Disseminating Agricultural Information via Twitter: Data Mining Content and Views

Tara Murphy[1], Tyler Wilson[1]
[1]USDA National Agricultural Statistics Service, 1400 Independence Ave. SW, Washington, DC 20250

Abstract

Every year, the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) produces hundreds of reports, providing those in agriculture critical information. Since 2006, Twitter has become a viable mode in which millions of people disseminate and collect information. Since 2009, NASS has used Twitter as a means to highlight relevant information about the agency and information found within the many reports it publishes. As NASS and other agencies have become more adept at storing assorted types of metadata associated with their Twitter accounts, analytic programs, such as SAS, JMP, and R, have incorporated features that facilitate examining the dynamics involved when a person 'views' or reads a tweet. In this analysis, a replicable classification framework is applied to a sample of NASS tweets to evaluate what types of content elicit higher or lower viewership. In addition, descriptive statistics, text mining, and other data mining techniques are used to examine what factors are associated with the most views. The results of the analyses are discussed.

Key Words: Twitter, social media, impressions, text explorer, topic analysis, decision trees

1. Introduction

As the use of social media continues its historic rise similar to the internet dotcom boom in the 1990's, Twitter remains a viable and relatively easy way to communicate and share information. Many government agencies, including the U.S. Department of Agriculture, use Twitter accounts to make their information more accessible to the general public. The USDA's National Agricultural Statistics Service (NASS) is charged with providing timely, accurate, and useful statistics in service to U.S. agriculture. Historically and presently, these statistics are derived from comprehensive surveys and made public via hundreds of reports published each year. Only recently (2009) has NASS begun to tweet out highlights from these reports and other pertinent information related to the agency.

Much of the literature surrounding Twitter data transpires in two veins: research studies on Twitter data's power to predict outcomes, such as elections or flu pandemics (Tumasjan et al. 2010; Aramaki et al. 2011), and reports from firms, such as *Simply Measured*, that facilitate and promote marketing via Twitter analytics. In regards to the former, a number of recent studies have begun to explore the categorization of Twitter message content (Naaman et. al. 2010), dynamics of a user unfollow (Kwak et. al. 2011), and topic models estimating future retweets (Hong et. al. 2010). This research examines the content within NASS Twitter data and considers what aspects of this content may help the agency increase the number of times the information is viewed (called impressions).

Both qualitative and quantitative analyses are used in this research and, in many methodological respects, this exploration seizes on the sociological tenets of Grounded Theory in which qualitative data is first collected, then sorted, categorized, and coded (Glaser 1968). Classification models are then considered to understand what, if any, content is related to higher or lower viewership. These methods are discussed in Section 2. Section 3 shares the findings and a discussion follows in Section 4.

2. Methods

2.1 Data Exploration

This study was conducted on a sample of 3,591 tweets posted by NASS over 22 months, from May 13, 2015 through February 28, 2017. The number of impressions per tweet ranges from zero to 17,944, with an average of 1,847 impressions. Along with the contents of each tweet, a date and time stamp were stored in an external database. Additional indicators for the presence of hashtags, mentions (denoted by @ symbol), exclamation marks, and links to pictures/videos/reports were manually coded into the data.

A qualitative classification scheme was used to categorize the tweets based on their content. Seven categories were realized using three independent coders (raters) who attained a high ($k > 0.9$) level of interrater reliability or agreement as measured by Cohen's Kappa (Cohen 1960; McHugh 2012). The most prevalent category was *Ag News*, followed by *Forecast, Event/Announcement, Conversation, Repeat/Other*, *Census,* and *Survey Request*, respectively. *Table 1* below provides a breakdown of these categories.

Table 1: Categorization of Tweets

Title	Total	Definition
Ag News	2430	General agriculture news and statistics
Forecast	359	Future tense, predictive agriculture news
Event/Announcement	299	USDA/NASS sponsored event announcement
Conversation	174	Starts with @, towards a singular Twitter handle
Repeat/Other	166	An exact repeat or abnormal tweet (e.g., Star Wars day tweet)
Census	117	Any content that references the Census of Agriculture
Survey Request	46	A request for people to respond to a NASS survey

This initial exploration identified a category of NASS tweets behaving differently than the other six categories. The intention of tweeting in a *Conversation* is to directly respond to a question or inquiry from an external follower, whereas the intention of the other categories is to disseminate NASS information to the broadest audience possible. Because this research's intent was to explore how NASS can expand viewership when tweeting, the *Conversation* category was eliminated from further analysis. The remaining six categories, along with the other indicators of date, time, at signs, exclamation marks, and pictures, were used in further analyses.

A term and phrase analysis provided lists of terms and phrases within the sample by frequency. This analysis found some frequent terms and phrases that were meaningless for the purposes of this research. For example, the term 'twitter.com' was within the top ten highest frequency counts, but added no insight into the specific content within a tweet and

was thus removed. In addition, terms, such as '2015', '2015!' and '2015.', were recoded so as to be recognized as a single stemmed term. A Word Cloud illustrated terms and phrases by their frequency and colored in respect to the number of impressions with which they were associated (Seifert et al. 2008). *Figure 1*, shown below, added insight into possible relationships in the data between the target variable (impressions) and the terms within a tweet. The larger the term the greater the frequency within the sample. The terms in blue are contained in tweets with higher than average impressions. The terms in red are contained in tweets with lower than average impressions.

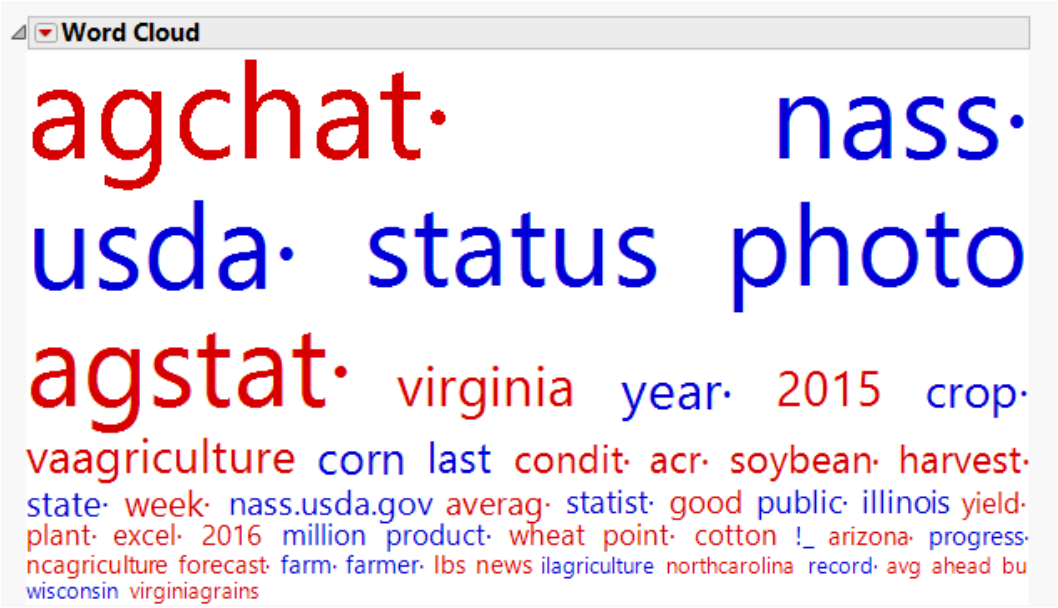

Figure 1: Word Cloud in Association with Impressions

A topic analysis (similar to factor analysis) was performed to explore similar semantics within and between tweets. Each term in each topic had positive or negative values scored to them, where negative values indicated less frequent occurrence in a topic compared to those with positive values. Twenty topics were formed from this analysis to be used as covariates in the classification models. Below is a list of the twenty topics the analysis identified. For a comprehensive list of the topics and their scores, see *Appendix A.*

Table 2: Topics Identified

ARMS[1]	Illinois	Wisconsin	Online/Local
Row Crop CAPS[2]	Arizona	Event	Release Date
North Carolina	West Virginia	CEAP[3]	Chickens
Booth Visit	Virginia	Forecast Yield	Vegetable/Organic
Event 2	Crop Condition	Missouri Soy	Kentucky

Many of the topics above reference a geographical location, while others reference events, crops/livestock, forecasts, online, and report releases. These 20 topics, in addition to the other hardcoded qualitative attributes mentioned above, were used as covariates to develop

[1] Agricultural Resource Management Survey, major NASS survey
[2] County Agricultural Production Survey, major NASS survey
[3] Conservation Effects Assessment Program, major NASS survey

classification models to identify and possibly predict impression levels. (See *Appendix B* for full list of covariates, identified as Term.)

2.2 Models

Classification trees were used to model which covariates lead to higher or lower impressions. Both bootstrap forest and boosted trees were assessed; however, due to much noise and small sample size, bootstrap forest was preferred (Dietterich 2000; Kotsiantis 2011).

Impressions were binned to create a categorical target variable, redundant covariates were eliminated, and pruning and trimming techniques were set, as briefly described below.

A capability analysis affirmed that a Normal 3 Mixture Distribution (Everitt 1985) best fit the distribution of impressions. Impressions were binned into *High*, *Medium*, and *Low* categories using the 0.25 and 0.75 quantiles of the Normal 3 Mixture Distribution, as defined as the lower and upper specification limits in *Figure 2* below.

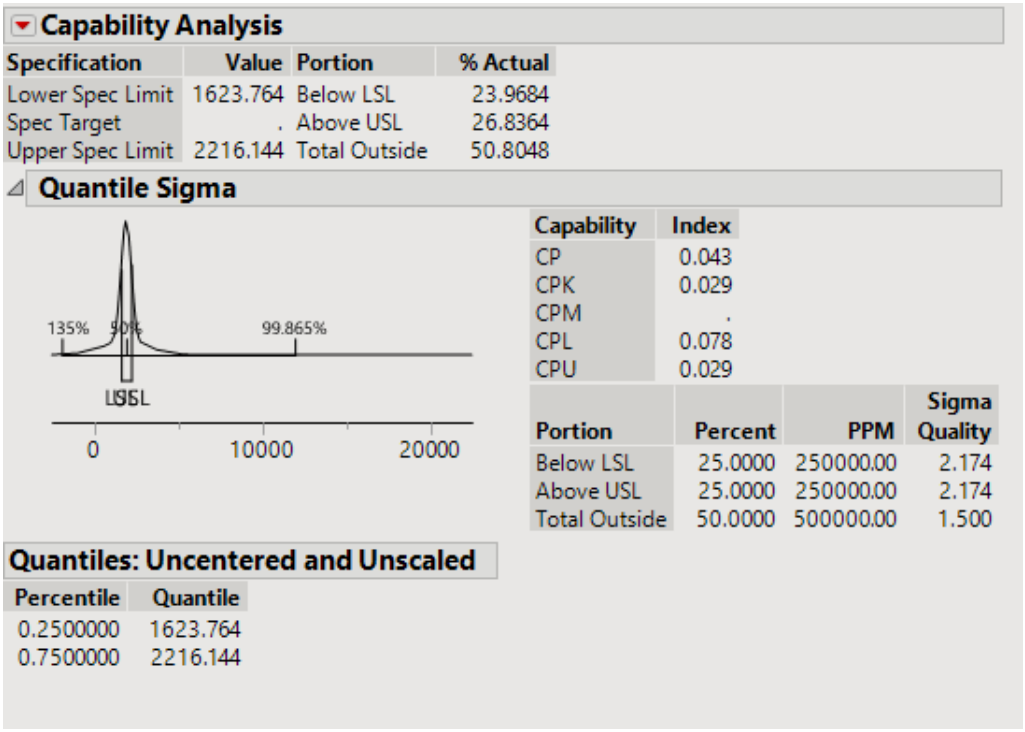

Figure 2: Capability Analysis of Distribution with Quantiles

In addition, the four lowest contributing variables were removed. These four variables did not include any of the aforementioned topics (see *Appendix B* for column contribution figure).

The minimum and maximum splits per tree were set at 10 and 2000, respectively. As *Figure 3* further shows, six predictors were set to be sampled at each split and at least five observations were needed at each tree node for it to be further split. Allowance of early stopping was also set for the bootstrap forest, which was employed in the selected model as shown in the following section.

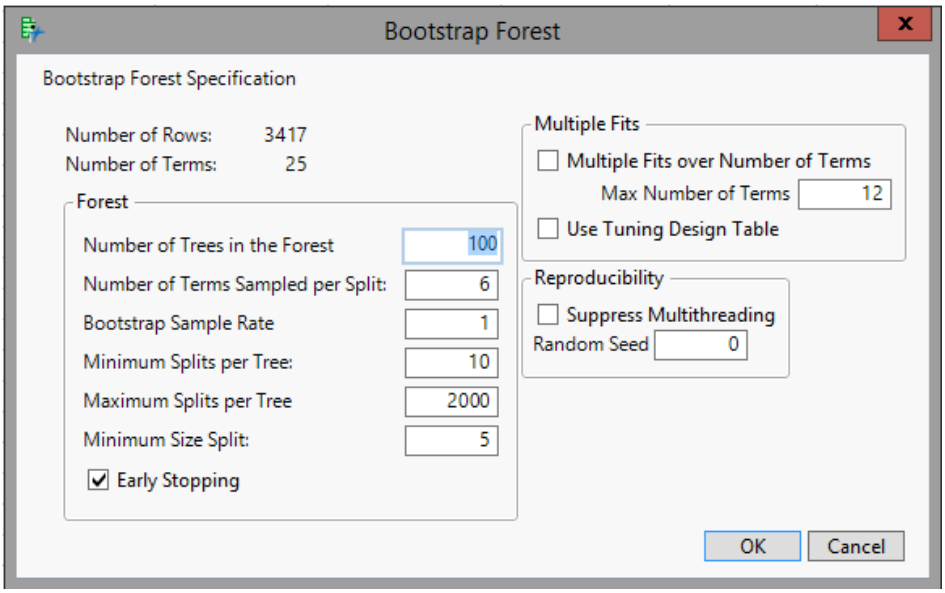

Figure 3: Bootstrap Forest Specifications

3. Findings

The Receiver Operating Characteristic (ROC) curve shown below shows the model having trouble classifying tweets in the validation data with medium levels of impressions (see green line in right graph of *Figure 4*).

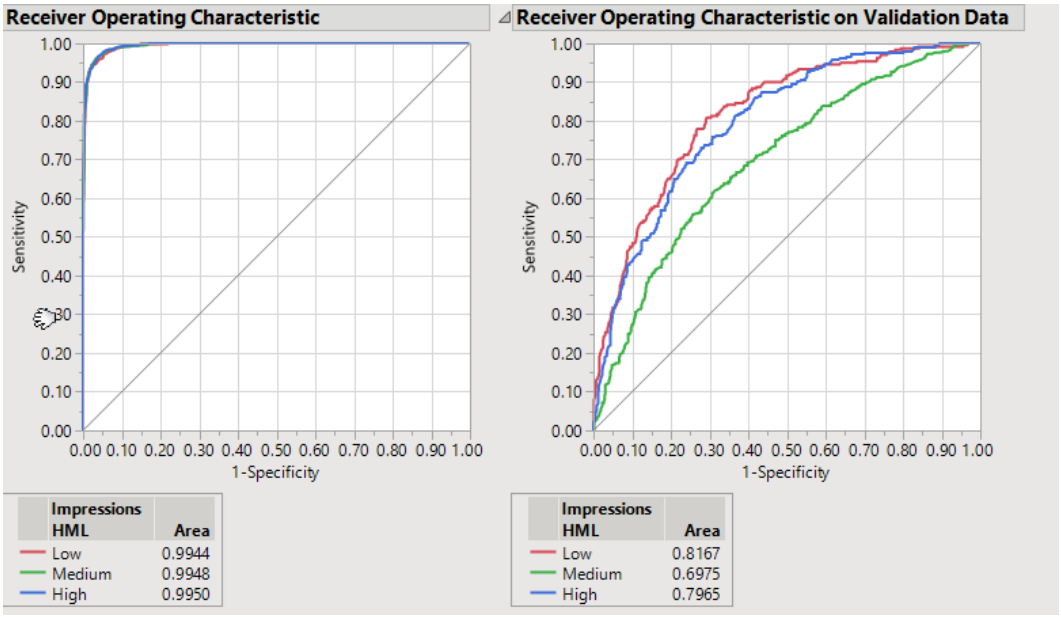

Figure 4: ROC Curves of Classification Model with Impressions binned as High, Medium, and Low

Tweets classified as possessing medium impressions were eliminated in order to find a more useful model that predicted 'good' and 'bad' tweets. Details of the final model are show below in *Figure 5*.

Bootstrap Forest for Impressions

Specifications

Target Column:	Impressions	Training Rows:	1201
		Validation Rows:	535
Number of Trees in the Forest:	34	Test Rows:	0
Number of Terms Sampled per Split:	6	Number of Terms:	25
		Bootstrap Samples:	1201
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.6034	0.3646	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7554	0.5287	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2742	0.4397	∑ -Log(ρ[j])/n
RMSE	0.2625	0.3719	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2269	0.3201	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0416	0.1925	∑ (ρ[j]≠ρMax)/n
N	1201	535	n

Figure 5: Classification model of Impressions Binned High and Low

The final model still shows signs of over-fit based on the calculated training versus validation entropy R-squared; however, the misclassification rate for the validation data is under 20 percent. As shown in *Figure 6*, the area underneath both ROC curves is approximately 0.9 (1 = perfect test) indicating that the model performs well classifying tweets with high and low levels of impressions.

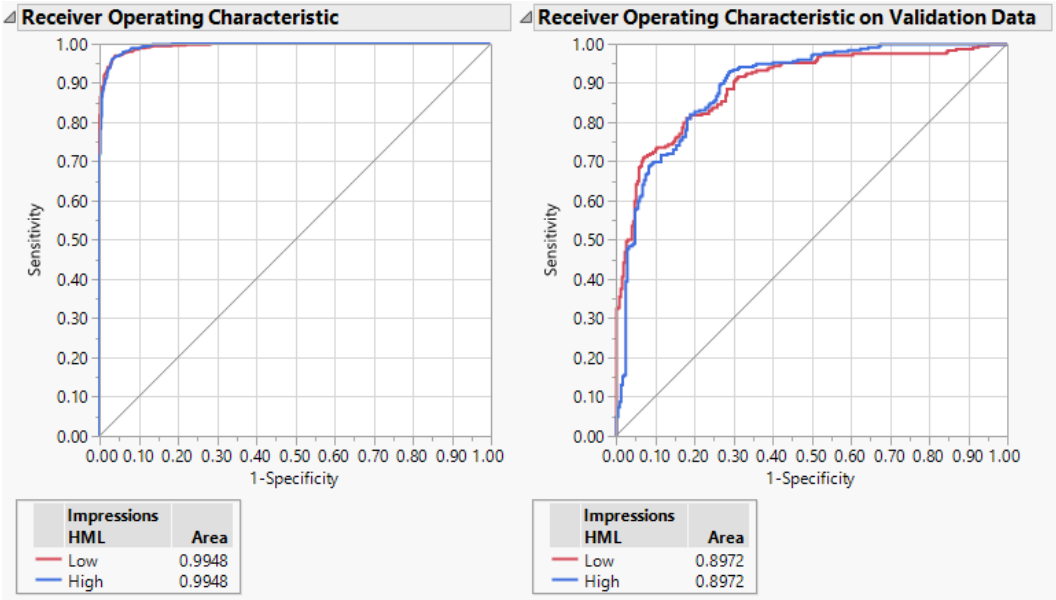

Figure 6: ROC Curves of Classification Model with Impressions Binned as High and Low

4. Discussion

The characteristics of NASS tweets and their relationship to impressions were explored. An attempt was made to classify and predict impressions based on their content and

surrounding metadata. A working model was produced, but limitations, such as small sample size and aggregate bias due to binning, make these results difficult to validate.

To ensure a high amount of interrater reliability between three coders, broad classification codes were necessary. When the starting point does not require focused specificity, this type of qualitative coding provides a good foundation for text mining research, assuming resources and document sample size are sufficient. Seven broad categories were realized, and an entire tweet category (*Conversation*) that functioned as a direct response to a tweet from an individual was removed from further analysis.

SAS JMP 13 offers a unique Word Cloud tool not often found within other text mining packages – the ability to categorize text by frequency and their association with another variable, here impressions. This Cloud tool provides a quick reference of words and phrases that may influence impressions in later classification models, in the same way as exclamation marks or photos within tweets. The bifurcation of colors can also inform the researcher of the general amount of topics to be formulated. At the start, a max of ten topics were set as the parameters; however, after analyzing the Cloud and the topics themselves, 20 topics were formulated to increase the level of subject specificity. For example, when set at 20, the 'Chickens' topic contained most of the words NASS associates with chickens and their corresponding surveys – 'broilers', 'hatcheries', 'incubators', 'eggs', and 'chicks'. Increased specificity of topics often leads to increased power of association with the target variable. In fact, none of the formulated topics were cut from the final model due to their relatively high contribution (see *Appendix B*).

To address the over-fit in the classification models, impressions were binned as binary. Binning any continuous variable results in a loss of power and a certain aggregate bias; however, binning to use classification models effectively is not without precedent (Sayad 2017). Binning highlights the initial purpose of this research – to explore what content relates to high and low viewership. Without a larger sample size and more influential predictors, it was determined that the noise contained near the center levels of impressions was too great. The ROC green line in *Figure 4* displays the noise relative to the high and low ROC curves. After eliminating tweets contained in this middle tier, both the high and low ROC validation curves were at 0.89. Assuming the model is correct, as the sample size increases, a fact that happens almost daily, we expect both the validation R-squared to increase and the misclassification rate to decrease.

A boosted tree will need to be considered in the future, as well. A boosted model was run once the tweets yielding medium-tier impressions were eliminated. This model indicated less over-fit; however, the entropy R-squared was lower and the validation model had a higher misclassification rate than the bootstrap forest model.

Here the sociological Grounded Theory approach to text mining and viewership levels of NASS tweets was taken. Only after analysis began were concepts formulated. NASS has been delivering objective, timely, and useful information to the public for over a century. NASS Twitter has become a useful method to highlight agency events, news, and findings from over 400 reports published annually to a broad audience interested in objective agricultural information. This study opens the door on how the contents of NASS tweets affect levels of impressions.

References

Aramaki, E., Maskawa, S., & Morita, M. (2011, July). Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1568-1576). Association for Computational Linguistics.

Bandari, R., Asur, S., & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. *arXiv preprint arXiv:1202.0332*.

Box, G. E. P., & Draper, N. R. (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, *40*(2), 139-157.

Everitt, B. S. (1985). Mixture Distributions—I. John Wiley & Sons, Inc..

Glaser, B. G., Strauss, A. L., & Beer, S. (1968). The discovery of grounded theory. na.

Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80-88). ACM.

JMP®, Version 13. SAS Institute Inc., Cary, NC, 1989-2017.

Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review*, *35*(3), 223-240.

Kwak, H., Chun, H., & Moon, S. (2011, May). Fragile online relationship: a first look at unfollow dynamics in twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1091-1100). ACM.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Naaman, M., Boase, J., & Lai, C. H. (2010, February). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 189-192). ACM.

Sayad, Saed., "Binning." *An Introduction into Data Mining*. Dr. Saed Sayad. June, 2010-2017.

Seifert, C., Kump, B., Kienreich, W., Granitzer, G., & Granitzer, M. (2008, July). On the beauty and usability of tag clouds. In Information Visualisation, 2008. IV'08. 12th International Conference (pp. 17-25). IEEE.

Simply Measured., "Simply Measured" June 26, 2017.

Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, *53*(8), 80-88.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, 10(1), 178-185.

Yu, B., Chen, M., & Kwok, L. (2011, March). Toward predicting popularity of social marketing messages. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 317-324). Springer Berlin Heidelberg.

Appendix

Appendix A Topics and Corresponding Scores

Topic Words

Topic1		Topic2		Topic3		Topic4		Topic5	
Term	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**
ii	0.35945	progress.pdf	0.33008	wi	0.37744	agcensus.usda.gov	0.33613	caps	0.36815
arms	0.32860	illinois	0.32274	wisconsin	0.36403	online	0.32023	matter·	0.35446
chem	0.31068	il	0.31502	16.pdf	0.27450	local	0.31395	row	0.33569
selected	0.29406	normal·	0.27873	15.pdf	0.25372	2012	0.30917	utextension	0.27348
practices	0.27601	ilagriculture	0.27830	details	0.20504	food·	0.29498	heard	0.26264
kycornfed	0.27311	crop·	0.23947	public·	0.17349	resourc·	0.29310	fill·	0.26248
prod·	0.22204	progress·	0.22160	statist·	0.16913	direct·	0.26403	estim·	0.24450
use·	0.20378	public·	0.17316	nass.usda.gov	0.16433	localfood·	0.22408	tnagriculture	0.21075
200	0.18256	statist·	0.16651	state·	0.16112	index.php	0.18303	counti·	0.20549
tell	0.18031	nass.usda.gov	0.15876	08	0.16106	sales	0.17095	tennessee	0.19800
grower·	0.17194	state·	0.15329	06	0.14466	consum·	0.16192	receiv·	0.18938
better	0.16866	now	0.14410	09	0.13794	agday365	0.16124	harvest16	0.17199
decisions	0.16487			dairy	0.13651				

Topic6		Topic7		Topic8		Topic9		Topic10	
Term	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**
az	0.3238	statchat	0.3031	2017	0.34357	pubs	0.35183	wvdeptofag	0.3260
azagriculture	0.3126	join·	0.2467	releas·	0.32964	usda.gov	0.35183	westvirginia	0.3194
azfb	0.2787	question·	0.2087	va.pdf	0.30582	todayrpt	0.35089	suitable	0.2854
azmilkproducers	0.2005	lanc·	0.2058	press	0.29403	ncagriculture	0.28404	wvuextension	0.2827
milk	0.1917	et	0.2043	current·	0.29302	northcarolina	0.28356	fieldwork	0.2497
arizonabeef	0.1546	tomorrow	0.1858	20	0.28773	ncgrown	0.27712	day·	0.2267
dairymen	0.1479	honig	0.1841	inventori·	0.20122	hog·	0.21788	goo.gl	0.1684
cwt	0.1249	today·	0.1650	goat·	0.19555	pig·	0.20828	nass·	-0.1473
growingarizona	0.1222	pm	0.1604	prjan17	0.17935	women4ag	0.13748	hay	0.1464
lbs	0.1175	send	0.1562	sheep	0.15822			status	-0.1376
missouriag	-0.1173	lh	0.1522	est	0.15452			photo	-0.1374
cow·	0.1172	answer·	0.1385	news	0.15182			usda·	-0.1314
		meet·	0.1342						
		agstat·	-0.1336						

Topic11		Topic12		Topic13		Topic14		Topic15	
Term	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**
conserv·	0.38881	broiler·	0.33414	ellison·	0.35177	vce	0.2932	yield·	0.3229
ceap	0.38604	hatcheri·	0.32216	herman	0.35177	vaagriculture	0.2884	forecast·	0.2898
phase	0.38591	week·	0.29773	visit·	0.32918	virginia	0.2863	ac	0.2868
farmersâ	0.35755	chicks	0.29552	booth	0.30806	news	0.2552	bu	0.2557
project·	0.31830	place·	0.29501	expo	0.27874	virginiagrains	0.2318	record·	0.1829
show·	0.29230	end·	0.29289	doubl·	0.25217	usda·	-0.2102	bushel·	0.1640
voic·	0.24724	incubators	0.23348	va	0.24653	status	-0.2033	aug	0.1421
make·	0.22238	egg·	0.23209	stat·	0.24129	photo	-0.2020	lbs	0.1414
heard	0.15614	set·	0.19718	ag	0.19276	nass·	-0.1884	high·	0.1299
survey·	0.14514	2015	0.15813			behind	0.1574	cured	0.1274
						yr	0.1440	flue	0.1274
						pts	0.1416	inventori·	-0.1273

Topic16		Topic17		Topic18		Topic19		Topic20	
Term	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**	**Term**	**Score**
guide	0.40381	audio	0.1952	excel·	0.38388	go.usa.gov	0.3287	kentucky	0.3139
veg	0.37094	ask·	0.1817	good	0.38374	soy	0.2905	annual	0.3132
survey·	0.30683	status	0.1811	poor	0.35219	mo	0.2822	bulletin	0.3051
look·	0.26857	photo	0.1809	fair·	0.34997	missouri	0.2662	kyag365	0.2911
count·	0.26146	nass·	0.1784	condit·	0.25288	full	0.2561	kentuckyag	0.2817
potato·	0.24856	usda·	0.1720	rate·	0.23468	missouricorn	0.2391	ky·	0.2306
amp	0.22442	hear	0.1699	ilagriculture	0.12489	rpt	0.1837	hear	0.1377
use·	0.21800	join·	0.1584	cond	0.12131	moagriculture	0.1826	statist·	0.1212
certifi·	0.19071	meet·	0.1552			agchat·	-0.1781	northcarolina	-0.1190
organ·	0.18594	live·	0.1532			report·	0.1656	audio	0.1167
		results	0.1500			ilsoybean	0.1397		
		access	0.1256			agstat·	-0.1381		
		question·	0.1245						
		agoutlook	-0.1185						

Appendix B Covariates

Covariates Reviewed

Column Contributions

Term	Number of Splits	G^2	Portion
Date	1648	240.527932	0.1076
Month	1202	180.920774	0.0810
Topic Virginia	1194	134.081707	0.0600
Topic Booth Visit	1205	101.370228	0.0454
Topic West Virginia	1133	84.4776815	0.0378
Topic North Carolina	1126	81.9866875	0.0367
Topic Wisconsin	1221	80.8090329	0.0362
Topic Kentucky	1079	77.6072432	0.0347
Topic Arizona	1171	77.0939091	0.0345
Topic Illinois	1162	76.7440549	0.0343
Topic Row Crop CAPS	1124	75.7781156	0.0339
Topic Event II	1040	75.173094	0.0336
Topic ARMS	1170	73.8645752	0.0331
Topic Missouri Soy	1085	72.8039873	0.0326
Topic CEAP	1087	72.2216126	0.0323
Topic Chickens	1082	71.5857747	0.0320
Topic Event	1070	70.3896914	0.0315
Topic Forecast Yield	1075	69.2055835	0.0310
Topic Online Local	1080	67.8603558	0.0304
Topic Veg Organic	1020	66.9160324	0.0299
Topic Crop Condition	997	64.8364648	0.0290
Topic Release Date	1030	64.7741789	0.0290
Time	1083	57.9273865	0.0259
#	992	55.5203398	0.0248
@	805	39.1573785	0.0175
Master Category	463	33.5077338	0.0150
Weekday	451	30.5998089	0.0137
Link 2	416	28.8894052	0.0129
!	185	8.28341436	0.0037

Final Model Covariates

Column Contributions

Term	Number of Splits	G^2	Portion
Date	209	97.9479325	0.1266
Month	211	89.9340054	0.1163
Topic Virginia	169	57.1512501	0.0739
Topic West Virginia	177	47.8310273	0.0618
Topic Booth Visit	152	43.6853994	0.0565
Topic Missouri Soy	141	26.0861665	0.0337
Topic Arizona	132	25.1735132	0.0325
Topic ARMS	130	24.4970147	0.0317
Topic Kentucky	126	23.6344565	0.0306
#	131	23.4322438	0.0303
Topic Chickens	131	23.1895738	0.0300
Topic Row Crop CAPS	129	23.1087579	0.0299
Topic Forecast Yield	137	22.9310059	0.0296
Topic Event II	110	22.7518662	0.0294
Topic North Carolina	117	22.5637391	0.0292
Topic Event	128	22.5540907	0.0292
Topic Wisconsin	138	22.3041001	0.0288
Topic Illinois	133	21.4123839	0.0277
Time	149	21.2080393	0.0274
Topic Release Date	127	21.176812	0.0274
Topic Crop Condition	124	20.4138985	0.0264
Topic Online Local	111	19.2105192	0.0248
Topic CEAP	112	17.9411484	0.0232
Topic Veg Organic	110	17.5799564	0.0227
@	101	15.6905303	0.0203