MANNHEIM BUSINESS SCHOOL

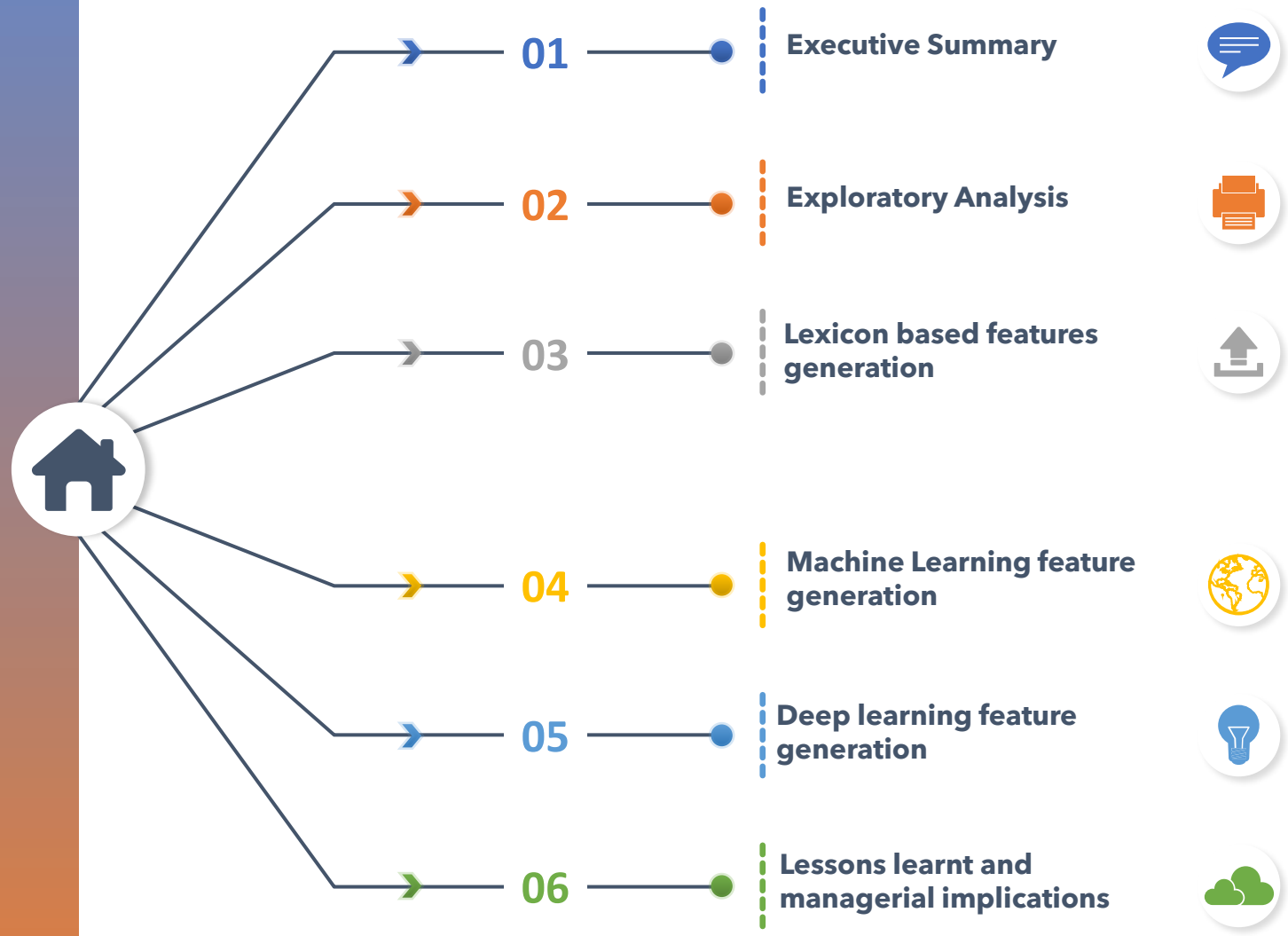
Predicting online success from tweets

Presented by: Wilfried Tcheumaha

JUNE 2022

TABLE OF CONTENT

Predicting success of Journalists on Twitter



Executive summary

The advent of social media has been a catalyst for making the whole world at our fingertips. Consequently, different people have made use of it for many reasons ranging from sharing their opinions to advertising their products or services to educate public. However, audience response seems to have been mainly associated with some language's patterns including but not limited to the processing fluency, the inclination of messages to specific topics, or their relative sentiment to name few.

In this report , we investigated the above hypothesized relationship in the context of journalism by analyzing tweets from a random selection of journalists or personalities within each of the three news organizations.

We extracted in total four textual features from the tweets. We first based this selection on a "buzz"-related custom dictionary including attention-grabbing words like breaking, news, president.., then we respectively derived the polarity and sentiment labelling from the AFFIN lexicon and the Roberta pretrained model and lastly we used the proportion of LIWC dictionary (operationalized as percentage of common words (1)) to capture in each text.

We then analyzed separately how each of these features perform in relation to their predictive utility using linear mixed models and controlling for random effects like the source and the author of the tweet.

Likes and retweets (natural log transformed) were in most cases, strongly associated in the positive direction. Their relationship with the features extracted was positive and statistically significant which suggest that, on average, news journalists tend to receive more engagements when Tweet content fall within the dictionary "categories" we identified

(1): David M. Markowitz and Hillary C. Shulman , The predictive utility of word familiarity for online engagements and funding

Exploratory analysis

3

Sources type: **Fox News, Associated Press, New York Times**

5

Journalists per source

13009

Tweets analyzed, 38.3% from fox news, 38.3% New York Times, 23.4% from AP

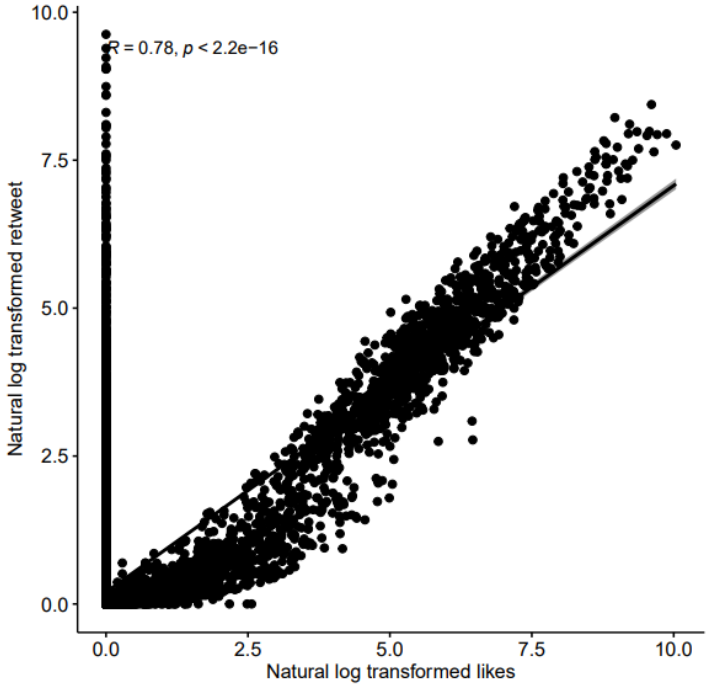
Descriptive statistics for Journalists

Likes and retweets are averages

Source	Journalist	Number_of_tweets	Likes	Retweets
AP	AlanSuderman	1	0	1
AP	JessicaLBruce	46	0.2826	64.348
AP	lhinnant	999	0.5506	325.92
AP	magancrane	999	1.2202	1581.7
AP	nekesamumbi	997	0.8445	5178.4
Foxnews	EricShawnTV	997	9.7262	99.011
Foxnews	LelandVittert	999	19.058	323.54
Foxnews	seanhannity	1000	6697.5	1951.1
Foxnews	ShannonBream	1000	1375.3	903.89
Foxnews	SteveDoocy	989	65.03	545.89
Nytimes	DanBarryNYT	1000	37.253	476
Nytimes	davidwchen	1000	26.995	301.53
Nytimes	DwightGarner	986	1.0791	9.7252
Nytimes	JulietMacur	996	32.984	508.71
Nytimes	perezpena	1000	3.303	1008.2

Likes Vs Retweets

In Natural log transformed



5 Top tweets

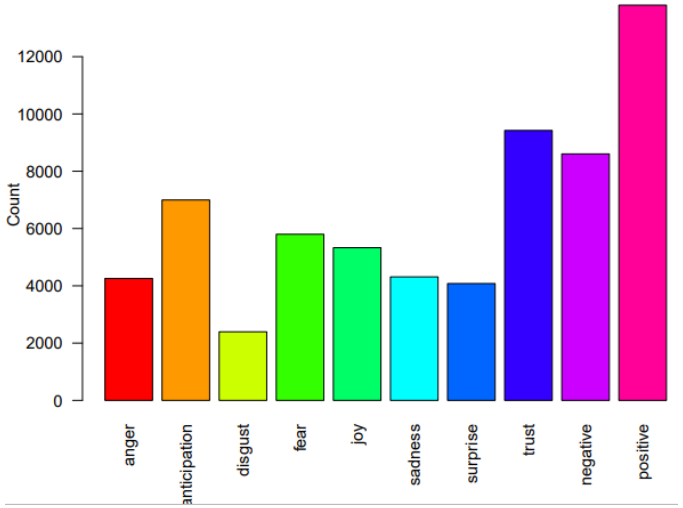
Favorite engagement level

1	seanhannity	"PRESIDENT TRUMP: â€˜I Will Never Give Up Fighting for You and Our Nationâ€™ https://t.co/CXjTtldxEc"	91819
2	ShannonBream	"If you're still catching your breath from today, I'm told next week will involve \"a bombshell\". I'm standing by ..~	83064
3	seanhannity	"Amazing and inspiring!! God please bless this country and â€œWe The Peopleâ€\u009d that make it Great. https://t.co/w~	82401
4	ShannonBream	"BREAKING: POTUS commutes Roger Stone's sentence. \n From WH: \"Roger Stone has already suffered greatly. He was tr~	67437
5	ShannonBream	"BREAKING FROM SCOTUS: Justice Alito has issued an Order than any ballots received after after 8pm on election day in~	67288

Exploratory analysis

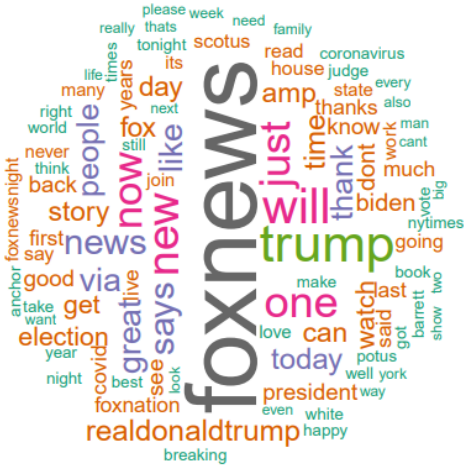
Sentiment classification

In the NRC Lexicon



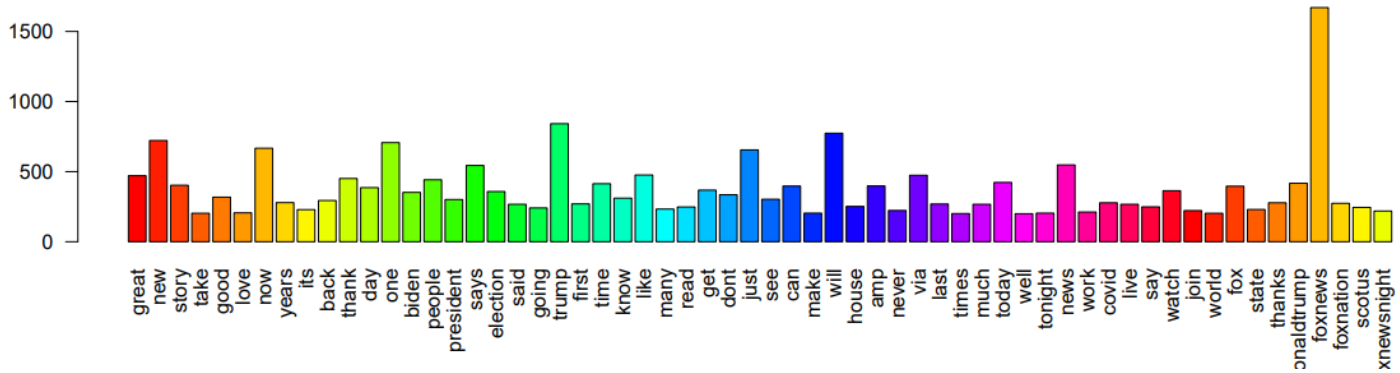
Word cloud

To 100 words used



Top word employed

Frequency above 200



The underlying sentiments of the most used words are split between **positivity, negativity, trust, anticipation, fear and joy**

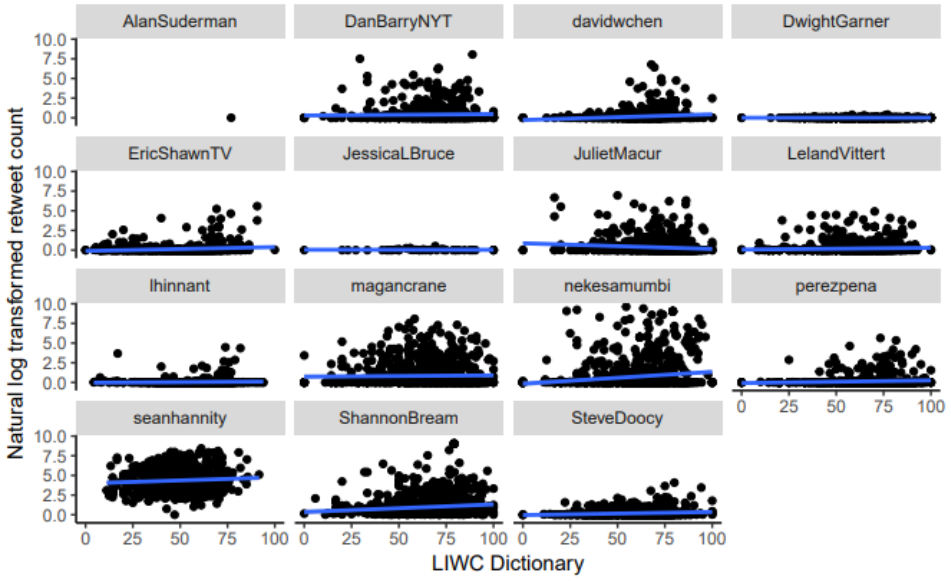
The words **Foxnews, Trump, will, can, now, news** dominate the spectrum of words tweeted. This revealed to a certain extent, **the political inclination** of the tweets amid the US **presidential election** that stir up passions. We were then intuitively prone to investigate if this pattern was associated with the audience engagement.

A custom lexicon based on words curated from few **buzz words** of the moment has been found to lead **to higher 'Likes'** engagement and this relationship was statistically significant (**p-value = 2.903e-11**). The same observation has also been found with word extracted from the **AFFIN Lexicon** (**p-value = 2.202e-09).**

Exploratory analysis

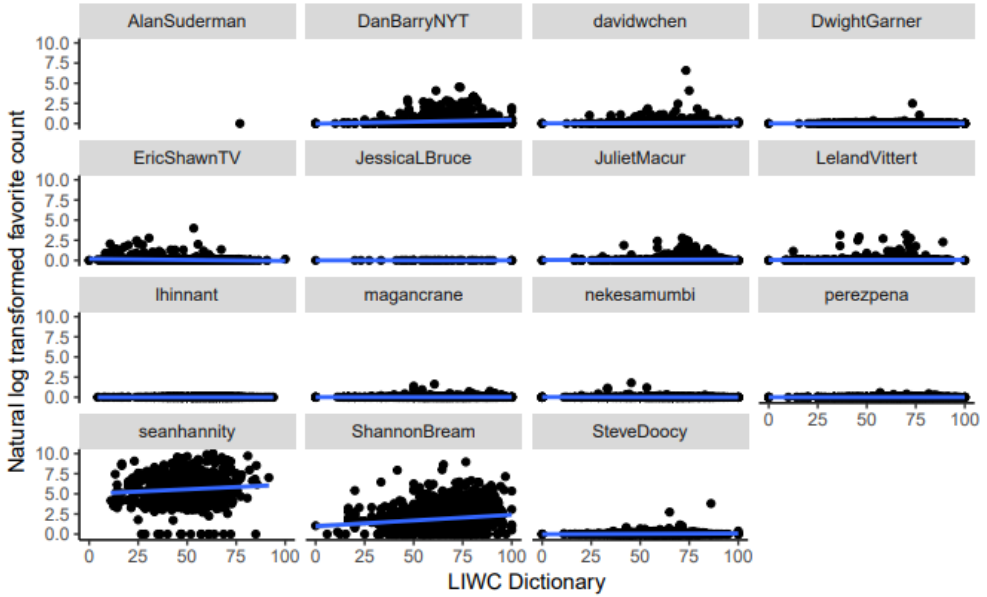
Retweet engagement

Natural log transformed retweet count by journalist



Likes engagement

Natural log transformed retweet count by journalist



The engagement level pattern varies across the **sources** and the **journalist names**. Therefore, we included fixed and random effects in our statistical models

Lexicon-based features generation

Linear mixed model results

Using lmer package in R

Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Likes	LIWC Dictionary	**1.88E-03**	**3.40E-04**	5.514	**3.57E-08**	0.000394	**0.852286**
	Intercept	4.32E-01	5.02E-01	0.86	0.48		
	Random effects	n	Variance	SD			
	Source	3	0.361	0.6009			
	Journalist	15	1.9422	1.3936			
Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Retweets	LIWC Dictionary	**4.26E-03**	**5.16E-04**	8.258	**<2e-16**	0.002518	**0.579267**
	Intercept	3.26E-01	3.18E-01	1.026	0.408		
	Random effects	n	Variance	SD			
	Source	3	0.05069	0.2251			
	Journalist	15	1.20855	1.0993			

The **rate of LIWC words** was positively associated with likes and retweets. Therefore, on average, news journalists tended to receive **more engagements** when **Tweet content** was **simple** compared to complex. However, one should note that we used the proportion of LIWC words in the tweets as a proxy of the processing fluency and hence, we admit with David and Hilary (1) the limitation of not using a direct measurement of the processing fluency.

Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Likes	"buzz" Lexicon	**5.75E-02**	**1.43E-02**	4.016	**5.95E-05**	0.00019	**0.85063**
	Intercept	5.37E-01	4.94E-01	1.089	0.39		
	Random effects	n	Variance	SD			
	Source	3	0.3384	0.5817			
	Journalist	15	1.9379	1.3921			
Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Retweets	"buzz" Lexicon	**1.92E-01**	**2.17E-02**	8.843	**<2e-16**	0.002646	**0.570162**
	Intercept	5.54E-01	2.97E-01	1.867	0.198		
	Random effects	n	Variance	SD			
	Source	3	0.0177	0.1331			
	Journalist	15	1.1943	1.0928			

The relationship between the **"buzz" custom lexicon based tweets** and the engagement at the **likes and retweets** levels was **positive** and significant. Therefore, on average, news journalists tended to receive **more engagements** when Tweet content was picked from "buzzed word". However, the extracted "buzz" feature did not account for the semantic meaning of the tweets and therefore would be prone to misclassification

Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Likes	AFFIN Polarity	**1.18E-04**	**1.14E-02**	0.01	**9.92E-01**	1.25E-09	**0.850879**
	Intercept	5.47E-01	4.96E-01	1.103	0.384		
	Random effects	n	Variance	SD			
	Source	3	0.3453	0.5876			
	Journalist	15	1.9387	1.3924			
Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Retweets	AFFIN Polarity	-0.02081	0.01733	-1.201	0.23	4.87E-05	**0.570383**
	Intercept	5.96E-01	3.04E-01	1.964	0.184		
	Random effects	n	Variance	SD			
	Source	3	0.02943	0.1716			
	Journalist	15	1.19614	1.0937			

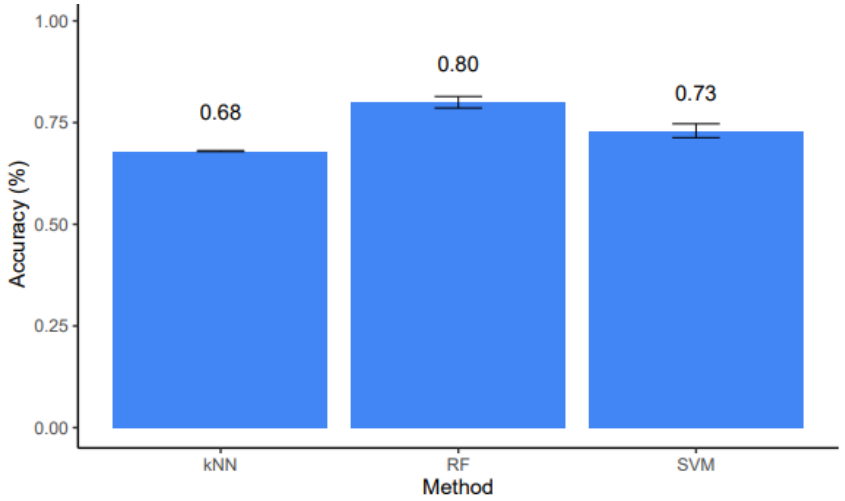
The rate of tweet with **AFFIN based positive sentiment** was **positively** associated with **only the likes engagement**. Its association with the retweets was not significant Therefore, on average, news journalists tended to receive more "likes" engagements when Tweet content was perceived as positive on the AFFIN lexicon. However, the AFFIN sentiment generated from the tweets did not account for the semantic meaning and therefore would be prone to misclassification

(1): David M. Markowitz and Hillary C. Shulman , The predictive utility of word familiarity for online engagements and funding

Machine learning feature generation

Accuracies comparison

Across various methods



We generated a textual feature called " **politically oriented**" to denote those tweets with a political tilt. We then sampled 500 observations , labelled them, split and trained them using various bag of word models , then selected the best model (Random Forest) to predict the new feature across the entire set.

However, the labelling setting was prone to introduce some biases as it did not include diverse perspectives.

Linear mixed model results

After extracting the "politically inclined" feature

Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Likes	Politically_inclined	**5.62E-02**	**1.83E-02**	3.068	**2.16E-03**	0.000118	**0.850117**
	Intercept	5.41E-01	4.94E-01	1.095	0.38714		
	Random effects	n	Variance	SD			
	Source	3	0.343	0.5856			
	Journalist	15	1.925	1.3876			
Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Retweets	Politically_inclined	**2.09E-01**	**2.78E-02**	0.02775	**2.78E-02**	0.002064	**0.564183**
	Intercept	5.66E-01	5.66E-01	0.5659	0.193		
	Random effects	n	Variance	SD			
	Source	3	0.02604	0.1614			
	Journalist	15	1.15984	1.077			

The **new feature** was positively associated with likes and retweets. Therefore, on average, news journalists tended to receive **more engagements** when **Tweet content** was **politically inclined.**

Deep learning feature generation

Linear mixed model results

After extracting the "Roberta sentiment" feature

Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Likes	Roberta label	**-3.28E-02**	**1.18E-02**	-2.784	**5.37E-03**	9.79E-05	**0.850696**
	Intercept	5.65E-01	4.97E-01	1.136	0.37308		
	Random effects	n	Variance	SD			
	Source	3	0.3503	0.5919			
	Journalist	15	1.9288	1.3888			
Engagement	Fixed effects	B	SE	t	p	Rm	Rc
Retweets	Roberta label	**-1.44E-01**	**1.78E-02**	-8.095	**6.23E-16**	0.00239	**0.568943**
	Intercept	6.67E-01	3.07E-01	2.173	2.173		
	Random effects	n	Variance	SD			
	Source	3	0.04108	0.2027			
	Journalist	15	1.16656	1.0801			

We generated a Roberta based positivity sentiment across the data set and fit a linear mixed model to investigate any association with the audience engagement. The **use of positive related tweets** was negatively associated with the likes and retweets.

Therefore, on average, news journalists tended to receive **more engagements** when **Tweet content** beard **negative sentiment.**

Lessons learnt and managerial implications

- This work sought to determine the **text features that explain success of journalists on Twitter. We examined tweets attention** and the results consistently revealed that texts that use common words capture attention more than complex texts.

- This observation is **consistent with prior similar work** (1) which attempted to empirically test "the **simple is better hypothesis**".

- Moreover, **politically tilted tweets** and other textual features extracted from custom lexicons were found to have generated **more engagement** than less politically inclined tweets.

- However, one could argue that the **political context of the tweets** would have driven the engagement as the **study did not control for the period**

- The study also revealed that the **variability of the engagement** was mostly explained by **the random effects** (sources and journalists) as depicted from the conditional Rc (0.57 – 0.85).

- The sentiment generated from **Roberta accounted for context** and therefore would be **less prone to misclassification** as compared to that generated from AFFIN lexicon. Therefore, one would give more credit to the observation that negative sentiment tended to lead to more engagement. This is consistent with the reality as we human like to take comfort in the positivity and anything negative is then subject to clarification, questions, reactions

(1): David M. Markowitz and Hillary C. Shulman , The predictive utility of word familiarity for online engagements and funding

Thank You