Aplicaciones de la Lingüística Computacional
Detección de robots en Twitter

Jesús Alonso
Máster IARFID

2 de julio de 2014

1. Introducción

Twitter es una popular red social y herramienta de microblogging lanzada en 2006. Su naturaleza abierta y su elevado número de usuarios han provocado que sea objetivo de programas automatizados conocidos como robots o bots. Al igual que los existentes en otro tipo de aplicaciones web (chats, blogs, juegos online, etc.), en Twitter es común su presencia.
La abundancia de estos programas en Twitter tiene dos caras: por un lado existen robots legítimos que generan gran cantidad de contenido benigno, como por ejemplo noticias y actualizaciones de blogs, lo que contribuye a la meta de Twitter de convertirse en una red de noticias e información. Por otro lado, no es desdeñable la presencia de otro tipo de robots más maliciosos y dedicados a difundir spam.
En este contexto, es interesante el desarrollo de técnicas para distinguir automáticamente a los robots de los humanos en esta red social. En esta memoria se describe una experimentación sencilla realizada con este objetivo y se comentan los resultados obtenidos.

2. Trabajo previo

El estudio más completo que se ha encontrado sobre detección de robots en Twitter es el que se describe en [1]. En él se define un sistema de clasificación de usuarios de Twitter entre humanos, cyborgs y robots, basándose en tres tipos de características:

- Componente de entropía: Detecta la regularidad de la frecuencia de tuiteo. Por regla general, los robots escriben más regularmente que los

humanos.

- Componente de detección de spam: Se basa en que la gran mayoría de los tuits que contienen spam son publicados por robots, por lo que la presencia de patrones de spam indica automatización.
- Componente de propiedades de la cuenta: Considera las siguientes propiedades: ratio de URLs, método de tuiteo (web, aplicación móvil, API,...), ratio seguidores / amigos, seguridad de los enlaces, verificación de la cuenta, fecha de registro, ratio de hashtags y ratio de menciones.

Mediante estas características se construye un bosque automático, que es el encargado de clasificar las cuentas de Twitter. Los resultados son prometedores, detectando correctamente más del 90 % de las muestras de cada uno de los grupos considerados.

3. Planteamiento del proyecto

Se propone en este proyecto una aproximación alternativa al problema de detección de robots, más sencilla que la descrita en el apartado anterior. Se toma como baseline un clasificador basado en bolsas de palabras de los contenidos de los tuits y Support Vector Machines. Posteriormente, con el ánimo de mejorar los resultados, se han complementado o sustituido las bolsas de palabras con otro tipo de características.
Los datos utilizados en el proyecto han sido proporcionados por la consultora *Autoritas* [1]. Consisten de 1052956 tuits, correspondientes a 231189 cuentas, divididos en 5 proyectos de la manera que muestra la tabla 1:

Proyecto	Tuits	Usuarios	Robots	Porcentaje robots
gasofa	339734	135167	50	0.04 %
castellon	1824	1178	86	7.30 %
senado	64014	29018	622	2.14 %
comite	323692	122296	393	0.32 %
rato	323692	122296	108	0.09 %

Tabla 1: Composición de los conjuntos de datos de los 5 proyectos

Se ha realizado el estudio de manera separada cada proyecto, con la metodología explicada a continuación.

[1] `http://www.autoritas.net/`

Las características utilizadas han sido las siguientes:

- Bolsas de palabras del contenido de los tuits. Las palabras utilizadas han sido las 400 más comunes (eliminando stopwords) en los tuits de cada proyecto.
- Bolsas de palabras de las biografías. En este caso, se han empleado únicamente las 20 palabras más comunes (igualmente eliminando stopwords) en las biografías de las cuentas de cada proyecto.
- Características adicionales. Son las siguientes:
  - Número de tuits
  - Ratio de URLs: nº de tuits que contienen URLs / nº total de tuits de la cuenta
  - Ratio de hashtags: nº de hashtags en los tuits / nº total de tuits de la cuenta
  - Ratio de menciones: nº de menciones en los tuits / nº total de tuits de la cuenta
  - Ratio amigos / seguidores
  - Nº de días desde el registro de la cuenta

Las muestras de cada proyecto se han dividido aleatoriamente en conjuntos de entrenamiento y test. Puesto que en determinados proyectos el número de robots es muy reducido, en comparación con el total de cuentas disponibles, se ha optado en estos casos por eliminar cuentas de humanos de los conjuntos de entrenamiento correspondientes, manteniendo una ratio aproximada de 25 humanos por robot. La tabla 2 detalla la composición de los conjuntos de entrenamiento y test para cada experimento.

Proyecto	Entrenamiento			Test			Total
	Total	Humanos	Robots	Total	Humanos	Robots	
gasofa	1040	1000	40	35127	35117	10	135167
castellon	858	792	66	320	300	20	1178
senado	12472	12000	472	8546	8396	150	29018
comite	10293	10000	293	32003	31903	100	122296
rato	2588	2500	88	32208	32188	20	122296

Tabla 2: Composición de los conjuntos de entrenamiento y test de cada uno de los experimentos realizados

Se han construido clasificadores mediante Support Vector Machines, utilizando el software *LibSVM* tanto para su entrenamiento como para la clasificación de las muestras de test. En particular, se ha empleado el script `easy.py`, distribuido con el toolkit, para el ajuste óptimo de los parámetros del modelo. Se han considerado los siguientes grupos de características, realizando experimentos separados para cada uno de ellos:

- Bolsas de palabras de los contenidos
- Bolsas de palabras de los contenidos + Bolsas de palabras de las biografías
- Bolsas de palabras de los contenidos + Características adicionales
- Bolsas de palabras de los contenidos + Bolsas de palabras de las biografías + Características adicionales
- Sólo características adicionales

Posteriormente a esta experimentación, se han realizado validaciones cruzadas consistentes en entrenar SVMs con los datos de un proyecto y evaluar la clasificación de los datos de los cuatro proyectos restantes. Para ello se han utilizado los 5 conjuntos de características definidos anteriormente. De nuevo, debido a la menor presencia de cuentas robots en los conjuntos de datos, se han eliminado aleatoriamente cuentas de humanos de los conjuntos de entrenamiento, quedando éstos distribuidos como muestra la tabla 3

Proyecto	Entrenamiento		
	Total	Humanos	Robots
gasofa	1300	1250	50
castellon	1178	1092	86
senado	16172	15550	622
comite	10218	9825	393
rato	2808	2700	108

Tabla 3: Composición de los conjuntos de entrenamiento en los experimentos de validación cruzada

Los conjuntos de test han correspondido en cada caso a la totalidad de los datos de los cuatro proyectos no considerados en el entrenamiento.

4. Resultados

Se han empleado las siguientes medidas de evaluación:

- Precision: mide el porcentaje de robots detectados que corresponden a robots reales. Se calcula de la siguiente manera:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Recall: mide el porcentaje de robots reales que el clasificador ha detectado correctamente. Se calcula de la siguiente manera:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- F-Measure: da una visión global de la bondad del sistema como algo equilibrado entre las medidas de Precision y Recall. Se calcula como la media armónica entre estas dos medidas, por tanto, su expresión es:

$$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

En cuanto a las validaciones cruzadas, los resultados se muestran en media para cada uno de los cinco conjuntos de características considerados.
Los resultados de los experimentos separados por proyectos se muestran en la tabla 4 y los de las validaciones cruzadas se muestran en la tabla 5.

Proyecto	Características	TP	FP	FN	Precision	Recall	F-Measure
gasofa	Cont.	10	0	0	100 %	100 %	1
	Cont. + bio.	9	1	1	90 %	90 %	0.9
	Cont. + adic.	9	0	1	100 %	90 %	0.95
	Cont. + bio. + adic.	9	0	1	100 %	90 %	0.95
	Adic.	10	8	0	55.56 %	100 %	0.71
castellon	Cont.	19	0	1	100 %	95 %	0.97
	Cont. + bio.	20	0	0	100 %	100 %	1
	Cont. + adic.	20	0	0	100 %	100 %	1
	Cont. + bio. + adic.	20	0	0	100 %	100 %	1
	Adic.	5	3	15	62.5 %	25 %	0.36
senado	Cont.	95	25	55	79.17 %	63.33 %	0.7
	Cont. + bio.	92	29	58	76.03 %	61.33 %	0.68
	Cont. + adic.	91	41	59	68.94 %	60.67 %	0.65
	Cont. + bio. + adic.	98	24	52	80.33 %	65.33 %	0.72
	Adic.	51	23	99	68.92 %	34 %	0.46
comite	Cont.	89	50	11	64.03 %	89 %	0.74
	Cont. + bio.	97	47	3	67.36 %	97 %	0.8
	Cont. + adic.	90	62	10	59.21 %	90 %	0.71
	Cont. + bio. + adic.	94	53	6	63.95 %	94 %	0.76
	Adic.	73	129	27	36.14 %	73 %	0.48
rato	Cont.	14	44	6	24.14 %	70 %	0.36
	Cont. + bio.	17	92	3	15.60 %	85 %	0.26
	Cont. + adic.	16	115	4	12.21 %	80 %	0.21
	Cont. + bio. + adic.	15	82	5	15.46 %	75 %	0.26
	Adic.	11	183	9	5.67 %	55 %	0.1

Tabla 4: Resultados de los experimentos realizados

Características	Precision	Recall	F-Measure
Cont.	11.45 %	33.88 %	0.13
Cont. + bio.	14.9 %	27.94 %	0.12
Cont. + adic.	16.14 %	26.03 %	0.12
Cont. + bio. + adic.	17.67 %	28.72 %	0.15
Adic.	10.66 %	27.61 %	0.14

Tabla 5: Resultados de los experimentos de validación cruzada

5. Conclusiones

- Las bolsas de palabras de los contenidos de los tuits han dado en general buenos resultados, llegando a un 100 % tanto de Precision como de Recall en el proyecto gasofa.
- Por el contrario, las características adicionales por sí solas han sido las que peores resultados han aportado, siendo superadas en todos los experimentos por los otros cuatro grupos de características.
- Las características que mejores resultados han proporcionado han sido diferentes para cada proyecto:
  - Proyecto gasofa: las bolsas de palabras de los contenidos han aportado un F-Measure de 1.
  - Proyecto castellon: en este caso se ha alcanzado igualmente el 1 de F-Measure. Sin embargo, las bolsas de palabras de los contenidos no han sido suficientes para ello, habiendo que combinarlas con con bolsas de palabras de la biografía, con las características adicionales, o con ambos tipos de características, para llegar a dicho 1.
  - Proyecto senado: en este experimento los resultados descienden con respecto a los dos anteriores. El conjunto de características más acertado en las tres medidas consideradas ha sido el que combina bolsas de palabras de los contenidos, bolsas de palabras de la biografía y las características adicionales.
  - Proyecto comite: las bolsas de palabras de los contenidos combinadas con las de la biografía han aportado en este caso los mejores resultados. Se obtiene un Recall superior al 95 %.
  - Proyecto rato: los resultados de todas las pruebas han sido muy pobres, no llegando en ningún caso al 0.4 de F-Measure. Esto es debido especialmente a los bajos valores de Precision obtenidos. El experimento que aporta mayor F-Measure ha sido el que considera únicamente las bolsas de palabras de los contenidos, aunque cabe comentar que agregándole información de las biografías se ha conseguido el Recall más elevado.
- En cuanto a los experimentos de validación cruzada, sus resultados son considerablemente menos prometedores que los comentados anteriormente. En un entorno real, por tanto, la tarea se complica, y los

clasificadores entrenados con información de un proyecto concreto no caracterizan correctamente robots pertenecientes a sectores diferentes. Las características que mejores resultados han aportado en estos experimentos han sido las que incorporan información de los contenidos de los tuits, de la biografía y las características adicionales, aunque la diferencia con el resto de características es muy ligera. Cabe comentar que en ningún caso los SVMs entrenados con datos del proyecto gasofa o castellon han conseguido detectar correctamente un solo robot, lo que ha hecho descender los resultados significativamente.

- De cara al trabajo futuro, se pueden incorporar al análisis nuevas características como frecuencias de actualización o información del grafo social, tratando de perseguir una mejora de los resultados. En este sentido, el número de agrupaciones de características posibles es muy elevado, y sería deseable realizar un estudio en profundidad sobre cuáles de ellas son verdaderamente discriminantes.

- Adicionalmente, una experimentación con métodos alternativos de clasificación podría igualmente hacer variar los resultados.

- Queda pendiente también replicar el método descrito en [1] con los usuarios del conjunto de datos de este proyecto. Es un procedimiento más complejo que el aquí descrito, que precisa de información adicional como un etiquetado de tuits como spam o no-spam o listas negras de URLs. Además, determinadas características utilizadas requieren de la extracción del timeline completo de las cuentas de usuario mediante la API de Twitter. En este proyecto se ha optado por una aproximación ceñida a los datos de que disponemos, aunque de cara al futuro es necesario comparar las prestaciones de ambas.

Referencias

[1] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, 9(6):811–824, 2012.