AI Based Opto-Lexical Pattern Analysis for Behavior Categorization

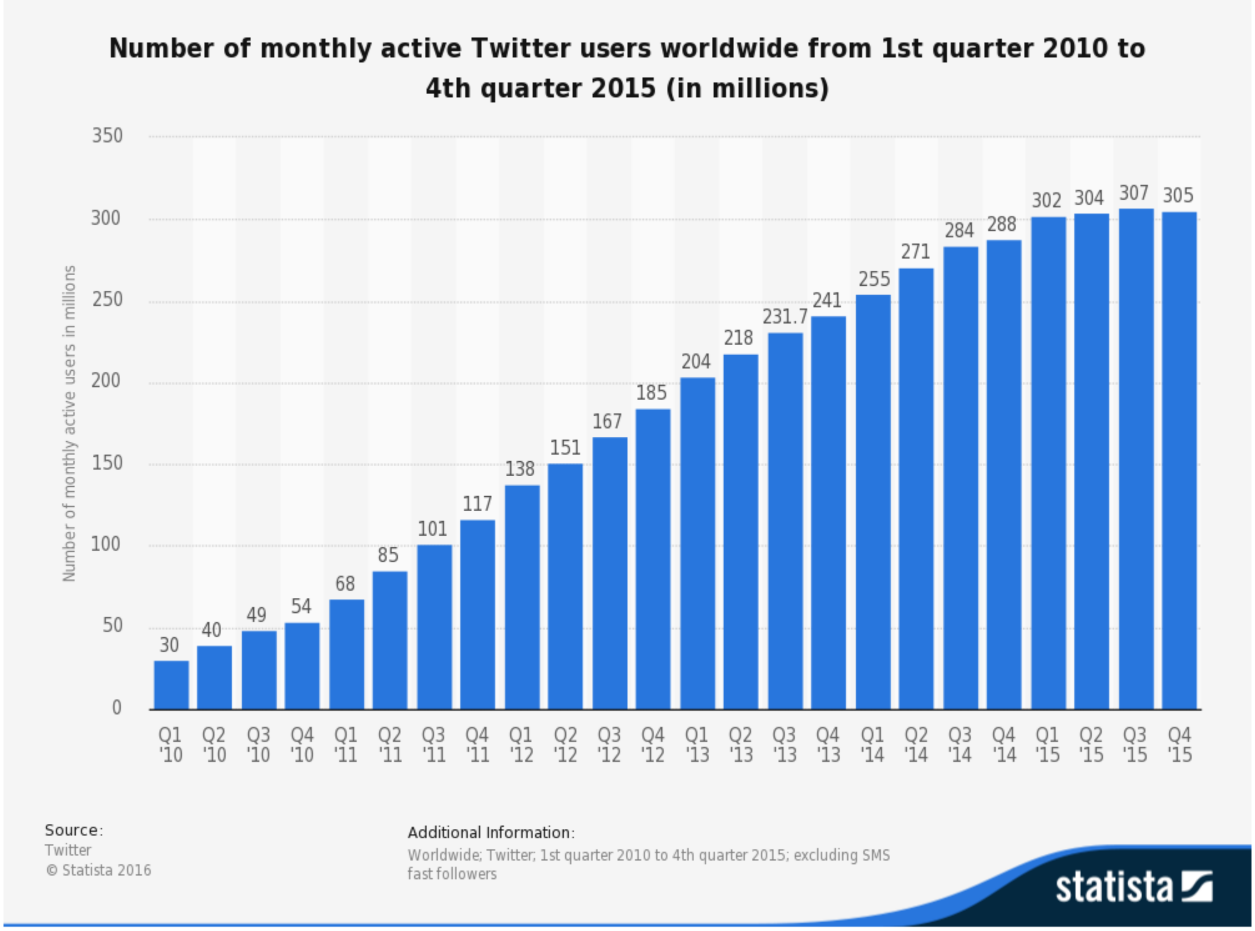
Data Book

Akshath Jain

11th Grade,
North Allegheny Senior High School,
akshath.r.jain@gmail.com

Table of Contents

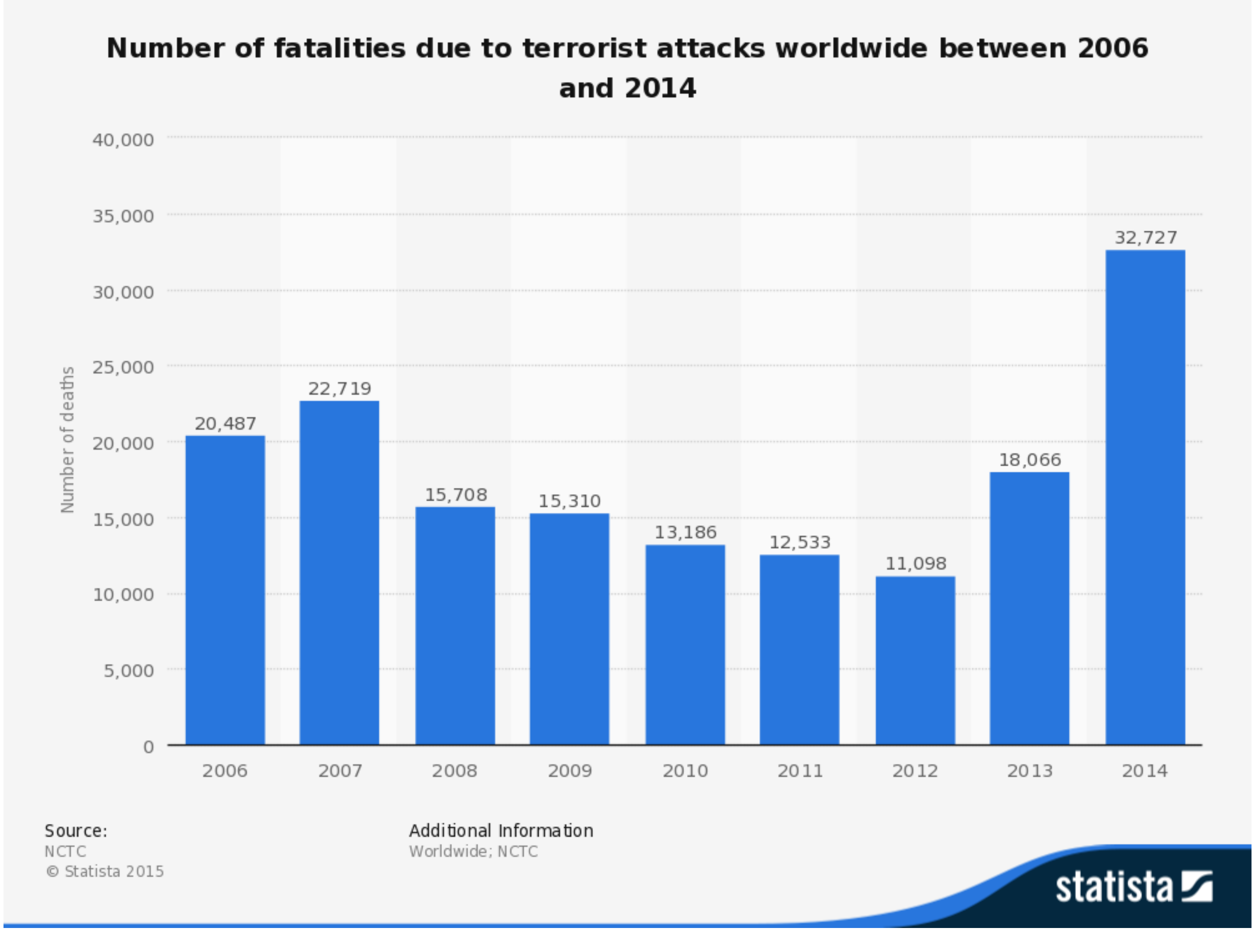
Number of Monthly Active Twitter Users



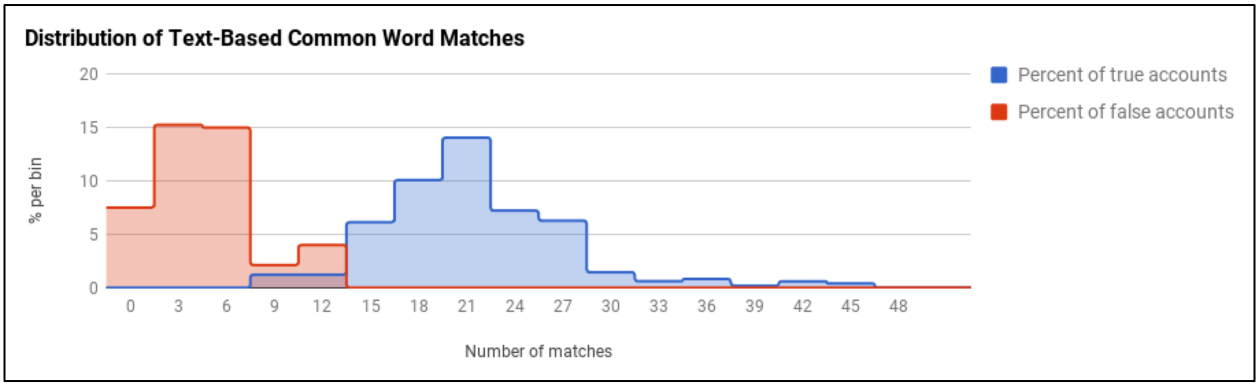
Source:
Twitter
© Statista 2016

Additional Information:
Worldwide; Twitter; 1st quarter 2010 to 4th quarter 2015; excluding SMS fast followers

Number of Fatalities Due to Terrorist Attacks

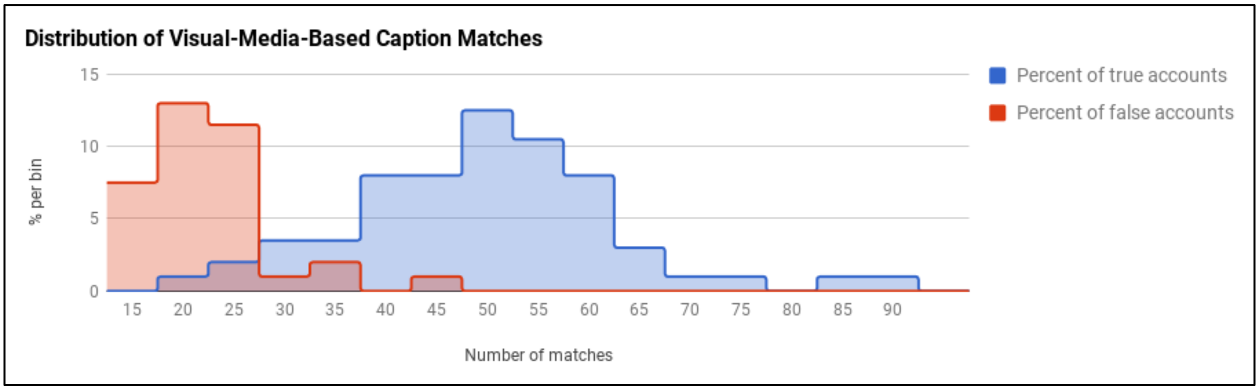


Distribution of Text-Based Common Word Matches



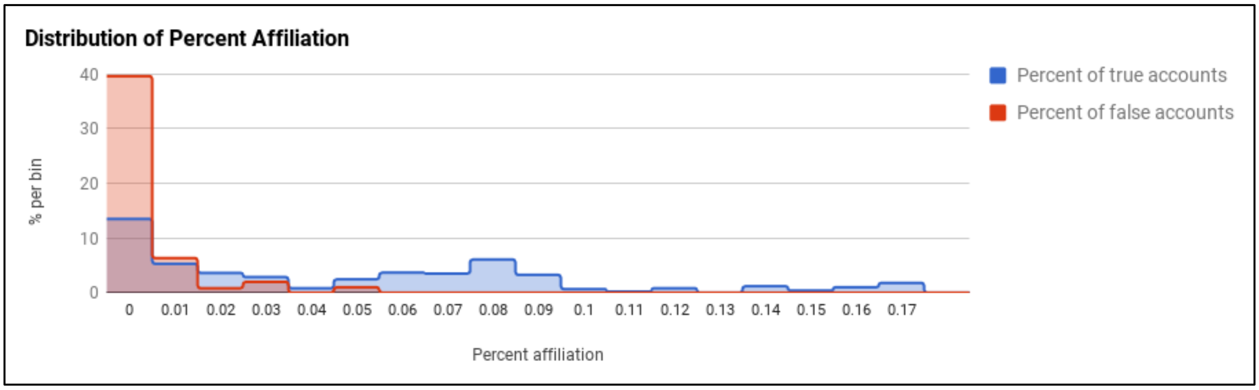
Bin	Percent of true accounts	Percent of false accounts
0	0	7.489795918
3	0	15.2244898
6	0	14.95918367
9	1.224489796	2.12244898
12	1.224489796	4
15	6.12244898	0
18	10.06122449	0
21	14.02040816	0
24	7.204081633	0
27	6.265306122	0
30	1.448979592	0
33	0.6163265306	0
36	0.8163265306	0
39	0.2	0
42	0.6	0
45	0.4	0
48	0	0
More	0	0

Distribution of Visual-Media Based Caption Matches



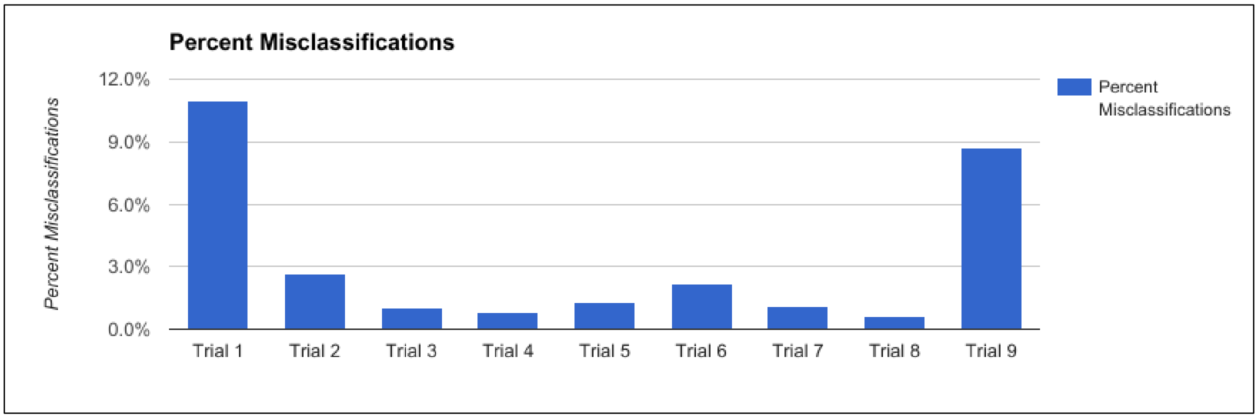
Bin	Percent of true accounts	Percent of false accounts
15	0	7.5
20	1	13
25	2	11.5
30	3.5	1
35	3.5	2
40	8	0
45	8	1
50	12.5	0
55	10.5	0
60	8	0
65	3	0
70	1	0
75	1	0
80	0	0
85	1	0
90	1	0
More	0	0

Distribution of Percent Affiliation



Bin	Percent of true accounts	Percent of false accounts
0	13.51020408	39.63265306
0.01	5.306122449	6.346938776
0.02	3.632653061	0.8163265306
0.03	2.857142857	2
0.04	0.8163265306	0
0.05	2.448979592	1
0.06	3.673469388	0
0.07	3.489795918	0
0.08	6.081632653	0
0.09	3.265306122	0
0.1	0.6734693878	0
0.11	0.2244897959	0
0.12	0.8163265306	0
0.13	0	0
0.14	1.2	0
0.15	0.4081632653	0
0.16	1	0
0.17	1.8	0
More	0	0

Percent Misclassifications



	True Accounts	False Accounts							
	Trial 1: Hostile	Trial 2: Brands and Products	Trial 3: Companies and Organizations	Trial 4: Local Businesses	Trial 5: Movies and Television	Trial 6: Music	Trial 7: Sports	Trial 8: Websites	Trial 9: Individual
Percent Misclassifications	11.0%	2.7%	1.0%	0.8%	1.3%	2.2%	1.1%	0.6%	8.7%

Training Set

Note that only a subset of the entire dataset used is shown because of resources limitations.

	Training Set	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
POSITIVE	Uncle_SamCoco	30	29	0.04
	RamiAlLolah	27	53	0.05
	warrnews	44	80	0.09
	WarReporter1	45	77	0.06
	mobi_ayubi	22	25	0.09
	_IshfaqAhmad	24	58	0.01
	wayf44rerr	44	37	0.02
	Nidalgazaui	17	27	0.12
	MaghrabiArabi	38	77	0.04
	melvynlion	29	78	0.11
	NaseemAhmed50	24	73	0.11
	ismailmahsud	37	39	0.03
NEGATIVE	_notmichelle	7	12	0
	jesseayye	5	13	0
	MrBrianLloyd	4	16	0.01
	sarahdorat_16	10	6	0.01
	wanderIustregui	0	13	0.01
	andhesonit	2	11	0
	Jas_Thxku	6	18	0
	KLitzau	2	8	0.02
	ThePettyHomo	5	28	0.02
	darkgreyxv	7	1	0.01
	michellemichmic	7	15	0.02
	FreeFX2	8	6	0.02
	RichLantos	3	3	0.01
	lonesomestereo	10	13	0.02
	topnotchcvlt	5	22	0
	m_e_lees	7	19	0

	HobiHobiHaz	3	9	0.01
	BeyonceStanBish	5	13	0.01
	mrtzchris09	7	20	0.02
	ramzax10	5	17	0
	a_sewart	7	24	0
	ambreezye	9	13	0.02
	trisetella	6	30	0.01
	aguilarmigs	3	23	0.02
	e_loochie	2	5	0.02
	ViktorSeriogin	8	22	0.02
	oliviaaajames	1	0	0
	ghoulslayers	4	12	0
	UrosRancic	4	10	0.02
	zunurain214	5	24	0.01
	Meagan_Grieger	6	20	0.01
	tytheintern	9	7	0
	WWMobileNews	10	6	0.01
	MCsnipes	6	20	0.02
	Naktadogg	8	30	0
	cooperjellybean	10	6	0.01
	taeftjimin	5	16	0.02
	BIEBERnRIRI	8	17	0.02
	fiona_langdon	7	14	0.01
	EmmanCabs	10	16	0.02
	hannahturner19x	7	15	0
	KARMAWOWZA	5	0	0.02
	lilacjaime	4	1	0
	V4apese	1	26	0.01
	ThePerfectPad57	0	22	0.01
	queermerlxn	5	0	0.01
	KellerDwight	5	4	0.01
	emailthebert	4	13	0.01
	eebeniro	9	20	0.02

	KiusRaie	4	16	0.02
	mjsty_	3	23	0
	Teri0328	3	12	0.02

Trial 1: Positive Accounts

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 1: Positive Accounts	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
Fidaee_Fulaani	35	78	0.01
__alfresco__	10	46	0.08
ro34th	39	68	0.07
AsimAbuMerjem	30	46	0.04
IbnKashmir_	13	36	0.03
MilkSheikh2	29	68	0.06
warreporter2	35	26	0.07
pleaoftheummah	45	75	0.01
murasil1	14	30	0.12
btt_ar	23	81	0.08
safiyaimback	27	65	0.03
MaghrebiQM	31	69	0.09
abuhumayra4	15	42	0.04
lNSlDEWAR	19	28	0.12
CXaafada2	30	67	0.1
FidaeeFulaani	15	78	0.11
Abu_Azzzam25	29	73	0.06
Witness_alHaqq	16	45	0.11
WhiteCat_7	22	64	0.09
abdlrhmn15	27	67	0.09
al_nusra	33	82	0.04
1Dawlah_III	17	74	0.1
abubakerdimshqi	29	61	0.01
Al_Battar_Engl	18	58	0.05
Alwala_bara	45	78	0.12
Bajwa47online	9	72	0.11
BaqiyaIs	34	28	0.03
Battar_English	40	34	0.11

Jazrawi_Saraqib	36	26	0.06
klakishinki	17	40	0.1
nvor85j	43	69	0.09
saifulakhir	43	46	0.1
al_zaishan10	29	52	0.03
MaghrebiWM	12	80	0.04
moustiklash	28	27	0.03
QassamiMarwan	20	82	0.08
AbuMusab_110	28	81	0.02
Baqiyah_Khilafa	16	81	0.1
Freedom_speech2	17	51	0.07
MaghrebiHD	19	41	0.05
maisaraghereeb	13	57	0.1
alamreeki4	13	33	0.08
EPlC24	16	63	0.07
thefIamesofhaqq	34	85	0.07
1515Ummah	16	25	0.1
war_analysis	43	84	0.08
abutariq040	11	81	0.01
st3erer	41	43	0.09
AbuNaseeha_03	17	61	0.12
Mountainjjoool	15	83	0.05
MhzBnt	15	35	0.12
freelance_112	31	40	0.08
432Mryam	14	35	0.11
DawlaWitness11	44	30	0.08
06230550_IS	32	34	0.08
wayyf44rer	33	41	0.1
YazeedDhardaa25	44	35	0.05
Jazrawi_Joulan	28	38	0.12
squadsquaaaaad	40	29	0.09
k_kid04	38	43	0.1
JoinISNation102	12	66	0.02

04_8_1437	44	31	0.09
Suspend_Me_fags	35	45	0.01
abutariq041	38	67	0.11
BilalIbnRabah1	39	57	0.12
grezz10	25	82	0.11
emran_getu	9	28	0.08
ALK___226	24	57	0.05
sonofshishan	15	63	0.08
AbdusMujahid149	9	74	0.06
ansarakhilafa	45	67	0.02
darulhijrateyni	19	77	0.09
___KU217_y	37	29	0.04
GunsandCoffee70	11	26	0.06
mustafaklash56	26	59	0.05
kIakishini5	36	32	0.08
abuayisha108	33	63	0.07
JohnsonsBot	28	68	0.03
MaghrebiQ	43	72	0.11
ManKhalfahum	11	39	0.12
Abu_Ibn_Taha	43	28	0.1
c0n0fj1had4_	32	35	0.01
mustaklash	16	32	0.11
abuhanzalah10	21	34	0.04
DabiqsweetsMan	26	65	0.1
Dieinurage308	10	80	0.08
wayff44rer	45	52	0.09
AbuLaythAlHindi	37	27	0.08
Afriqqiya_252	12	72	0.03
baaqiya_01	40	64	0.03
dieinurage29__7	14	59	0.1
almuhajirun9	44	70	0.09
bintraveller	9	85	0.07
ks48a174031	40	28	0.12

Mosul_05	39	75	0.05
Abdul__05	19	62	0.08
abuayisha102	20	33	0.09
fahadslay614	21	68	0.04
newerajihadi61	22	62	0.1

Trial 2: Brands and Products

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 2: Brands and Products	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
MercedesBenz	8	29	0.02
Audi	4	3	0.02
DJIGlobal	5	13	0.01
Polaroid	3	23	0
Chanel	9	16	0.02
hm	6	26	0
marcjacobs	3	16	0.02
android	3	19	0
microsoft	8	4	0.01
googlechrome	2	6	0.02
windows	10	24	0.02
gopro	2	3	0.01
sony	9	30	0.02
maccosmetics	0	2	0
gfuelenergy	3	0	0.02
nike	6	12	0
jumpman23	10	16	0.02
adidas	3	22	0.02
unam_mx	9	6	0
harvardhealth	5	2	0

Trial 3: Companies and Organizations

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 3: Companies and Organizations	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
3m	1	11	0.01
mahindrarise	2	15	0.01
tatacompanies	7	19	0.01
generalelectric	10	11	0
virigin	2	16	0
bayer	7	30	0
coopuk	10	10	0.02
dsm	1	6	0
lvmh	6	21	0
geindia	6	8	0.01
honeywell	8	25	0.01
utc	1	10	0
tupperwareww	0	26	0.01
mudocomtr	2	21	0
borusanholding	10	9	0
dubaiholding	7	22	0.01
ecoATM	10	23	0.01
google	1	26	0.02
apple	2	27	0
youtube	10	28	0
hdfclife	8	24	0
bupaarabia	8	23	0.01
mr_social10	8	29	0.02
esurance	7	0	0.01
exitoesvivir	8	13	0.01
axaindonesia	2	14	0
alrajhitakful	9	12	0.01

Trial 4: Local Businesses

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 4: Local Businesses	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
recordingacad	4	26	0.01
tomorrowland	8	9	0.01
e3	7	13	0
ultramusicfestival	4	30	0.02
comic_con	9	24	0.02
museummodernart	3	15	0.02
designmuseum	9	10	0.02
tate	7	15	0
guggenheim	0	26	0.01
metmuseum	4	16	0.01
babylon	7	13	0.01
niconico_bar	9	4	0
garajistanbul	1	6	0
saloniksv	4	1	0.01
sugarhut	8	4	0.02
hobhouston	5	20	0.02
livmiami	8	30	0
facesessex	6	4	0
hobnola	10	21	0.02
xscapesd	3	6	0.02
okuzofficial	3	19	0.02
hoborlando	2	9	0.01
storymiami	6	3	0
lavony	3	21	0.02
vortexjazz	6	22	0
amgniightlife	4	27	0.01
avnightclub	6	12	0.02
ukraveupdates	5	22	0

coyotefly	2	16	0.01
guaba	4	24	0.01
thehamiltondc	2	24	0.02

Trial 5: Movies and Television

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 5: Movies and Television	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
twiitertv	7	17	0
imdb	9	25	0.01
itunestrailers	5	30	0.02
rottentomatoes	5	4	0
cinepremier	4	19	0
yovoyalcine	2	18	0
korseries	3	28	0.02
semuafilm	4	6	0.01
filmsteria	5	16	0.01
boxofficeindia	7	5	0
srkuniversie	7	8	0.02
sibi_sathyaraj	6	30	0
robpattznews	9	6	0.01
cinema21	1	25	0.01
disneypixar	2	11	0
wbpictures	6	18	0
pokemon	6	13	0.02
sonpicsteens	10	24	0.02
starwars	4	20	0.02
universalpics	4	22	0

Trial 6: Music

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 6: Music	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
katyperry	0	10	0
justinbieber	10	3	0
rihanna	5	21	0.02
pink	7	4	0
niallofficial	0	4	0
brunomars	6	12	0.02
adele	5	1	0
maroon5	7	16	0.02
npr	9	17	0.02
nrjhitmusiconly	3	27	0.02
caracolradio	8	29	0
wradiocolombia	8	23	0
js100radio	5	6	0
pramborsradio	1	0	0.02
capital967	1	27	0
studio92	9	29	0.01
radioacktiva	6	18	0
radioadn	9	17	0.02
rfi	3	5	0
5fm	8	25	0.01
hot97	6	2	0.01
rds	5	28	0

Trial 7: Sports

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 7: Sports	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
cristiano	6	20	0.01
kingjames	8	0	0.02
neymarjr	7	13	0
kaka	0	4	0.02
kdtrey5	2	22	0.01
sachin_rt	0	5	0
imvkohli	5	25	0
nba	5	18	0.02
nfl	1	0	0
championsleague	10	3	0
fifacom	0	26	0.01
wwe	4	6	0
arsenal	8	13	0
lfc	4	17	0.01
lakers	4	10	0.01
steelers	2	4	0.02
ravens	8	30	0.01
warriors	10	14	0.01
chivas	9	7	0
mls	9	10	0.01
dallascowboys	2	4	0.02
bvb	3	13	0.02
chennaiipl	6	20	0.02

Trial 8: Websites

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 8: Websites	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
pewdiepie	4	3	0
theonion	4	9	0.02
9gag	0	12	0.02
yuyacast	10	8	0.02
detikcom	9	2	0.02
enews	0	13	0.01
sabqorg	5	17	0.02
huffingtonpost	8	12	0.01
g1	1	2	0.02
xhnews	7	22	0.02
la_patilla	1	23	0
gmanews	1	7	0
buzzfeed	5	16	0
nyimesbooks	7	4	0.02
tmz	4	15	0
wikileaks	3	29	0
espncricinfo	5	24	0
portalr7	6	23	0
zaytung	10	19	0.01
vivacoid	10	19	0.01
younghollywood	10	3	0.01

Trial 9: Individual

Note that only a subset of the entire dataset used is shown because of resources limitations.

Trial 9: Individual	Number of Common Word Matches	Number of Visual Media Based Caption Matches	Percent Affiliation
princesssam93	0	30	0.02
Holliewrightx	0	28	0.02
emmkov	7	19	0.01
celtic_stephen	9	6	0
xJanosklansx_UK	3	20	0.02
iam_ybnl	10	12	0
randomafstuff	4	14	0.01
be975n_	1	8	0.01
mattate9	7	14	0
coolhikingblog	5	6	0.01
themaddywong	4	24	0.01
buff_n3rd	4	18	0.01
RiotSquirrrl	6	23	0.01
JVP_28	4	14	0
DamonEarp	0	3	0.02
barackobama	9	13	0.02
kitkhatami	4	24	0
akshathjain	8	18	0.02
benlandis	4	9	0.01
mcthapreacha	3	27	0