13 de agosto de 2018.

EXAMEN GENERAL DE CONOCIMIENTOS

PARA OBTENER EL GRADO DE MAESTRO EN CIENCIAS

ING. BRIAN ALFONSO SÁNCHEZ VÁZQUEZ

1. RESUMEN

Para este trabajo se presenta la tarea de identificación de un texto anónimo de la Época de la masacre estudiantil de 1968, “El Móndrigo”, supuestamente una bitácora del movimiento de la época.

Se presumen 5 posibles autores: Jorge Joseph, Emilio Uranga, Gregorio Ortega Hernández, Gregorio Ortega Martínez y Roberto Blanco Moheno.

Se hace frente al problema como un problema clásico de clasificación de textos, en este caso a la identificación de autoría.

Para esto se hace uso de Support Vector Machines, mediante la clasificación multiclase OVR (one vs rest) haciendo una predicción para cada tipo de kernel y utilizando los vectores de características que provee la herramienta SAUTEE.

2. INTRODUCCIÓN

Para poder lograr la identificación del o de los autores del texto en cuestión se tienen que considerar rasgos de estilo del escritor, se intenta hacer una clasificación de los textos proporcionados basándonos marcadores estilométricos, entre los cuales destacan características tales como signos de puntuación, n-gramas, distribución y patrones de clases semánticas y sintácticas, arboles de análisis, métricas de complejidad y riqueza de vocabulario, e incluso características del discurso. [1]

Los métodos clásicos funcionan razonablemente bien cuando se tiene una gran cantidad de textos disponibles, y en este caso cuando hay unos pocos candidatos posibles, y se conoce al autor verdadero. [1]

Se considera en [1] que la representación vectorial de características como n-gramas tiene cierta ventaja con respecto a otras formas de representación más complejas en pruebas estilométricas, y esto es por una buena razón, ya que compensa el tamaño reducido de características con la cantidad información que contiene.

En este ejercicio de atribución de autoría se tiene un pequeño grupo de candidatos posibles y una cantidad limitada de texto para entrenar cada clase, lo cual no es lo ideal, pero es lo que se tiene, y dado que el enfoque de machine learning ha dado resultados alentadores, usaremos SVM, ya que nos permite usar la representación vectorial de características que conservan la mayor parte de información útil [2].

3. METODOLOGÍA

Apoyándonos en la herramienta SAUUTE y el la cual utiliza un servidor FREELING para realizar el etiquetado automático para obtener los vectores de características utilizados.
Para el ejercicio utilizamos los siguientes marcadores estilométricos:

Para poder hacer la clasificación de cada con cada uno de los marcadores estilométricos, se utilizó la herramienta scikit-learn, la cual también incluye LIBSVM, con scripts hechos en Matlab y Python.

México resistió a la C.I.A. y a Fidel Castro	Gregorio Ortega Hernández	1968
Un chamaco de Tepito da una lección de civismo	Gregorio Ortega Hernández	1968
Un injusto artículo de Blanco Moheno	Gregorio Ortega Hernández	1968
El Móndrigo	Anónimo	1968
La noticia detrás de la noticia 2	Roberto Blanco Moheno	1966
La noticia detrás de la noticia 3	Roberto Blanco Moheno	1966
Picaluga Vende a los Rojo Gomistas por una Curul Senatorial	Jorge Joseph	1961
Pitoloco en Palacio	Jorge Joseph	1961
Aviso	Jorge Joseph	1961

Pruebas de la existencia de la Atlántida	Jorge Joseph	1965
Esoterismo en Anahuac	Jorge Joseph	1965
¿Hacia la extrema derecha?	Gregorio Ortega Molina	1968
La guerra de los zombies	Gregorio Ortega Molina	1968
La sociedad del átomo	Gregorio Ortega Molina	1968
Universidad Popular	Emilio Uranga	1968
La Represión Extremada	Emilio Uranga	1968
La Ambigüedad Universitaria	Emilio Uranga	1968
Ornato y Orden	Emilio Uranga	1968
Viaje del Canciller	Emilio Uranga	1968
Prólogo de Astucias literarias	Emilio Uranga	1971
De Astucias literarias 1	Emilio Uranga	1971
De Astucias literarias 2	Emilio Uranga	1971
Lectura de galeras de De Astucias literarias	Emilio Uranga	1971
Tlatelolco. Historia de una infamia	Roberto Blanco Moheno	1969
La noticia detrás de la noticia 1	Roberto Blanco Moheno	1966

i. Para verificar este método de clasificación, al tener 24 textos de autores conocidos, se utilizaron 12 como entrenamiento para el modelo.

ii. Se obtuvieron con ayuda de SAUTEE los vectores característicos de cada texto.

Con ayuda de SAUTEE se seleccionaron 15 marcadores estilométricos:

ID_MARCADOR	MARCADOR
BC	Bigramas de caracteres
TC	Trigramas de caracteres
UPF	Unigramas de palabras funcionales
BPF	Bigramas de palabras funcionales
BPF2H	Bigramas de palabras funcionales con hasta 2 huecos
TPF	Trigramas de palabras funcionales
TPF2H	Trigramas de palabras funcionales con hasta 2 huecos
LO	Longitud de oraciones
LP	Longitud de palabras
UPOS	Unigramas de etiquetas POS
BPOS	Bigramas de etiquetas POS
TPOS	Trigramas de etiquetas POS
CGIO	Categoría gramatical al inicio de la oración
CGFO	Categoría gramatical al final de la oración
UPOSNF	Unigramas de etiquetas POS no fino

iii. Se seleccionaron aleatoriamente 12 elementos como entrenamiento y el resto como datos de prueba o validacón.

iv. Se usó validación cruzada para obtener los mejores parámetros de kernel para el clasificador y generar el modelo optimo para estimar la probabilidad de pertenencia a cada clase de los vectores de los marcadores asociados al elemento desconocido.

Para realizar el entrenamiento del set de datos de entrenamiento, se hizo uso de SVC (Support Vector Clasiffier), la implementación de LIBSVM en Matlab que utiliza SVM implementando el paradigma OVR (One vs. Rest), que es en si una clasificación binaria que hace pares con todos los elementos del conjunto, esto para minimizar errores de clasificación y obtener los mejores parámetros para la generación de un modelo.
Los parámetros que buscamos con este método son :

K	Función de Kernel, se contemplan 4, Lineal, Polinomial, RBF y Sigmoide.
C	Es la función de costo, a menor costo, se tiene cierta holgura para hallar un hiperplano de mejor ajuste.
γ	Es el coeficiente de la función del Kernel
d	Degree o grado de la función del Kernel, para el Kernel polinomial, irrelevante para las demás.

v. Se corrieron 100 iteraciones por función de Kernel y marcador estilométrico, se contemplaron 4 funciones de Kernel : lineal, polinomial, rbf y sigmoidal, se contemparon también para hallar el grado de ajuste de la funcion de kernel polinomial del grado 3 al grado 7.

vi. En el caso que hubiese empate, se consideró el kernel que tuviera una función de costo menor, esto para evitar la penalizacion y se tuvieran el mayor número de vectores de soporte.

vii. Una vez obtenidos los mejores parámetros para el modelo del clasificador, se corrieron dichos parámetros con el kernel:

	LINEAL [0]	POLI [1]	RBF [2]	SIGMOIDE[3]
BC	83.3333%, : C=8.7241, gamma=0.15328 d=3	83.3333%, : C=32, gamma=0.014148,d=3	83.3333%, : C=279.17, gamma=0.021819	83.3333%, : C=181.0193, gamma=0.19035
TC	83.3333%, : C=76.1093, gamma=0.007831	75%, : C=5.6569, gamma=0.00015868	91.6667%, : C=76.1093, gamma=0.028724	83.3333%, : C=181.0193, gamma=0.0013843
UPF	75%, : C=32, gamma=0.010989	83.3333%, : C=32, gamma=0.026136, d=3	75%, : C=279.17, gamma=0.00034341 d=4	75%, : C=430.539, gamma=0.00014438

BPF	75%, : C=76.1093, gamma=0.011101	83.3333%, : C=76.1093, gamma=0.0019624	75%, : C=430.539, gamma=0.000535,	83.3333%, : C=13.4543, gamma=0.00082508
BPF2H	58.3333%, : C=5.6569, gamma=0.07891	66.6667%, : C=5.6569, gamma=0.013949 d=3	75%, : C=181.0193, gamma=0.013949	83.3333%, : C=76.1093, gamma=0.0024659
TPF	66.6667%, : C=5.6569, gamma=0.0080469	66.6667%, : C=13.4543, gamma=0.0052177 d=3	66.6667%, : C=76.1093, gamma=0.0033833	66.6667%, : C=181.0193, gamma=0.0033833
TPF2H	66.6667%, : C=13.4543, gamma=0.018572	66.6667%, : C=32, gamma=0.0078087, d=3	83.3333%, : C=181.0193, gamma=0.0050633	66.6667%, : C=117.3765, gamma=0.0050633
LO	66.6667%, : C=1, gamma=0.0052083	75%, : C=20.7494, gamma=0.0021898, d=3	75%, : C=32, gamma=0.0052083	75%, : C=13.4543, gamma=0.00092071
LP	75%, : C=5.6569, gamma=0.05	83.3333%, : C=49.3507, gamma=0.021022 d=4	83.3333%, : C=20.7494, gamma=0.013631	83.3333%, : C=76.1093, gamma=0.021022
UPOS	83.3333%, : C=8.7241, gamma=0.024595	83.3333%, : C=32, gamma=0.024595, d=4	83.3333%, : C=49.3507, gamma=0.008327	75%, : C=0.013139, gamma=0.010341,
BPOS	83.3333%, : C=32, gamma=0.0012452	91.6667%, : C=13.4543, gamma=0.00022012, d=4	75%, : C=32, gamma=0.0012452	75%, : C=1, gamma=0.00022012
TPOS	75%, : C=181.0193, gamma=0.040529	75%, : C=181.0012, gamma=0.0071645, d=3	58.3333%, : C=32, gamma=0.0019532	58.3333%, : C=117.3765, gamma=0.0012665
CGIO	75%, : C=117.3765, gamma=3.3116	83.3333%, : C=13.4543, gamma=1.7291, d=3	83.3333%, : C=32, gamma=0.083333	83.3333%, : C=2.3784, gamma=6.3424
CGFO	75%, : C=2.3784, gamma=0.1, d=3	75%, : C=20.7494, gamma=3.2, d=4	75%, : C=13.4543, gamma=0.23784	75%, : C=5.6569, gamma=3.2
UPOSNF	83.3333%, : C=76.1093,	83.3333%, : C=181.0193,	83.3333%, : C=2.3784,	75%, : C=5.6569, gamma=0.090909

	gamma=0.0028409	gamma=0.00050221, d=3	gamma=0.016071

De manera intuitiva se seleccionaron aquellas funciones y parámetros de Kernel que presentaba un grado mayor de precisión, para cada marcador estilo métrico y en caso de que todas las funciones de kernel tuvieran un empate, se selecciona aquella con una función de penalización C para que el clasificador tenga una mayor holgura al buscar el mejor hiperplano de separación.

4. RESULTADOS

Se hizo un ranking con los mejores resultados de las funciones de kernel, identificando en el entrenamiento el que obtuviera la mayor presición.

Tabla de predicción de autoria basada en SVM y marcadores estilométricos.

ME	K	ORTEGA MOLINA	EMILIO URANGA	BLANCO MOHENO	ORTEGA HERNANDEZ	JORGE JOSEPH
BC	LIN	10.61%	3.94%	1.82%	14.01%	69.62%
TC	RBF	8.28%	9.43%	4.25%	11.72%	66.31%
UPF	POLY	10.05%	8.85%	6.90%	13.30%	60.91%
BPF	SIG	16.74%	17.45%	12.10%	14.24%	39.48%
BPF2H	SIG	9.67%	24.36%	14.59%	9.19%	42.19%
TPF	LIN	9.02%	29.94%	11.41%	10.36%	39.26%
TPF2H	RBF	9.16%	35.30%	15.41%	17.47%	22.66%
LO	SIG	14.70%	42.58%	8.80%	12.37%	21.55%
LP	RBF	9.40%	18.15%	3.58%	14.15%	54.72%
UPOS	LIN	15.36%	3.43%	7.54%	14.02%	59.65%
BPOS	POLY	19.59%	13.22%	13.96%	16.44%	36.79%
TPOS	POLY	32.08%	8.08%	14.42%	29.59%	15.83%
CGIO	SIG	6.73%	44.50%	9.35%	19.41%	20.01%
CGFO	LIN	13.10%	36.80%	5.99%	17.28%	26.83%
UPOSNF	RBF	13.04%	12.51%	15.63%	12.12%	46.71%

5. CONCLUSIONES

Dadas las características que se presentan en las tablas de resultados, podemos concluir que es muy posible que la limitada cantidad de datos nos esté provocando un sesgo, ya que la máxima precisión para la mayoría de los modelos es del 83.3%, a excepción del caso en donde se utilizó como vectores de características aquello generados con el marcador trigramas de caracteres y un kernel RBF, dando una confiabilidad del 91.6%.

Basámdonos en la tabla de resultados, es altamente probable que el texto "El Móndrigo" haya sido escrito por Jorge Joseph, con alguna contribución de Emilio Uranga, ya que según la predicción, la distribución de palabras funcionales, se acercan en cierta medida a él, así como las Categorías Gramaticales al Final e Inicio de la Oración.

Podemos concluir que un entrenamiento mas extensivo con una mayor cantidad de muestras de texto, y mejor balanceadas, y tal vez buscando una combinación de marcadores estilométricos podrían ayudarnos en esta tarea.

Sería importante realizar el ejercicio de identificación de autoria con una colección de textos más extensa y en donde se tenga una plena identificación del candidato.

6. REFERENCIAS

[1] Daelemans, W. (2013).
Explanation in computational stylometry.
In Gelbukh, A., editor, Computational Linguistics and Intelligent Text Processing, volume 7817 of Lecture Notes in Computer Science, pages 451–462. Springer Berlin Heidelberg.

[2] Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60(1), 9–26.

[3] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[4] Authorship Attribution with Support Vector Machines. Applied Intelligence, 2003, Volume 19, Number 1-2, Page 109. Joachim Diederich, Jörg Kindermann, Edda Leopold, Gerhard Paass.

[5] Se adaptaron funciones Matlab escritas por Kittipat "Bot" Kampa, Integrated Brain Imaging Center, UW Medical Center, Seattle, UW.

[6] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.