Identificación de la Variedad del Lenguaje para la Mejora del Geoposicionamineto en Social Media

Autor: Raül Fabra Boluda

Directores: Dr. Paolo Rosso, Universitat Politècnica de València
Dr. Francisco Rangel, Autoritas Consulting

Trabajo de Final de Máster en Inteligencia Artificial,
Reconocimiento de Formas e Imagen Digital

30 de septiembre de 2016

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Índice I

Medios Sociales y Geoposicionamiento

Medios sociales:

- Plataformas *online* orientadas al intercambio de información entre sus usuarios.
- Se encuentran en plena proliferación.

Problema del geoposicionamiento:

- Fronteras geográficas desdibujadas.
- Apenas un 2 % de los usuarios georreferencia su contenido.

Aplicaciones:

Marketing

Segmentación geográfica o demográfica de las opiniones de los usuarios al lanzar un nuevo producto.

Seguridad y lingüística forense

Tratar de identificar rasgos del autor de una amenaza a partir de mensajes sospechosos.

Geoposicionamineto mediante Identificación de la Variedad del Lenguaje

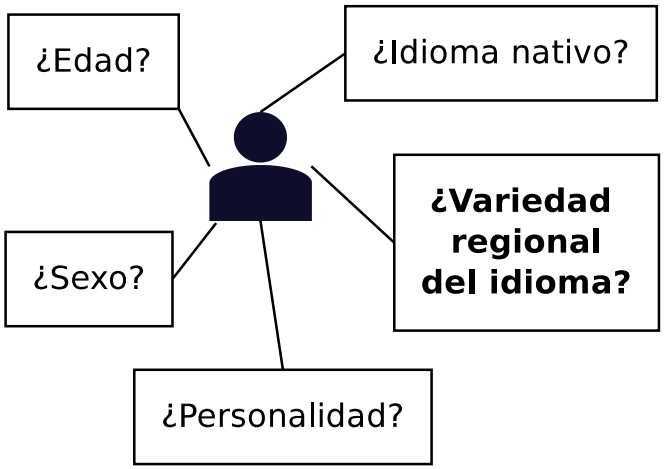
Objetivos:

- Geoposicionamiento mediante la identificación de la variedad del lenguaje.
- Metodología para la construcción de corpus anotados con la variedad del idioma, garantizando la ausencia de sobreajustes por autor.
- Construcción y evaluación de un corpus para identificación de variedades del español en Twitter.

Author Profiling

Author Profiling:

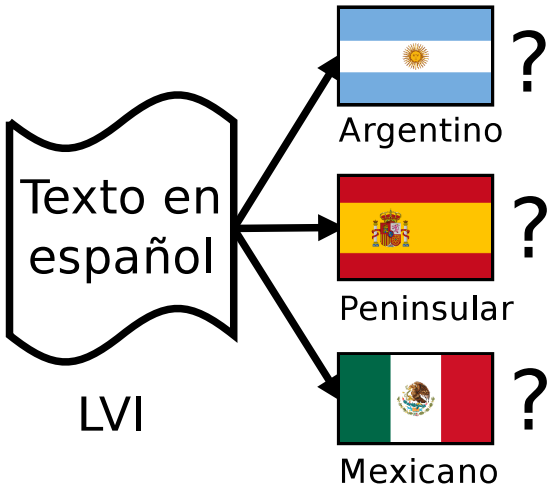
Los textos de un individuo reflejan sus creencias e influencias socioculturales y educacionales.

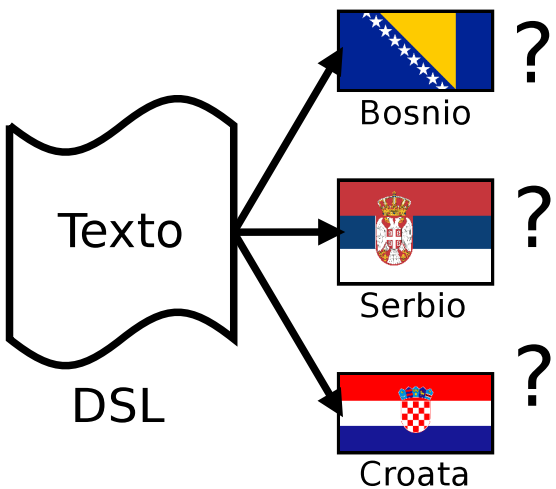


Tareas organizadas en los últimos años:

- PAN 2013/14/15/16: reconocimiento de la edad, el sexo, la personalidad en Twitter y a nivel cross-genre.
- myPersonality: reconocimiento de la personalidad.
- PR-SOCO: reconocimiento de la personalidad de los programadores a partir de sus códigos fuentes.
- NLI 2013: identificación del idioma nativo.
- DSL 2015: discriminación de idiomas similares (DSL) y variedades del idioma (LVI).

LVI y DSL





Características	LVI	DSL
n-gramas (caracteres)	X	X
n-gramas (palabras)	X	X
n-gramas (sílabas)	X	

Clasificación	LVI	DSL
SVM	X	X
Modelos de lenguaje	X	
Modelos Ocultos de Markov	X	
Máxima Entropía	X	X
Naïve Bayes	X	X
Árboles de Decisión		X

Escasez de Recursos para LVI/DSL

DSLCC v.2.0

No se separan por autor los textos en los conjuntos de entrenamiento y test.

Training
252.000

Development
28.000

Test*
14.000

*Dos conjuntos de test, idénticos pero uno de ellos sin Entidades Nombradas.

Grupo	Idioma/ Variedad
Eslavos del sureste	Búlgaro Macedonio
Español	Argentino Peninsular
Portugués	Brasileño Europeo
Eslavos de suroeste	Bosnio Croata Serbio
Austronesios	Indonesio Malayo
Eslavos del oeste	Checo Eslovaco
Otros	

HispaBlogs

 Argentina

 Chile

 España

 México

 Perú

Entrenamiento
2.250 blogueros
(450 por país)

Test
1.000 blogueros
(200 por país)

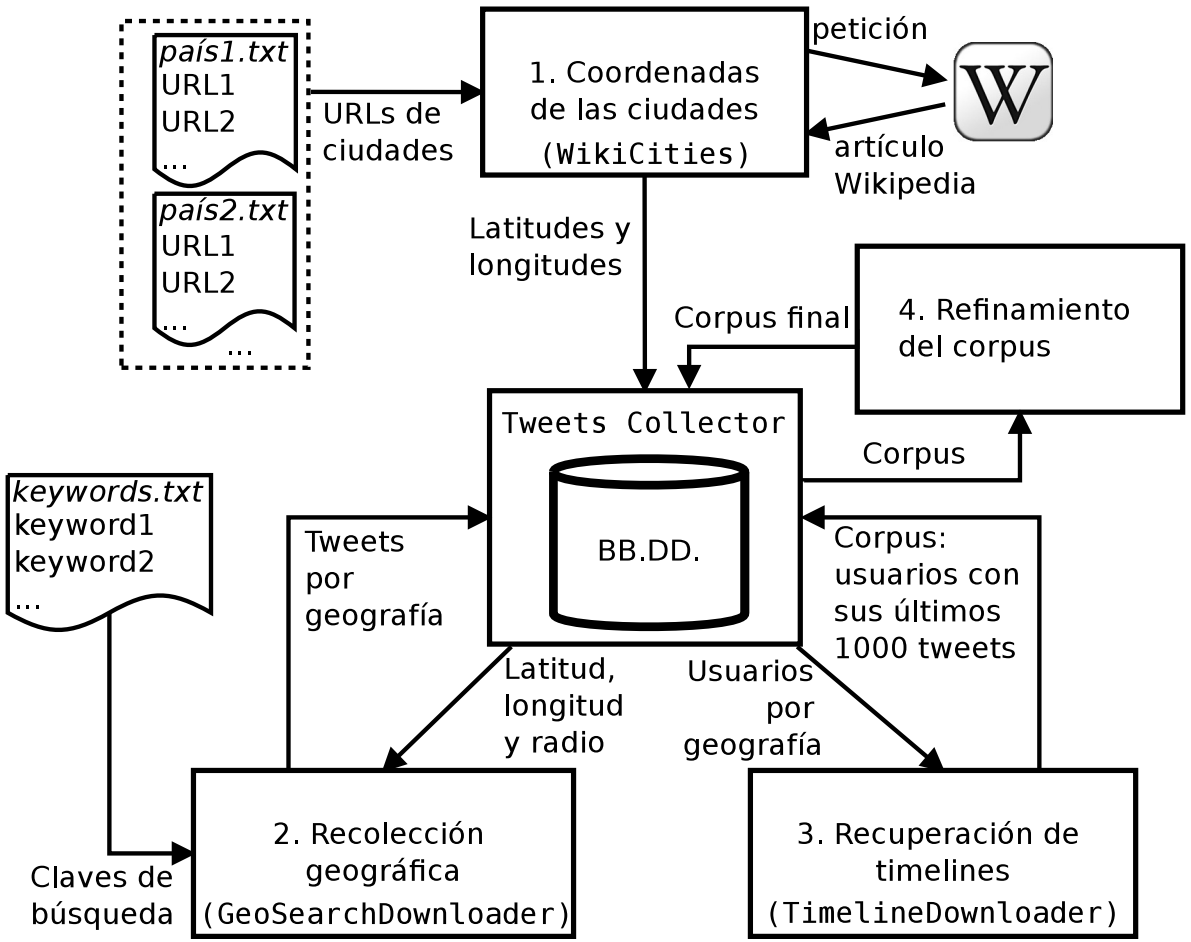
Los autores aparecen sólo en uno de los dos conjuntos

Tipología específica de textos

Metodología para la Construcción de Corpus para LVI

1. Metodología y sistema desarrollado.
2. Corpus en bruto.
3. Refinamientos: filtrado geográfico, temporal y por frecuencia.
4. Corpus final: HispaTweets

Descripción de la Metodologia y el Sistema



Software liberado: https://github.com/autoritas/RD-Lab/tree/master/src/

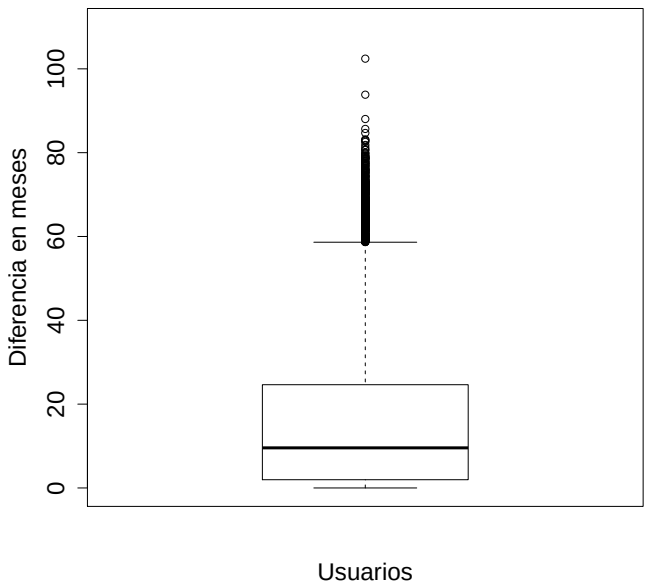
Corpus tras el Proceso de Descarga

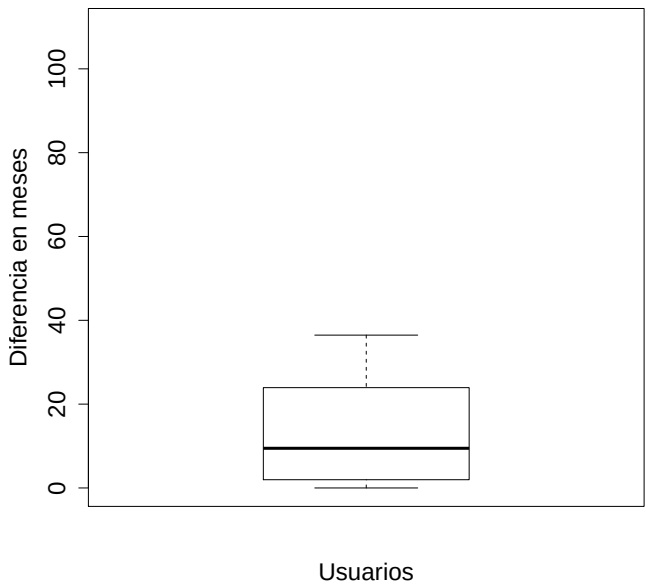
País	Usuarios	Tweets (*timelines*)	Tweets/ usuario	Longitud media Palabras	Caracteres
Argentina	2.393	1.684.662	704,00	10,72	65,18
Chile	1.378	1.059.724	769,03	11,71	77,45
Colombia	1.719	1.084.965	631,16	12,13	79,96
España	1.909	1.148.523	601,64	12,81	84,12
México	2.684	1.571.859	585,64	11,82	76,87
Perú	1.113	676.623	607,93	11,34	72,87
Venezuela	1.400	715.987	511,42	13,55	90,33
Total	12.596	7.942.343	630,88	11,87	76,81
Media	1.799,44	1.134.620,43	630,12	12,01	78,11
SDev	529,37	355.970,97	**77,58**	0,87	7,42

Corpus obtenido tras la búsqueda geolocalizada y la descarga de las *timelines*.

Refinamientos

1. Filtrado geográfico: prescindimos de ciudades fronterizas. Proceso manual.
2. Filtrado temporal: conservamos tweets entre: 1 de enero de 2013 y 1 de enero de 2016





3. Filtrado por frecuencia: conservamos aquellos usuarios con más de 500 tweets.

Corpus Final

País	Usuarios	Tweets (*timelines*)	Tweets/ usuario	Longitud media Palabras	Caracteres
Argentina	650	566.113	870,94	10,63	64,32
Chile	650	569.190	875,68	11,71	77,51
Colombia	650	559.619	860,95	12,17	80,34
España	650	558.253	858,85	12,72	83,91
México	650	567.734	873,44	11,75	76,42
Perú	650	564.203	868,00	11,26	72,35
Venezuela	650	552.978	850,74	13,54	90,94
Total	4.550	3.938.090	865,51	11,96	77,91
Media	650	562.584,29	865,51	11,97	77,97
SDev	0,00	5.412,43	**8,33**	0,89	7,83

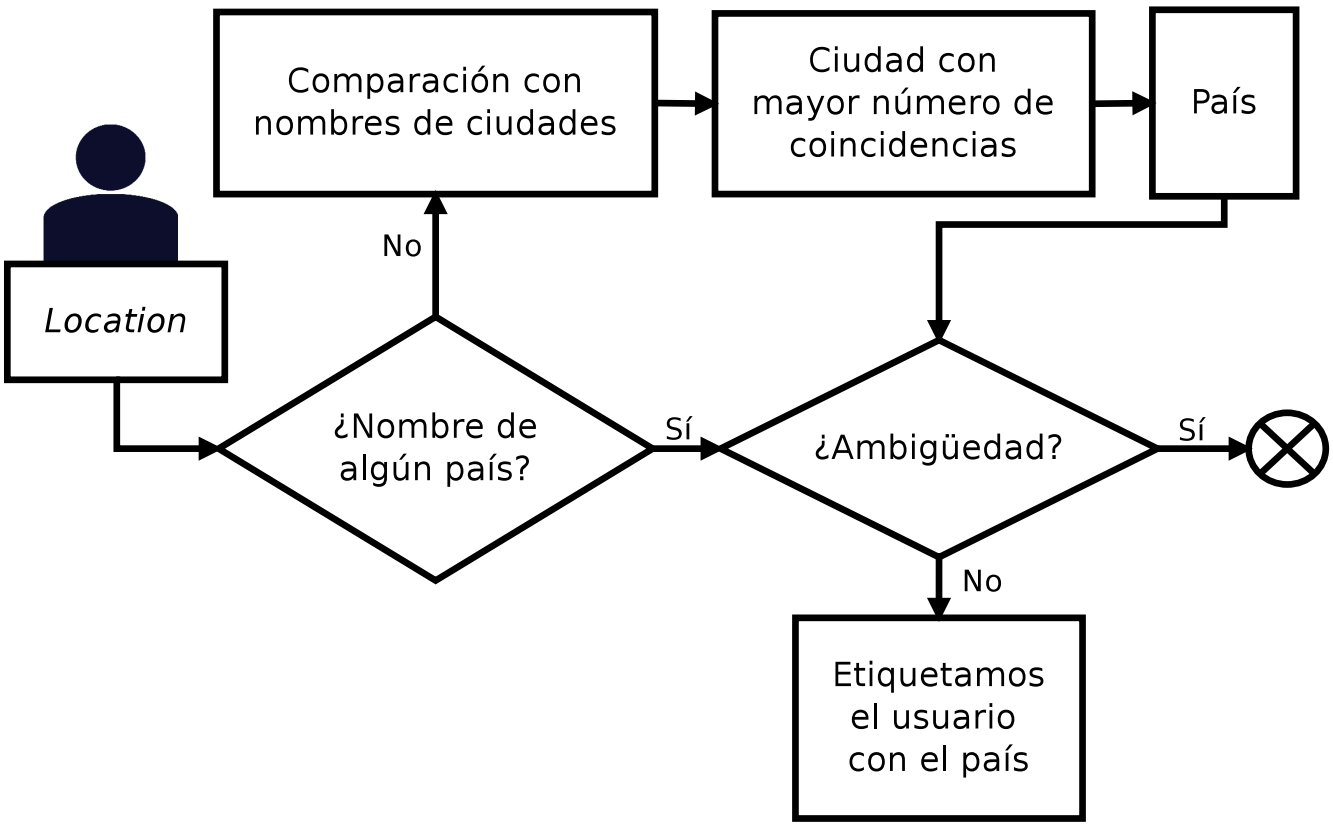
Corpus final tras los tres filtrados.

Marco de Evaluación y Resultados Experimentales

1. Algoritmo de localización por perfil.
   - Utiliza únicamente información disponible en perfil del usuario.
2. Técnicas para LVI
   - Basadas el análisis de los textos con técnicas del estado del arte.

Localización por Perfil

Evaluación de HispaTweets con la información disponible en el perfil del usuario. Referencia para comparar con técnicas de LVI.



LVI: Representaciones

TF: Cada usuario es representado por un vector con la frecuencia de los términos que utiliza.

TF-IDF: Cada usuario es representado por un vector de pesos. Cada peso indica la relevancia de su término asociado en el documento (usuario).

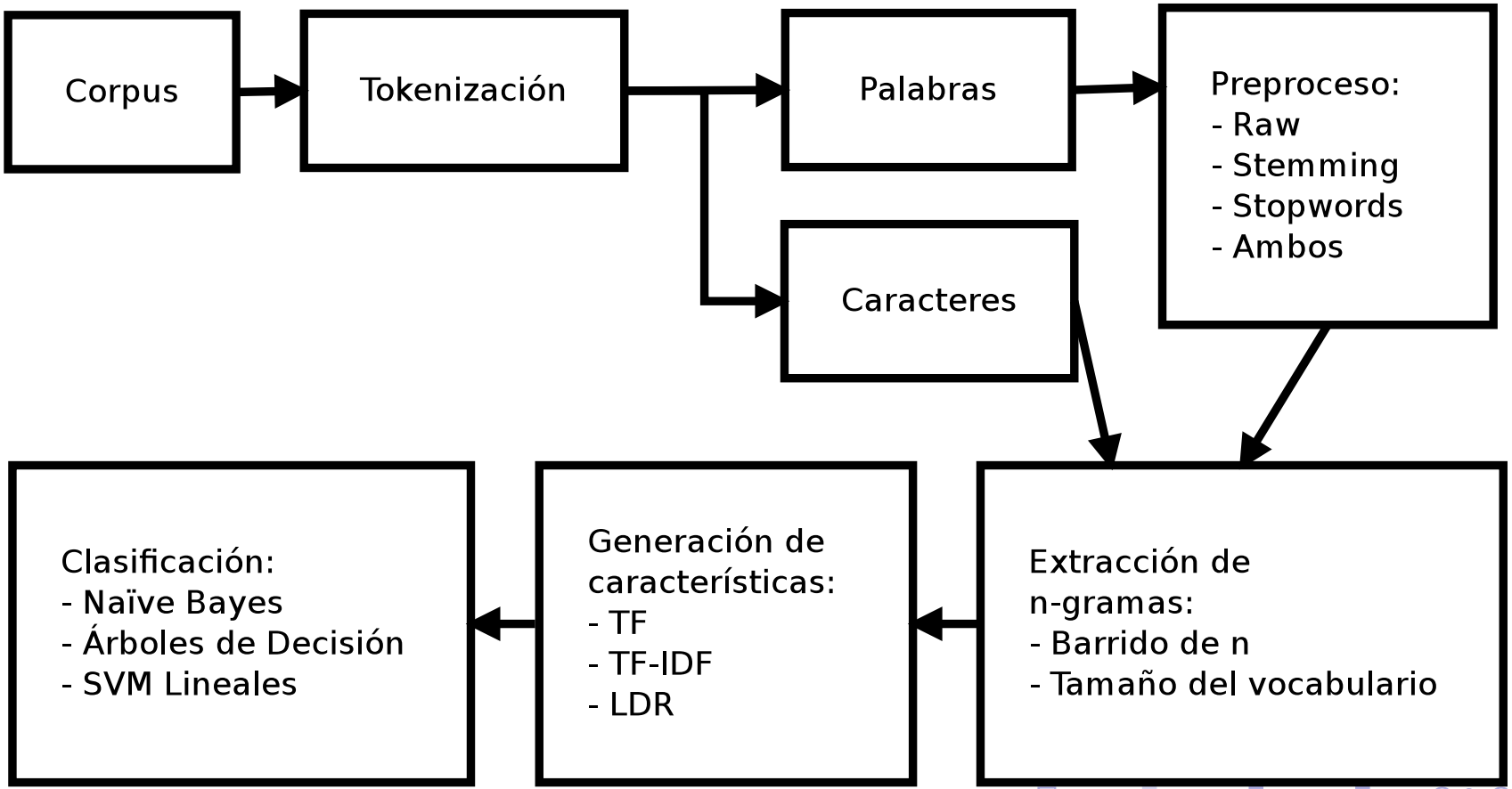
LDR o *Low-Dimensionality Representation* [Rangel et al., 2016]:

1. Cálculo de la matriz TF-IDF.
2. Ponderado de términos dependientes de las clases.
   - Para cada clase $c \in C$ (variedad) y cada término $t \in T$, se calcula una puntuación que indica con que confianza el término t pertenece a la clase c.
3. Representación dependiente de las clases.
   - Para cada clase, se calculan 6 medidas estadísticas en base a las puntuaciones anteriores.
     - avg, std, min, max, prob y prop
   - El número de características con qué representamos cada usuario es $6 \times |C|$.

F. Rangel, M. Franco-Salvador, and P. Rosso, "A low dimensionality representation for language variety identification," in 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2016, Springer-Verlag, LNCS(), (2016).

Marco Experimental

Evaluación con validación cruzada en 5 bloques, asegurando siempre la separación de los usuarios de entrenamiento y validación durante todo el proceso.



Resultados del Algoritmo de Localización por Perfil

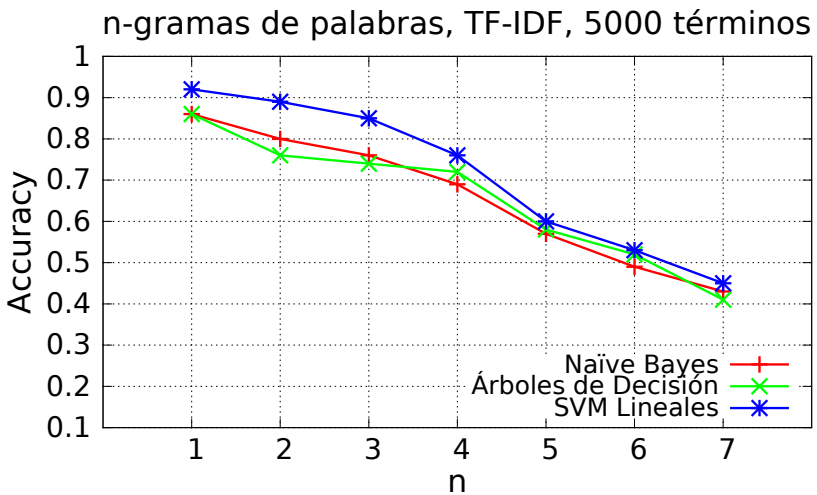
Predicción	Usuarios	% Usuarios
Correcta	2.731	60,02
Incorrecta	129	2,84
Indefinible	1.690	37,14
Total	4.550	100,00

Considerando únicamente usuarios etiquetados (62,86 % del total)

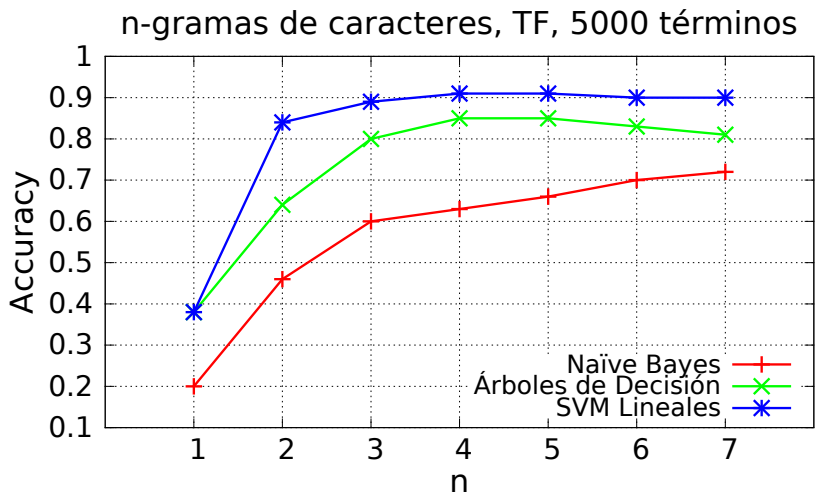
Accuracy: 0.95

País	Número usuarios	*Precision*	*Recall*	*F-score*
Argentina	358	0,91	0,96	0,93
Chile	403	0,97	0,97	0,97
Colombia	474	0,98	0,96	0,97
España	335	0,95	0,94	0,94
México	387	0,93	0,97	0,95
Perú	470	0,97	0,91	0,94
Venezuela	433	0,96	0,99	0,97
Total	2.860	0,96	0,95	0,95

Mejores Resultados con *n*-gramas



Mejor resultado: 92 % con 1-gramas de palabras y SVM lineales.



Mejor resultado: 91 % con 4-gramas de caracteres y SVM lineales.

HispaTweets vs HispaBlogs

Accuracy HispaTweets (HT): 0,96
Accuracy HispaBlogs (HB): 0,71

País	*Precision*		*Recall*		*F-score*	
	HT	HB	HT	HB	HT	HB
Argentina	0,95	0,66	0,96	0,72	0,96	0,69
Chile	0,97	0,67	0,96	0,76	0,97	0,71
Colombia	0,95	-	0,95	-	0,95	-
España	0,95	0,71	0,96	0,77	0,95	0,74
México	0,95	0,78	0,98	0,66	0,96	0,71
Perú	0,98	0,77	0,90	0,66	0,94	0,71
Venezuela	0,94	-	0,98	-	0,96	-
Total	0,96	0,71	0,96	0,71	0,96	0,71

Resultados alineados para HispaTweets (HT) e HispaBlogs (HB).
LDR sobre uni-gramas de palabras y SVM.

Participación en la Tarea de Discriminación de Idiomas Similares

1. Descripción y motivación de la tarea.
2. Aproximaciones.
3. Resultados experimentales.

Tarea DSL 2015

- Tarea de discriminación de idiomas similares y variedades del idioma (DSL 2015).
- Recortes de noticias (entre 20 y 100 tokens).
- Los textos de los diferentes conjuntos no están separados por autor.

DSLCC v.2.0

Training
252.000

Development
28.000

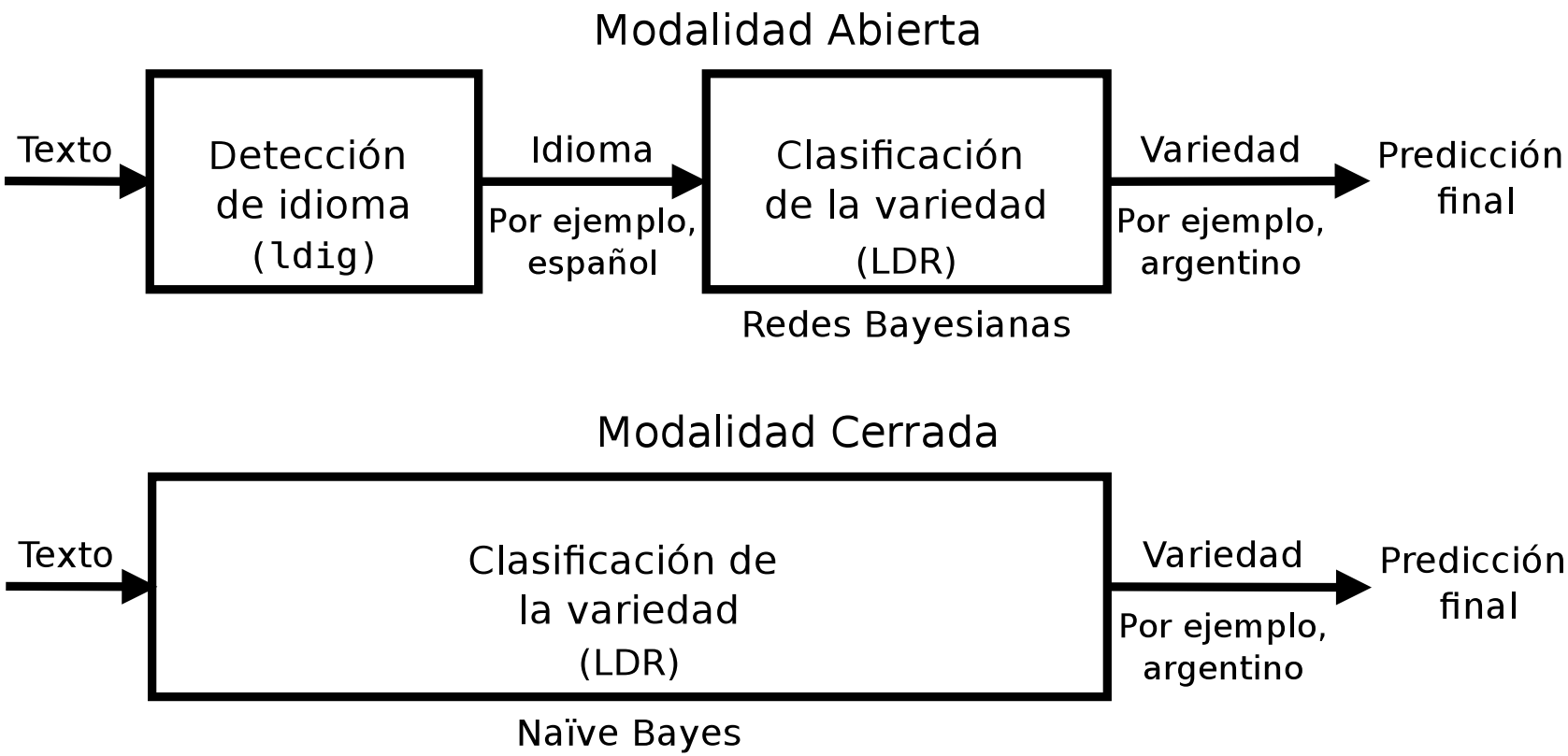
Test*
14.000

*Dos conjuntos de test, idénticos pero uno de ellos sin Entidades Nombradas.

Grupo	**Idioma/ Variedad**
Eslavos del sureste	Búlgaro Macedonio
Español	Argentino Peninsular
Portugués	Brasileño Europeo
Eslavos de suroeste	Bosnio Croata Serbio
Austronesios	Indonesio Malayo
Eslavos del oeste	Checo Eslovaco
Otros	

Aproximación

Objetivo: evaluación del rendimiento con LDR en dos sistemas diferentes.



Resultados

Variedad	Accuracy		
	Validación	Test A	Test B
bg*	**99,80**	**99,90**	**99,80**
mk*	**100,00**	**99,90**	**100,00**
es-ES	88,00	84,70	79,50
es-AR*	87,50	88,00	87,70
pt-PT	88,60	87,40	94,00
pt-BR	90,10	90,03	68,50
bs*	78,35	78,00	74,40
hr*	86,15	85,80	85,40
sr**	86,40	86,40	82,70
id	99,40	99,40	92,90
my*	99,45	99,20	99,50
cz*	99,70	99,80	99,40
sk*	99,60	99,30	99,60
xx*	**99,90**	**99,90**	**99,70**
global	93,07	92,71	90,22

Modalidad abierta.

Variedad	Accuracy		
	Validación	Test A	Test B
bg	98,15	97,50	95,10
mk*	98,95	98,20	98,20
es-ES	87,55	84,80	48,70
es-AR**	67,05	70,00	74,10
pt-PT	82,15	81,20	58,30
pt-BR	72,45	72,50	65,90
bs	55,70	54,30	86,20
hr	80,85	78,88	13,10
sr	74,40	74,70	7,80
id	97,75	97,60	92,00
my	94,25	93,60	97,60
cz	98,45	98,40	94,40
sk	98,80	97,60	79,30
xx*	98,55	98,50	98,80
global	86,08	85,57	72,11

Modalidad cerrada.

¿Es posible que exista sobreajuste por autor?

Conclusiones: Localización por Perfil *vs.* LVI

Algoritmo de localización por perfil:

- 60,02 % de los usuarios en total etiquetados correctamente.
- Considerando usuarios etiquetados, acierta el 95 % de las veces.

Obtenemos un *accuracy* del 96 % con LDR sobre uni-gramas de palabras, SVM lineales y con un vocabulario de 5.000 términos.

- Aumento de n en los n-gramas de palabras empeoran el resultado
- n-gramas de caracteres funcionan mejor con $n \geq 4$.

Conclusiones: HispaTweets *vs.* HispaBlogs

Comparativa de HispaTweets con HispaBlogs:

- Diferencia entre *accuracies*: en HispaTweets disponemos de mucha más información por usuario.
- Perú: el clasificador no suele equivocarse (alta *precision*), pero le cuesta más detectarlos (bajo *recall*).
- En ambos corpus, Chile, España y en menor medida Argentina presentan un *recall* alto (más fáciles de detectar).

Conclusiones: Tarea DSL 2015

Participación en la tarea DSL 2015 [Fabra et al., 2015]:

- El sistema en dos pasos obtiene mejores resultados.
- Los grupos más complicados de diferenciar han sido:
  - Bosnio, serbio y croata (*accuracy* medio de 83,63 %).
  - Español: argentino y peninsular (*accuracy* medio de 87,75 %).
  - Portugués: de Portugal y Brasil (*accuracy* medio de 89,35 %).
- Accuracies superiores al 99 % para el resto de idiomas/variedades.

Fabra R., Rangel F., Rosso P. NLEL_UPV_Autoritas Participation at Discrimination between Similar Languages (DSL) 2015 Shared Task. In: Proc. of the RANLP Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), Hissar, Bulgaria, 10 September, pp. 52-58

Trabajo Futuro

Para un entorno de producción en la empresa hay que tener en cuenta la aplicabilidad.

- Necesidad de obtención de conocimiento en tiempo real.
- Reducción de la cantidad de información por usuario:
  - Identificación a nivel de tweet en lugar de usuario.
  - Limitar la cantidad de tweets por usuario en el entrenamiento.

Aplicación de la metodología para la construcción de corpus para otros idiomas/variedades.

- Para la tarea PAN en la conferencia CLEF 2017 se elaborarán corpus para portugués, inglés y árabe.