelo ANOVA para el Escenario 4 con 2 factores

ACCURACY

HSD Tukey[a,b]

CLASSNGRAM	N	Subconjunto 1	2	3	4	5	6	7	8	9	10	11
TREE-5	60	,091944444444418										
SVM-5	60	,093906250000015										
NAIVE-BAYES-5	60		,101961805555517									
TREE-4	60			,120763888888922								
NAIVE-BAYES-4	60			,121215277777862								
SVM-4	60			,123107638888878								
NAIVE-BAYES-3	60				,171874999999957							
TREE-3	60					,182083333333367						
SVM-3	60						,190607638888917					
NAIVE-BAYES-2	60							,227395833333267				
TREE-2	60								,251579861111133			
SVM-2	60									,286319444444400		
TREE-1	60										,301406250000000	
NAIVE-BAYES-1	60										,303871527777717	
SVM-1	60											,381145833333383
Sig.		1,000	1,000	,999	1,000	1,000	1,000	1,000	1,000	1,000	,999	1,000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,000.
a. Utiliza el tamaño de la muestra de la media armónica = 60,000.
b. Alfa = ,05.

Figura B.23: Pruebas post-hoc para el Escenario 4 con el factor CLASNGRAM

ACCURACY

HSD Tukey[a,b]

VOCABCARACT	N	Subconjunto 1	2	3	4
500-TF-IDF	150	,174750000000054			
500-TF	150	,174840277777786			
1000-TF-IDF	150		,189798611111101		
1000-TF	150		,192805555555529		
5000-TF-IDF	150			,219909722222217	
5000-TF	150				,227569444444418
Sig.		1,000	,256	1,000	1,000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.
Se basa en las medias observadas.
El término de error es la media cuadrática(Error) = ,000.

a. Utiliza el tamaño de la muestra de la media armónica = 150,000.

b. Alfa = ,05.

Figura B.24: Pruebas post-hoc para el Escenario 4 con el factor VOCABCARACT