Diffusion and emotion: The effect of sentiment of retweeting. Notes to an article by Lars Kai Hansen, Adam Arvidsson, Elanor Colleoni, Michael Etter and Finn Årup Nielsen

Finn Årup Nielsen

June 13, 2017

These note provide some extra information for our "Good Friends, Bad News - Affect and Virality in Twitter" article (Hansen et al., 2011).

1 Method

348862 tweets were downloaded at with the Twitter sprinkler streaming API. They were downloaded within several intervals, the first tweet was from 2010-09-09 10:13:12, the last at 2010-09-14 12:07.

For the COP15 data set 207782 tweets were downloaded in December 2009 by querying the search Twitter API (search.twitter.com) with the query 'cop15'.

The sentiment analysis only work on English tweets. Downloaded tweets also shows an indication of the language with a special field. However, for many non-English tweets the field is set to English. We therefore constructed a language filter: After download a language detector scored each tweet for *englishness*. The language detector used a wordlist with words scored from -3 to $+3$ indicating an estimate on how English each word is. The englishness of the words in a tweet were summed, and tweets with a positive sum were regarded as English. For the streaming Twitter data 106719 remained after extracting tweets where the user had set the language to English and the language detector also detected the tweet as English. For the COP15 data set 136262 tweets remained after language filtering.

For English tweets sentiment analysis scored individual tweets via an English word list specifically constructed for Twitter. The present list associates 1446 words with a valence between -5 and $+5$ all scored by FÅN with example words from the begining listed below:

```
abandon         -2
abandons        -2
abandoned       -2
absentee        -1
absentees       -1
```

```
aboard          1
abducted        -2
abduction       -2
```

Short word lists of obscene words helped to define some words (Baudhuin, 1973; Sapolsky et al., 2008). The sentiment analysis results in a *valence*, an *ambivalence* and an *arousal* value for each tweet. The valence of a tweet v is computed as the sum of the valences of the words in the tweet v_n. Arousal a is computed as the sum of the absolute value of the valences of the words.

$$
\begin{align}
v &= \sum_n v_n & \text{valence} \tag{1}\\
a &= \sum_n |v_n| & \text{arousal} \tag{2}\\
b &= a - |v| & \text{ambivalence} \tag{3}
\end{align}
$$

The extracted features were included in a generalized linear model (GLM) setup in R (`glm`) and Matlab (`glmfit`) modeling a binomial distribution with the standard link function, see, e.g., (Dalgaard, 2008). That model treats correlated Twitter data as independent.

As in (Suh et al., 2010) we add other variables to the GLM: "hashtag", "mention" and "url". "hashtag" indicates that the tweet contains one or more of the patterns "#hashtag", while mentions are of the type "@user". "Url" indicates that the tweet contain a web link "http://" A retweet is detected if the text has a pattern with "RT" or "via" followed by a user name "@user". User names that are part of a retweet indicator are not counted as a mention. We also count the length of the tweet text and let that enter as a variable, both as the raw full length and as the length discounted by the length of any detected web link and mention.

For further analysis a binary bag-of-words matrix was constructed. A stop word list on 571 words excluded common words and the 10'000 most frequent words were extracted in the corpus were used as terms. For the streaming data "lol", "love" and "good" were the most frequent while for the COP15 data the most frequent terms were "obama", "world", and "deal".

1.1 "Newsness"

To determine the "newsness" of a tweet we trained a classifier on the sentences of the Brown corpus with the NLTK toolkit (Bird et al., 2009). The news category of the Brown corpus has 4623 sentences, while the other categories have 42511 sentences (when excluding the 'editorial' category). Splitting 75% to a training and the rest to a test set the NLTK naïve Bayes classifier reports an accuracy on 84% on the test set. We apply the trained classifier on the tweet data getting a probability of "newsness" for each tweet. Based on the set of probabilities for all tweets we calculate the ratio of news (how many tweets have a probability larger than 0.5), the mean and median probability.

To evaluate the classifier and to provide statistics on the news in Twitter we (FÅN) manually labeled 1000 tweets sampled randomly among the streaming tweets classifier as English. The tweets were classifier into 4 categories: non-English, strictly news, broadly news and other.

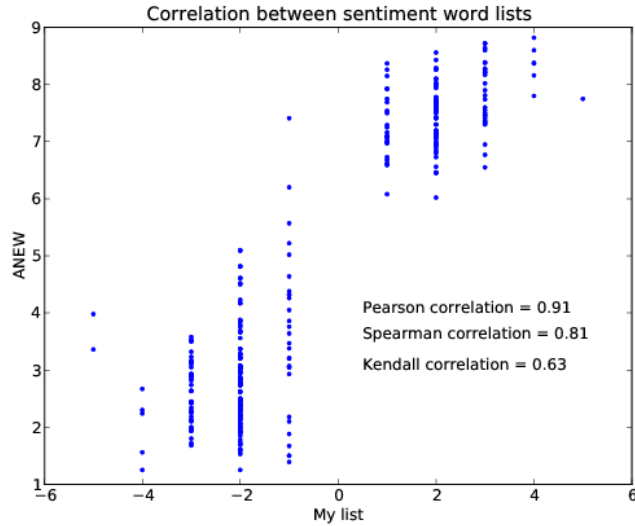The newsness variable is included in GLM model.

Figure 1: Scatter plot of valence for ANEW and our word list.

2 Results

2.1 Sentiment analysis

2.2 Summary statistics for Twitter stream

```
      id                retweet              url               mention
Min.   :2.327e+10   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:2.404e+10   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :2.440e+10   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :2.427e+10   Mean   :0.1767   Mean   :0.2521   Mean   :0.3929
3rd Qu.:2.441e+10   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :2.447e+10   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
   hashtag            followers          followees           length
Min.   :0.0000   Min.   :      0   Min.   :     0.0   Min.   :  1.00
1st Qu.:0.0000   1st Qu.:     54   1st Qu.:    65.0   1st Qu.: 51.00
Median :0.0000   Median :    170   Median :   158.0   Median : 81.00
Mean   :0.1441   Mean   :   1063   Mean   :   544.4   Mean   : 83.54
3rd Qu.:0.0000   3rd Qu.:    496   3rd Qu.:   386.0   3rd Qu.:119.00
Max.   :1.0000   Max.   :2911065   Max.   :580438.0   Max.   :239.00
 englishness          valence            arousal           normvalence
Min.   : 1.000   Min.   :-43.0000   Min.   : 0.000   Min.   :-0.999975
1st Qu.: 4.000   1st Qu.:  0.0000   1st Qu.: 0.000   1st Qu.: 0.000000
Median : 8.000   Median :  0.0000   Median : 0.000   Median : 0.000000
Mean   : 9.503   Mean   :  0.4849   Mean   : 1.504   Mean   : 0.006178
3rd Qu.:14.000   3rd Qu.:  1.0000   3rd Qu.: 3.000   3rd Qu.: 0.014706
Max.   :60.000   Max.   : 60.0000   Max.   :60.000   Max.   : 0.642853
  normarousal
Min.   :0.00000
```

Variable	Covariate	COP15		Stream	
		(en)	English	All	English
n		147041	136262	335236	106719
news ratio		0.303	0.305	0.226	0.233
newsness mean		0.328	0.326	0.288	0.271
newsness median		0.147	0.142	0.105	0.096
Negativity	—	4.889	4.649	2.775	-0.024
	log(length)	-4.263	-4.165	-21.84	-8.373
	length	-4.479	-4.339	-20.72	-7.904
	log(lengthnourl)	2.427	2.544	NA	-5.618
	lengthnourl	3.639	3.725	-15.46	-4.691
	url*length	-4.338	-4.194	-20.68	-7.917
	url*lengthnourl	3.860	3.959	-15.48	-4.904
Negativity-newsness	—	2.275	1.471	-6.019	3.904
	log(length)	3.957	3.077	8.082	4.984
	length	3.862	2.986	6.496	4.628
	log(lengthnourl)	2.761	1.884	NA	4.538
	lengthnourl	2.475	1.612	2.068	4.178
Only tweet with Arousal > 0					
n		44611	42087	53473	51929
Negativity	—	3.276	2.372	-9.725	-9.374
	log(length)	1.886	0.867	-8.902	-8.431
	length	1.675	0.725	-9.073	-8.604
	log(lengthnourl)	3.169	2.271	-9.585	-9.150
	lengthnourl	3.336	2.484	-9.612	-9.189
Negativity-newsness	—	1.125	0.180	1.179	1.239
	log(length)	0.974	0.286	0.685	0.763
	length	0.973	0.273	0.756	0.836
	log(lengthnourl)	1.120	0.198	0.874	0.938
	lengthnourl	1.126	0.155	0.893	0.960

Table 1: GLM Z-value/t-value for binary "negative" sentiment variable with the co-variates hashtag, mention and url predicting retweet. The "covariate" column denotes the extra column added as independent variable. The "lengthnourl" variable is the text length minus the length of any URL and any "mentions". "url*length" adds url, length and their interaction effect as covariates. "Negativity-newsness"

```
 1st Qu.:0.00000
 Median :0.00000
 Mean   :0.01786
 3rd Qu.:0.02586
 Max.   :0.99998
```

2.3   Twitter stream

```
glm(formula = retweet ~ hashtag + mention + url + length + valence +
    arousal, family = binomial)

              Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.4658054  0.0274688 -126.173  < 2e-16 ***
hashtag      0.5657288  0.0206364   27.414  < 2e-16 ***
mention     -0.6491782  0.0189536  -34.251  < 2e-16 ***
url         -0.5758612  0.0202971  -28.372  < 2e-16 ***
length       0.0237179  0.0002643   89.727  < 2e-16 ***
valence      0.0366561  0.0032107   11.417  < 2e-16 ***
arousal     -0.0189167  0.0033709   -5.612 2.00e-08 ***
```

Without the *length* variable the effect of the URL and the arousal changes on the retweeting:

```
glm(formula = retweet ~ hashtag + mention + url + valence + arousal,
    family = binomial)

              Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.664911   0.014341 -116.092  < 2e-16 ***
hashtag      0.762230   0.019725   38.643  < 2e-16 ***
mention     -0.570829   0.018187  -31.386  < 2e-16 ***
url          0.089287   0.018499    4.827 1.39e-06 ***
valence      0.036925   0.003023   12.214  < 2e-16 ***
arousal      0.061159   0.003084   19.834  < 2e-16 ***
```

2.4   Non-stationarity

To examine nonstationarity of the data we split the data set in several windows and performed the analysis independently for the data in the window. Figure 2 shows the generalized linear modeling regression coefficient and their standard error. In this plot each window consists of 25'000 tweets. The standard error is grossly underestimated perhaps due to the dependence among tweets.

2.5   News classification

The manual news classification found 33 of 1000 tweets to be non-English, showing that the automated English classifier is not perfect. Of the remaining 967 tweets 54 (5.6%) were news in a strict sense, while the result went up to 82 (8.5%) when a more broader
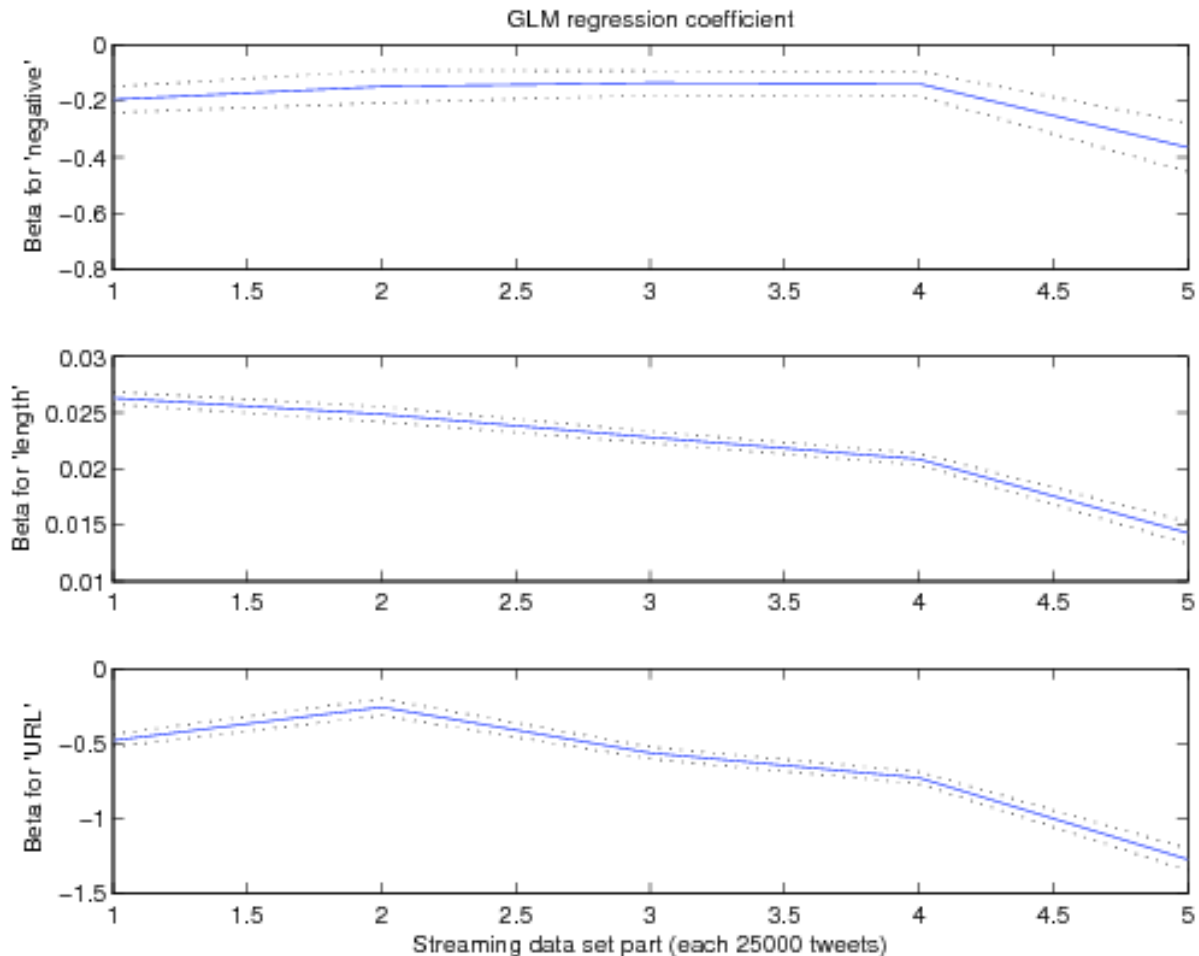
Figure 2: Evolution of the GLM regression coefficient

scope for news was applied. With the manually labeled data set we evaluated the classifier trained on the Brown corpus. Figure 3 shows the ROC curve for this news classifier. When the output of the classifier is thresholded on 0.5 then 222 (23%) tweets are classified as news. The confusion matrix at that level is displayed in Table 2. If we adjust the threshold so 82 tweets are classified as news we need to set it at 0.963 with the corresponding confusion matrix displayed in Table 3.

Table 4 shows examples of the tweets that the classifier reports as most likely news tweets (positives), while Table 5 display examples of tweets most likely not news. Among the 50 first in the list no news tweets occur, while for the list with tweets classified as news several non-news tweets occur. Some of these tweets contain elements of news in

	News (strict)	News (broad)	Other	Total
Classified as news	39	59	163	222
Classified as other	15	23	722	745

Table 2: Confusion matrix for news classifier and manually labeled Twitter corpus. With threshold on $P = 0.5$.
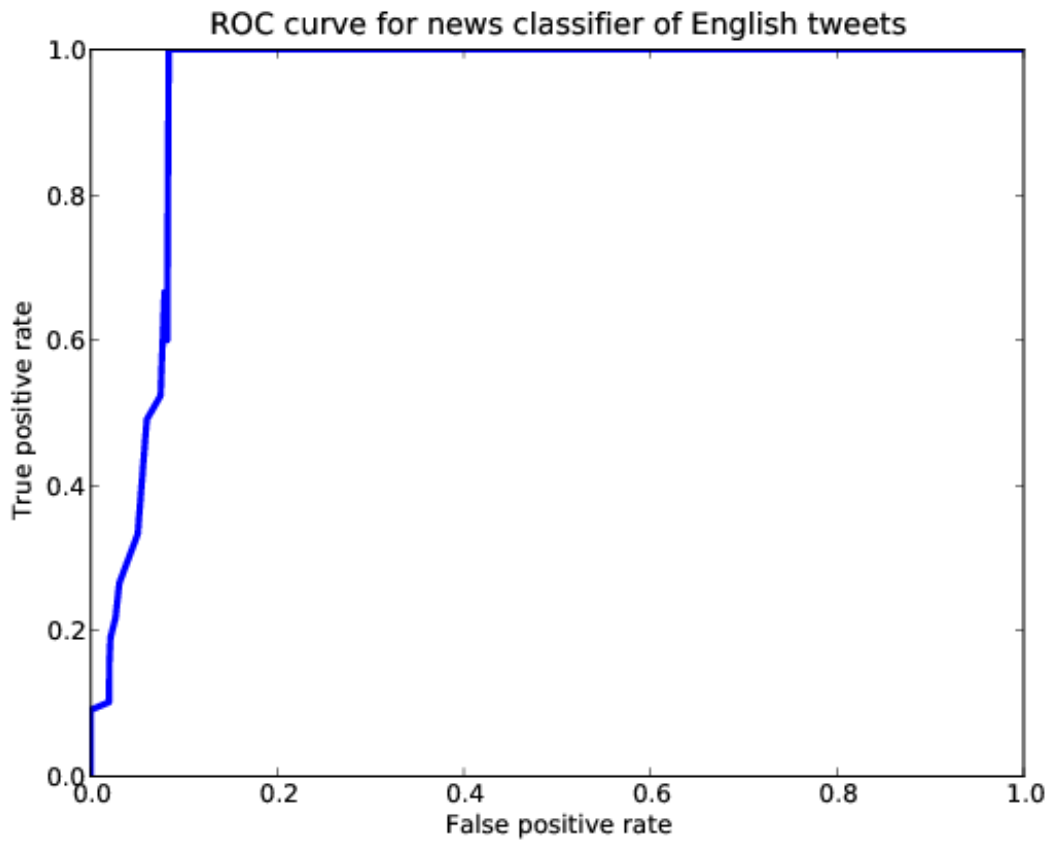
Figure 3: ROC curve for news classifier

	News (strict)	News (broad)	Other	Total
Positive	21	31	51	82
Negative	33	51	834	885

Table 3: Confusion matrix for news classifier and manually labeled Twitter corpus. With threshold on $P = 0.963$.

them, e.g.,

> Tonight is the night the start of the NFL season....stuck here in whodat nation with all these crazy Saints fans....Go Vikings.....

that combines a news and personal comment. Another is an announcement of an event:

> RT @SHO_TheBigC: New The Big C tonight at 10:30PM ET/PT on Showtime! http://s.sho.com/agfyq1Check-in with @clicker for an exclusive award!

References

Baudhuin, E. S. (1973). Obscene language and evaluative response: an empirical study. *Psychological Reports*, 32.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly, Sebastopol, California.

Dalgaard, P. (2008). *Introductory Statistics with R.* Statistics and Computing. Springer, New York, second edition.

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., and Etter, M. (2011). Good friends, bad news — affect and virality in Twitter. In Park, J. J., Yang, L. T., and Lee, C., editors, *Future Information Technology*, volume 185 of *Communications in Computer and Information Science*, pages 34–43, Berlin. Springer.

Sapolsky, B. S., Shafer, D. M., and Kaye, B. K. (2008). Rating offensive words in three television program contexts. BEA 2008, Research Division.

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In *2010 IEEE International Conference on Social Computing (SocialCom10)*. IEEE.

#	Label	Probability	Tweet
1	2	1.00000	Navy awards $40M Guam contract to Honolulu company - KPUA.net: HONOLULU (AP) — The Navy has awarded a $40 million ... http://bit.ly/9W9nNg
2	2	1.00000	AFL-CIO Sends Out Two Million Mailers; Blast Palin, Angle And Paul: Tea Party Senate candidate Sha... http://tinyurl.com/34mshs8 #HARRYREID
3	0	1.00000	#weekend Glee: The Complete First Season Starring: Matthew Morrison, Lea Michele Format: DVD (59)Release Date: Se... http://bit.ly/b8riP2
4	0	1.00000	I'm at Grand Central Terminal (87 E 42nd St, Park Ave, New York) w/ 7 others. http://4sq.com/2DsYgh
5	2	0.99999	Team USA Keepers for 2012 Olympics: The USA "B" Team won Gold at the 2010 FIBA World Championships. Which p... http://tinyurl.com/32ylgzh
6	2	0.99998	RT @goal_intl: #soccer #football Rio Ferdinand's Manchester United Reserves Omission Hints At Everton Comeback: And ... http://tmi.me/1cLDD
7	2	0.99998	Stuart Robbins Memorial Basketball Tournament (17th Sept 2010 ...: Sports News Ireland is a website providing Spor... http://bit.ly/8Xtjeh
8	2	0.99998	Jerry Jones demands Gene Wojciechowski be fired from ESPN. Picks Dallas to finish third in NFC East. http://es.pn/9mw1YD
9	0	0.99992	Tonight is the night the start of the NFL season....stuck here in whodat nation with all these crazy Saints fans....Go Vikings.....
10	0	0.99989	RT @MSKEYKEY: Correction meet thedonkryder and @TRINArockstarr The Diamond Princess @themaineventhwy85 on Friday Sept. 24th, 2010
11	2	0.99985	Less than a year after 9 month old James Dawson drowned in the bathtub of his Mother's Tulsa apartment, the baby's... http://fb.me/EHSUpe82
12	0	0.99979	RT @DJ_daDON Chris RT @DaGaryMasonShow: Adrian Peterson or @ChrisJohnson28? <<<<Chris hands down. AP just fumble 2 damn much.
13	0	0.99963	Thursday it's "Locals Night" at Garfinkels! Hip-hop, Dance, Electro, House, the mix start at 9.30pm. See you there! @garfswhistler #party
14	2	0.99951	NFLHuddle- Video: TODAY: Saints fans gear up for duel with Vikings - http://bit.ly/9dQA7A
15	1	0.99951	Trinidad vs Vargas highlight by Gorilla Productions http://boxeo.pl/2010/09/09/trinidad-vs-vargas/
16	0	0.99949	RT @Wes_HKNY: RT @BLIZZY_BL: Download @BLIZZY_BL- We Gettin Money (BOSS LIFE)...RT RT http://twiturm.com/6v8av RT RT RT RT
17	0	0.99943	@cpcc Cato Dean's Forum - When: Wed Sep 29, 2010 11am to 12pm EDT Where: Cato II Atrium Event Status: confirm... http://ht.ly/18WecM
18	2	0.99940	Alleged Meth Lab Near Lock Haven: For the second time since June, police in Clinton County have discovered alleged... http://bit.ly/c54v6l
19	1	0.99918	Come to "Clutter Busting Seminar" Wednesday, October 6 from 7:00 pm to 8:30 pm. Brooks Palmer, author of, "Clutter... http://fb.me/JpnO47DB
20	0	0.99917	RT @SHO_TheBigC: New The Big C tonight at 10:30PM ET/PT on Showtime! http://s.sho.com/agfyq1Check-in with @clicker for an exclusive award!
21	0	0.99909	good meeting with Josh & Steve from StageWorks GR this am, got a plan for the HD projector & sound for #24hourfilmfest screening Oct 16th
22	2	0.99897	Headline News - Pakistan drone attack kills eight: Eight militants die as missiles fired by a suspected US drone h... http://bbc.in/d4XCwT
23	2	0.99877	HH Nick O'Leary, Jack Nicklaus' grandson, suspended for making obscene gesture, report says: Dwyer High School tigh... http://es.pn/a7KmAP
24	1	0.99876	RT @LemuriaBooks: Curtis Wilkie's Fall of the House of Zeus: The Mysterious Planter from Greenwood http://wp.me/pjyCv-3kJ
25	1	0.99875	RT @JacksGuitarBar: Jack Stevens 9/18 Swamp Grass Willies JACK STEVENS SHOW & VIDEO SHOOT! - 10:00pm Promenade Shopping Center Palm Be ...
26	0	0.99859	Childbirth Class Part 2 — at Baylor Frisco - Professional Bldg II http://gowal.la/c/2ta24?137
27	0	0.99856	I'm at Friedl House (1300 Chain St., Milwaukee Rd., Beloit). http://4sq.com/9BffHE
28	0	0.99844	I'm at Eureka Presbyterian Church (2155 Farm Life School Rd, McCaskill Rd, Whispering Pines). http://4sq.com/cjJENK
29	0	0.99825	RT @KDthunderup: In new york bouta link up with my brother @jHARD13 pause!! Jay n eminem concert...officially Turnt up for 2 weeks str8 ...
30	1	0.99820	RT @WildcatClub: Arizona Athletics is calling for a "RED OUT" on Saturday for the Iowa game. Be sure to come to Arizona Stadium... http ...
31	0	0.99816	Beer Dinner at Lola's on Harrison http://bit.ly/auCr1F
32	1	0.99807	SOUTH AMERICA WANTS NEW KIDS ON THE BLOCK @NKOTB IN ROCK IN RIO 4 @Rockinrio ! www.nkotb.com XO, NKOTB BRASIL - OFFICIAL FAN CLUB IN BRAZIL
33	2	0.99797	Gah...RT @KarenWagg Rises in global food commodity prices may lift annual CPI food inflation to 7
34	0	0.99769	Tonights Homework: Martin Luther King, Jr. "War is a poor chisel to carve out tomorrow". Look into the down side of the Marshall Plan.
35	0	0.99760	1032sf Brand New bldg in Tai Hang (Mid-Levels East) $36000 3bd: Brand new luxury building located in the quiet ar... http://bit.ly/bSCx8m
36	2	0.99732	Tim Tebow, Jacksonville Jaguars Fans Talk: What B/R Overheard at Jaguars-Broncos http://bit.ly/acVCBq
37	2	0.99713	Apple's App Store Review Guidelines Offer Remarkably Candid View of App Store Philosophy http://j.mp/cAEVxM
38	2	0.99660	Air race friends pay tribute to pilot who died this year: There were hugs and smiles on Monday, but they were mixe... http://bit.ly/bq2gwi
39	2	0.99656	TV overhaul fueling battle over net neutrality: Netflix, meanwhile, is teaming up with game console makers to brin... http://bit.ly/9a46Ug
40	-1	0.99639	@Roza_N Du får ha det så bra<3 då är två av min vänner bort på semester och jag sitter här i den kalla Sverige:(
41	2	0.99635	RT @SkySports: #racing Absinthe suffers setback: Cambridgeshire contender Absinthe has been hit with a minor setback and faces a ... htt ...
42	0	0.99625	finally reached home:D dunno u hv reached bangkok ont cant sms u,dis weekend wil b vy long 4 me..enjoy ur trip n remember to miss me o
43	0	0.99603	RT @pollyswalking: Denver Tattered Cover Reading & Signing for 3mph: The Adventures of One Woman's Walk Around the World. Tues 9/14 7:30 ...
44	0	0.99598	@saffrontaylor Great week, Saffron. Rainy, dreary day here: http://youtu.be/rmeFrNPGXlc
45	0	0.99526	cant believe how villa got robbed last night...gutted.
46	0	0.99493	I'm at Mall Kartini (Jl. Kartini, Bandar Lampung) w/ 4 others. http://4sq.com/aZB3RF
47	1	0.99422	Come to "COMMUNITY, CULTURE, RESISTANCE. From the Crescent City to Chitown" Wednesday, September 22 from 8:00 pm... http://fb.me/HV7J7PdB
48	0	0.99403	Come to "(BOOKED) Pink and Black themed 18th Birthday " Tomorrow from 5:00 pm to 9:00 pm. BOOKED By Ms. Kay... http://fb.me/FuMJeMMX
49	2	0.99364	NYSE Euronext's Amex options CEO expects SEC nod on deal with Wall Street firms "soon," sees further boost to trading -Reuters interview
50	0	0.99242	[Contract] Web Designer / Developer (IT Professional II) at MRE Consulting, Ltd. (LM) http://dlvr.it/5BBJp

9

Table 4: Example of news classified tweets. Labels: "2" is news in strict sense, "1" in broad sense, "0" is non-news, "–1" is non-english.

#	Label	Probability	Tweet
1	0	0.00000	@imjustcreative Cool. I suppose this culture is indeed the reason for Apple's great success; holding them to the highest of standards :)
2	0	0.00001	@devikabhagat Seriously! Devika is clearly girls name. Guess Airtel needs to work on customer skills or hire people with common sense.
3	0	0.00002	@JeanCyrus And? Whats your point? Someone dmed me saying that I wasn't fit and that they want me dead.
4	0	0.00002	@daveyk317 I was told by someone never to talk politics or religion. I find It hard not to talk politics but I (cont) http://tl.gd/61hou5
5	0	0.00003	In Perth, Australia. It's fucking freezing cold. And the Internet is slow as hell. Can't even stream radio from europe without buffering...
6	0	0.00003	@Jothel jail..sounds like a scary scary place!! aww thats a shame about your fav band, i know what you mean it probably doesnt seem so bad
7	0	0.00004	A delivery breakfast service needs to be created because I like to eat breakfast in my pajamas and you can't do that when u get it yourself
8	0	0.00005	I still hear your voice when you sleep next to me I still feel your touch in my dreams
9	0	0.00005	want you to make me feel like im the only girl in the world, like im the only one that you'll ever love, the only one that knows your heart!
10	0	0.00006	Thoughts? A second ascending visual pathway from the optic tectum to the ...: that cells within this region are vi... http://bit.ly/bYgoLn
11	0	0.00006	@justinbieber I thought you were different but what now I see it is not so deseccion =(
12	0	0.00008	Out the door i go just thought of sum cold shit i guess today may be better than others
13	0	0.00009	@CheckxYesJuliet my eyes hurt. (I mean you tell me what's going on but I'm not you. I don't know everything you're feeling.)
14	0	0.00009	hay cutie :) hah. , people really need to shut up eh? D: — Oh, hey :) & yeah that they do ! Lmao. http://4ms.me/9C0lQF
15	0	0.00010	Som classic one liners from my ma over the wkd "sure, your grandfather used to drink methlyated spirits".
16	0	0.00010	@MaseDouglas haha.. you've never ecountered someone as strange as me! I LOVE THAT!
17	0	0.00010	@MARLONLWAYANS hey somebody said you holding your imaginary girlfriend in your twitter pic. i promise it wasn't me lol
18	0	0.00010	Why is terrorism always linked to religion wtf is wrong with there assholes fuck this is annoyin
19	0	0.00011	@luciahoff LaRosa and Cronin are my professional reference out of Spain. Bodo write too long articles. Good ones, but so extensive.
20	0	0.00013	What A System... I Have Never Seen Anything SO EASY and Yet Powerful! http://bit.ly/aHWUDW?=mtcw
21	0	0.00014	Your mind is a muscle and you can train it. http://bit.ly/c6isPC /via @whitneygeorge
22	0	0.00014	You usually know if its gon be a good day once you open your eyes in the mornings
23	0	0.00015	Obtain Management of Your Skin Problems — Personal Development Tips: Nobody is familiar with for sure what exactly... http://bit.ly/ai0h8E
24	0	0.00015	what shall we say to these things?thank you lord!
25	0	0.00015	[LOL! I can tell you love 'em too!] RT @SheSimmers: @kattebelletje Eat! Eat! Eat!
26	0	0.00017	@cantergirl omg yes ! Lmao and then you have a big patch of sticky hair.. I hate it
27	0	0.00019	Overheard: cheesy. XD - Girl: How much do you love me? Boy: I love you very much. Girl: If you really love... http://tumblr.com/xwgi24ckb
28	0	0.00021	#lirik I remember when I walked out, I remember when I said, I hated you But some how deep inside still lovin' you.. So sad and lonely
29	0	0.00022	This tweet right here makes me like Nick Jonas again... so fucking epic. RT @nickjonas: We will make it up to you. Loge you all
30	0	0.00022	@Leila_SPIN I am so seeing you this weekend , I have to, okay shall I bring lunch on SAT for you both-can that work?
31	0	0.00026	living my life happily without my phone for at least 4 days already, but not certain i have not lost it somewhere..
32	0	0.00026	@Focus_JRJ tell me y is @LostnNYCsir sending me text message saying I don't hit him up.....remind him who used to front of us always
33	0	0.00027	Eye am ash why oh you am why puh are ay and sea
34	0	0.00027	@changeling0711 Um, trust me, you didn't use the website, unless they suddenly became cool and give links now.
35	0	0.00028	@Mikeys_Cheeks haha Your always makeing me Want to eat makeing me not fit my jeans anymore Dx
36	0	0.00029	I can't believe that people with brainless morale still exist. -.-
37	0	0.00030	@LupeFiasco I wrote a poem like this. But the meanings are opposite.
38	0	0.00034	you know what ? i trusted you . so i tell you those things . if THEY get to know . you get into trouble . cos u bitched about THEM to me
39	0	0.00038	@swashbuckled I have no idea what you are talking about (no, please don't explain) But well done anyway.
40	0	0.00039	"No, you know what, hey. It's just a dance. It's your reunion. Let's do it. Let's dance." -Booth. #BonesMarathon
41	0	0.00039	You fkn alchyy!! Drink some water for a change :) RT @sayyG: $1.75 for a 32oz of miller highlife is that a deal or is that a fuckin deal???
42	0	0.00042	RT @gerardway: How much can you boys and boys and girls blow this up- or do we have to lip-sync?http://youtu.be/ixAq6dWnh1I
43	0	0.00043	I spend my life doing things for others no doubt hell bless me!!
44	0	0.00045	What about sunriseWhat about rainWhat about all the th... — ...all the blood we have shed before..did you ever sto... http://4ms.me/cbXiNp
45	0	0.00046	@Trancendance This is true. Although there are other pictures where I'm not naked, too!
46	0	0.00048	If someone doesn't trust you... I'm sure you gave him/her a reason not to!
47	0	0.00049	@anamuan they're usually cheeper than paper. And you can port them to a computer and read them there. Plus, fully searchable and noteable.
48	0	0.00049	RT @DamnItsTrue: I gave my heart to you in a box. I think you didn't see the word FRAGILE on it. #DamnItsTrue
49	0	0.00051	Why must it always be, so cold, I wanna stay in bed!!
50	0	0.00053	@32Red_Bingo I am looking forward to seeing how bad Nathan gets.

Table 5: Example of tweets classified as not news. Labels: "2" is news in strict sense, "1" in broad sense, "0" is non-news, "–1" is non-english.

A Code

A.1 Hansen2010Diffusion_glm.R

```
# $Id: Hansen2010Diffusion_glm.R,v 1.6 2010/12/02 20:25:05 fn Exp $

detach(d)
data = read.csv('Hansen2010Diffusion_glm.csv')
data_news = read.csv('mat_rows_newsness.txt')
attach(data)
attach(data_news)

binarousal <- arousal > 2
negative <- valence < 0
positive <- valence > 0

summary(glm(retweet ~ hashtag + mention + url + length + negative, family = binomial))
# Coefficient for NegativeTRUE is negative

summary(glm(retweet ~ hashtag + mention + url + length + negative*newsness, family = binomial))

i = abs(data$arousal)>1
d = data[i,]
dnewsness = data_news[i,]
detach(data)
attach(d)
dnegative <- valence < 0

summary(glm(retweet ~ hashtag + mention + url + dnegative, family = binomial))

summary(glm(retweet ~ hashtag + mention + url + length + dnegative*dnewsness, family = binomial))




############


r <- glm(retweet ~ hashtag + mention + url + followers + followees + log(followees+1) + length + eng
summary(r)

r <- glm(retweet ~ hashtag + mention + url + length + arousal, family = binomial)
summary(r)

r <- glm(retweet ~ hashtag + mention + url + length + binarousal, family = binomial)
summary(r)

r <- glm(retweet ~ hashtag + mention + url + length + negative, family = binomial)
summary(r)
```

A.2 Hansen2010Diffusion_glm.m

```
% $Id: Hansen2010Diffusion_glm.m,v 1.3 2011/05/04 13:35:03 fn Exp $

M = brede_read_csv2mat('Hansen2010Diffusion_glm.csv', ...
                       'method', 'dlmread', ...
                       'rowhead', 'no');

binarousal = M.matrix(:,13) > 0;
negative = M.matrix(:,12) < 0;
positive = M.matrix(:,12) > 0;

% retweet ~ hashtag + mention + url + length + arousal
% [b, dev, stat] = glmfit(M.matrix(:,[5 4 3 8 13]), M.matrix(:,2), 'binomial');

% glm(retweet ~ hashtag + mention + url + length + negative, family = binomial)
[b, dev, stat] = glmfit([M.matrix(:,[5 4 3 8]) negative], M.matrix(:,2), ...
                        'binomial');


clear beta se
window = 25000
for n = 1:ceil(size(M.matrix,1) / window)
  I = (1+window*(n-1)):min(window*n, size(M.matrix,1));
  X = [M.matrix(I,[5 4 3 8]) negative(I)];
  Y = M.matrix(I,2);
  [b, dev, stat] = glmfit(X, Y, 'binomial');
  beta(:,n) = stat.beta;
  se(:,n) = stat.se;
end

subplot(3,1,1)
plot(1:size(beta,2), beta(end,:), 'b-', ...
     1:size(beta,2), beta(end,:)+se(end,:), 'k:', ...
     1:size(beta,2), beta(end,:)-se(end,:), 'k:')
ylabel('Beta for ''negative''')
title('GLM regression coefficient')

subplot(3,1,2)
plot(1:size(beta,2), beta(5,:), 'b-', ...
     1:size(beta,2), beta(5,:)+se(5,:), 'k:', ...
     1:size(beta,2), beta(5,:)-se(5,:), 'k:')
ylabel('Beta for ''length''')

subplot(3,1,3)
plot(1:size(beta,2), beta(4,:), 'b-', ...
     1:size(beta,2), beta(4,:)+se(4,:), 'k:', ...
     1:size(beta,2), beta(4,:)-se(4,:), 'k:')
ylabel('Beta for ''URL''')
xlabel(sprintf('Streaming data set part (each %d tweets)', window))
```

```
% print -depsc /home/fn/fnielsen/eps/Hansen2010Diffusion_glm_part.eps
```