Canadian Charity Twitter Analysis

An analysis of tweets related three International Development charities

Introduction

Since the recession of 2008, Canadian charities have seen a steady decline in donations. New generations (Millennials and Gen-Z) display distinctly different charitable giving habits as compared previous generations. The future success of Canadian charities depends on understanding the sentiment associated with their organizations, and in engaging with new generations in meaningful ways through social media. These interactions will help to build trust with potential donors, which has been shown to influence future charitable giving.

Primary research questions include:
-*Topic identification*: What are the most common topics in social media posts associated with **World Vision Canada (WVC)** and how do they compare to other charities?
-*Sentiment Analysis*: What is the sentiment of public social media posts associated with WVC, and how do they compare to other charities?

The answers to these questions will be found using Text Mining/Natural Language Processing (NLP) to identify named entities, identify topics, and perform sentiment analysis.

Literature Review

Sentiment Analysis and Text Classification

Synthesis and Summary

- Sentiment Analysis has typically consisted of assigning a positive or negative value ('polarity') to a document, but more nuanced measures have been developed. Multi-class values are being applied using a wide range of sentiments (ex. love, joy, hate, anger). Researchers are also adding nuance to their analysis by applying different sentiment tags to different parts of documents (ex. sentences, entities), and representing the document sentiment as the most common sub-sentiment or representing all sentiments with some measure of occurrence or importance.
- Sentiment Analysis could be approached as a supervised or unsupervised problem – with the wide usage of existing pre-labeled datasets to train models or labeled sentiment lexicons of words to score documents. One paper demonstrated that when one lacks labeled data, a technique can be utilized where a small subset of the data is manually labeled, then used to train the remaining data probabilistically, then a training classification model is iterated to converge on a reasonable accuracy.
- Many of the papers compared different algorithms and approaches to find the best performance. It was found that Feature Selection (often n-grams) had a significant impact on model performance, and that unigrams outperformed bigrams in many cases. In algorithm comparisons, variations of Naïve Bayes or Support Vector Machines led to the most positive evaluation measures – with some researchers employing deep learning algorithms and Ensemble techniques with further success in accuracy and f-measures.

Papers reviewed

Trends on Sentiment Analysis over Social Networks (2018)
Christos Troussas, Akrivi Krouska and Maria Virvou

Opinion Mining and Sentiment Analysis (2008)
Bo Pang, Lillian Lee

Comparative Evaluation of Algorithms for Sentiment Analysis over Social Networking Services (2017)
Krouska, Troussas, Virvou

Sentiment Analysis and Classification: A Survey (2015)
Shailesh Kumar Yadav

Various Machine Learning Algorithms for Twitter Sentiment Analysis (2018)
Rishija Singh, Vikas Goel

Multi-Class Sentiment Analysis in Twitter (2018)
Mondher Bouazizi

Using Sentiment Analysis to Determine Users' Likes on Twitter (2018)
Yo-Ping Huang ; Nontobeko Hlongwane ; Li-Jen Kao

Combining Classical and Deep Learning Methods for Twitter Sentiment Analysis (2018)
Mohammad Hanafy, Mahmoud I. Khalil, Hazem M. Abbas

A Twitter Sentiment Analysis Using NLTK and Machine Learning Techniques (2017)
Bhagyashri Wagh, J. V. Shinde, P. A. Kale

(Article) Emotion and Sentiment Analysis: A Practitioner’s Guide to NLP (2018)
Dipanjan Sarkar

Classification of Short Text Using Various Preprocessing Techniques: An Empirical Evaluation (2018)
H. M. Keerthi Kumar, B. S. Harish

Distant supervision for relation extraction without labeled data (2009)
Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky

Text Classification from Labeled and Unlabeled Documents using EM (2000)
Nigam, McCallum, Thurn, Mitchell

A Comparison of Document Clustering Techniques (2000)
Michael Steinbach, George Karypis, Vipin Kumar

<u>R, Books, Tutorials, and other Resources</u>

Synthesis and Summary

Various resources will be utilized to further understand Text Analysis in practice, with Python-based resources referenced primarily for theory, and R-based resources for both theory and code.

Resources

tidytext: Text Mining and Analysis Using Tidy Data Principles in R (2016)
Julia Silge, David Robinson

Text Mining with R, A Tidy Approach (2018)
Julia Silge and David Robinson

Natural Language Processing with Python (2009)
Steven Bird, Ewan Klein, and Edward Loper

Applied Text Analysis with Python (2018)
Benjamin Bengfort, Tony Ojeda, Rebecca Bilbro

Analyzing Linguistic Data
Harald Baayen

A Statistical MT Tutorial Workbook (1999)
Kevin Knight

Dataset

I scraped Twitter tweets using a Python library called TwitterScraper (https://github.com/taspinar/TwitterScraper) to download tweets related to 3 Canadian charities:
-World Vision Canada
-Plan Canada
-Compassion Canada

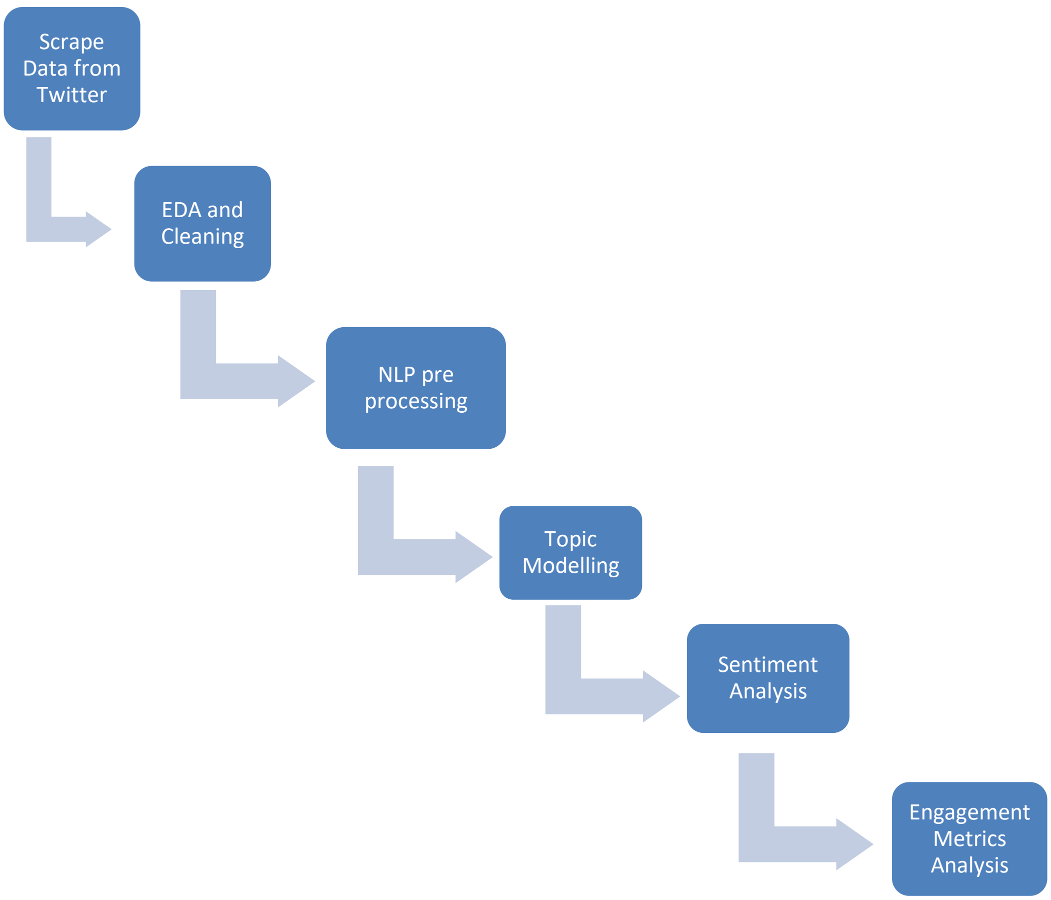
The tweets span the 2-year time-period from 2016-11-01 to 2018-10-31. The TwitterScraper library was called through an Anaconda command-prompt terminal, and tweets were extracted to CSV files for each of the 3 charities in 6-month intervals. The resulting CSV files were combined to create the final dataset containing 10,099 rows.

The dataset contains 10 Attributes:
user – text field; unique username in the twitter database
fullname – text; name as it appears currently on tweets and profile
tweet-id – numeric; unique numeric ID
timestamp – timestamp; date and time when the tweet was posted
url – text; URL extension where tweet can be displayed (following an initial: https://twitter.com...)
likes – numeric ; number of likes the tween has received to-date
replies – numeric ; number of replies the tween has received to-date
retweets – numeric; number of retweets the tween has received to-date
text – text; actual written content/message of the tweet
html – text; content of the tweet in HTML format

The analysis will be centered around the 'text' attribute. In a 'bucket-of-words' approach, this is the only attribute necessary to determine descriptive statistics like word frequencies, after sub-setting data related to the 3 charities. Further analysis will utilize the 'user' attribute to investigate qualities of tweets being generated by the 3 charities themselves, and to find specific users-of-interest (ex. 'most positive/negative sentiment'). Some further sentiment analysis will look at relationships between sentiment in the 'text' attribute and time-period ('timestamp' attribute); and sentiment in 'text' related to engagement metrics: 'likes', 'replies', and 'retweets'.

Approach

Create a block diagram for the steps of your approach to clearly provide an overview. For example, if you first scrapped twitter, second applied NLP techniques to extract keywords, third labelled the tweets as positive and negative using a set of keywords, and fourth build a classifier, then you should create a box for each of the steps with arrows connecting one step to the next one. A sample block diagram is shown below.



Step 1: Scrape Data from Twitter

Using the TwitterScraper python library, I will collect tweets related to World Vision Canada and 2 of its 'Peer' organizations, based on the presence of a hashtag or '@' mention. These 3 sets of tweets will be combined to form the full dataset for analysis.

Step 2: Conduct exploratory analysis and data cleaning

In this step, I will generate an in-depth understanding of the variables collected from twitter. I will then use descriptive statistics and plots to clean inconsistencies, making decisions about missing values and outliers.

Step 3: Apply NLP preprocessing techniques

In this step I will removing stop words and engaging in stemming or lemmatization. I will then 'tokenize' sub-sets of data for the following steps of analysis

Step 4: Conduct topic modelling based on keywords

After extracting keywords, I will analyze topics related to each of the 3 charity organizations of focus. Techniques employed will include word counts and TF-IDF. I will then create visualizations to present the results

Step 5: Perform Sentiment Analysis

Using external libraries (specifically the R 'sentiments' dictionaries), I will conduct a sentiment analysis and compare the results from the 3 charity organizations of focus. The results will be visualized in histograms and word clouds.

Step 6: Perform an analysis of engagement metrics

Finally, I will investigate the relationship between number of likes, retweets, and replies as related to the tweet topic and sentiment.

Results

<u>Notes</u>

In the data and in plots, the names of charities were shortened to the following:

cc = Compassion Canada (or 'Compassion')
pc = Plan Canada (or 'Plan')
wv = World Vision Canada (or 'World Vision')

<u>Exploratory Data Analysis</u>

The initial cleaned dataset contains 11 variables with 10,099 observations.

```
Observations: 10,099
Variables: 11
$ user      <fct> worldvisioncan, worldvisioncan, worldvisioncan, ch...
$ fullname  <fct> World Vision Canada, World Vision Canada, World Vi...
$ tweet.id  <dbl> 7.98654e+17, 7.98624e+17, 7.98614e+17, 7.98584e+17...
$ timestamp <date> 2016-11-15, 2016-11-15, 2016-11-15, 2016-11-15, 2...
```

```
$ url       <chr> "/worldvisioncan/status/798654414508605440", "/wor...
$ likes     <int> 1, 9, 1, 1, 4, 1, 4, 4, 3, 0, 1, 1, 117, 280, 32, ...
$ replies   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0,...
$ retweets  <int> 0, 5, 4, 2, 2, 1, 1, 1, 0, 0, 1, 3, 179, 378, 25, ...
$ text      <chr> "Proud to have one of our development managers spe...
$ html      <chr> "<p class=\"TweetTextSize js-tweet-text tweet-text...
$ tag       <fct> wv, wv, wv, wv, wv, wv, wv, wv, wv, wv, wv, wv, wv...
```

Most of the tweets came from the organizations themselves, with some other top-tweeters who appear to be staff of those charities (based on the content of their tweets).

Compassion Canada

```
Observations: 1,819
Variables: 11
              user                           fullname
 CompassionCA   :1014   Compassion Canada        :1014
 Mama2GreatKids :  86   Aimee Esparaz            :  86
 VisJeremy      :  33   Jeremy Vis               :  33
 Amyv93         :  29   Amy Smart                :  29
 FOTLProductions:  26   FrontOfTheLineProductions:  26
 loveismoving   :  25   Love is Moving           :  25
 (Other)        : 606   (Other)                  : 606
```

```
     likes           replies          retweets
 Min.   : 0.000   Min.   :0.0000   Min.   : 0.0000
 1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.: 0.0000
 Median : 2.000   Median :0.0000   Median : 0.0000
 Mean   : 2.518   Mean   :0.1056   Mean   : 0.6223
 3rd Qu.: 3.000   3rd Qu.:0.0000   3rd Qu.: 1.0000
 Max.   :67.000   Max.   :7.0000   Max.   :70.0000
```

Plan Canada

```
Observations: 3,502
Variables: 11
           user                              fullname
 PlanCanada  : 660   Plan Int'l Canada            : 660
 criseboro   : 110   Caroline Riseboro            : 110
 somto_ugwu  :  60   Ugwu Somto                   :  60
 skochschulte:  56   Sarah Koch-Schulte Â ðŸ‡¨ðŸ‡¦:  56
 PlanZimbabwe:  28   Plan Int' Zimbabwe           :  28
 1962sue     :  24   Sue Anderson                 :  24
 (Other)     :2564   (Other)                      :2564
```

```
     likes            replies           retweets
 Min.   :  0.00   Min.   : 0.0000   Min.   :  0.000
 1st Qu.:  0.00   1st Qu.: 0.0000   1st Qu.:  0.000
 Median :  2.00   Median : 0.0000   Median :  0.000
 Mean   :  4.72   Mean   : 0.2487   Mean   :  1.935
 3rd Qu.:  5.00   3rd Qu.: 0.0000   3rd Qu.:  2.000
 Max.   :421.00   Max.   :27.0000   Max.   :146.000
```

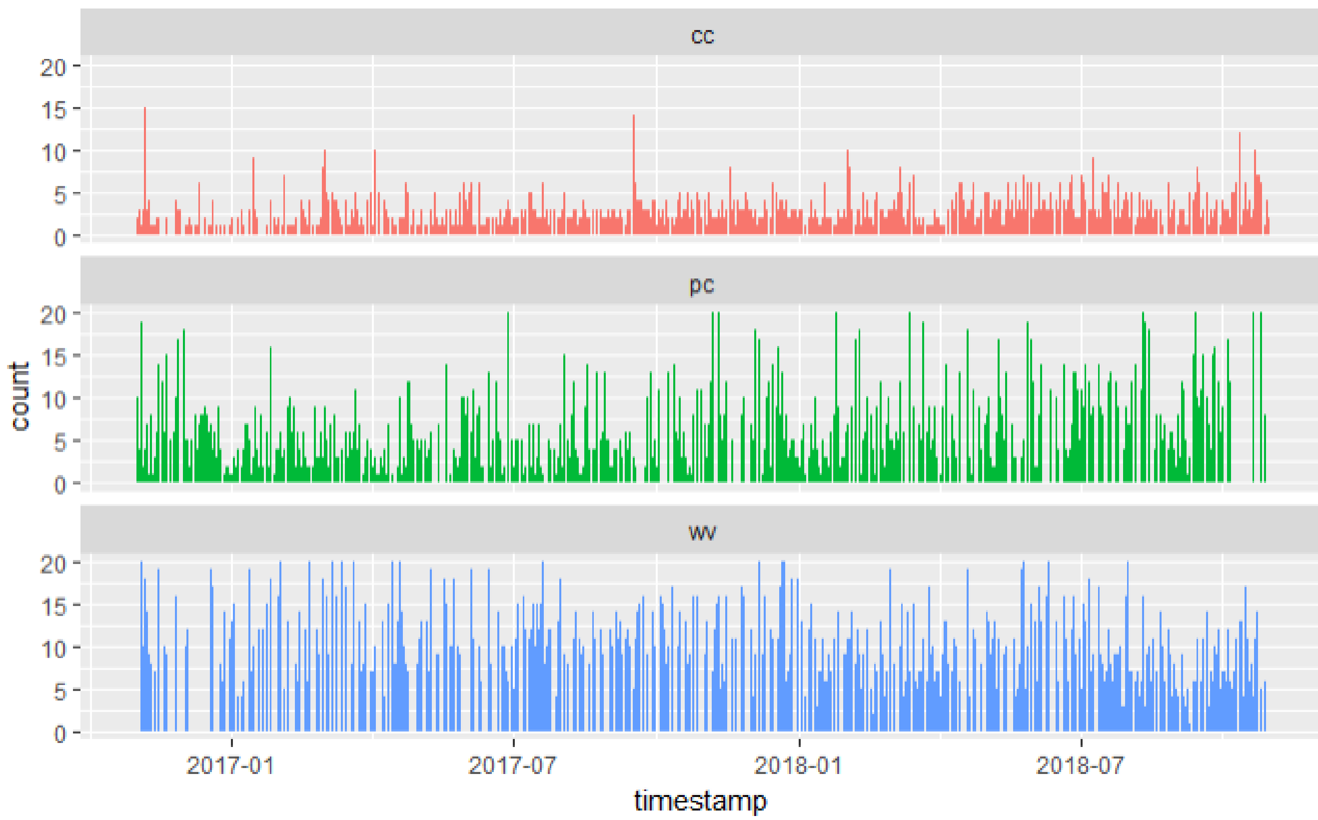
World Vision Canada

```
Observations: 4,778
Variables: 11
              user                     fullname          tweet.id
 worldvisioncan :2334   World Vision Canada:2334   Min.   :7.944e+17
 mjmessenger    :  56   Michael Messenger  :  56   1st Qu.:8.689e+17
 WVHungerFree   :  42   HungerFree         :  42   Median :9.321e+17
 BlogsbyFriends :  41   Friends            :  41   Mean   :9.315e+17
 WorldVisionJobs:  41   We Are World Vision:  41   3rd Qu.:9.958e+17
 WorldVisionLAC :  41   World Vision LAC   :  41   Max.   :1.057e+18
 (Other)        :2223   (Other)            :2223
```

```
    likes            replies           retweets
Min.   :  0.000   Min.   : 0.000   Min.   :  0.000
1st Qu.:  1.000   1st Qu.: 0.000   1st Qu.:  0.000
Median :  2.000   Median : 0.000   Median :  1.000
Mean   :  5.333   Mean   : 0.185   Mean   :  2.208
3rd Qu.:  6.000   3rd Qu.: 0.000   3rd Qu.:  3.000
Max.   :540.000   Max.   :14.000   Max.   :378.000
```

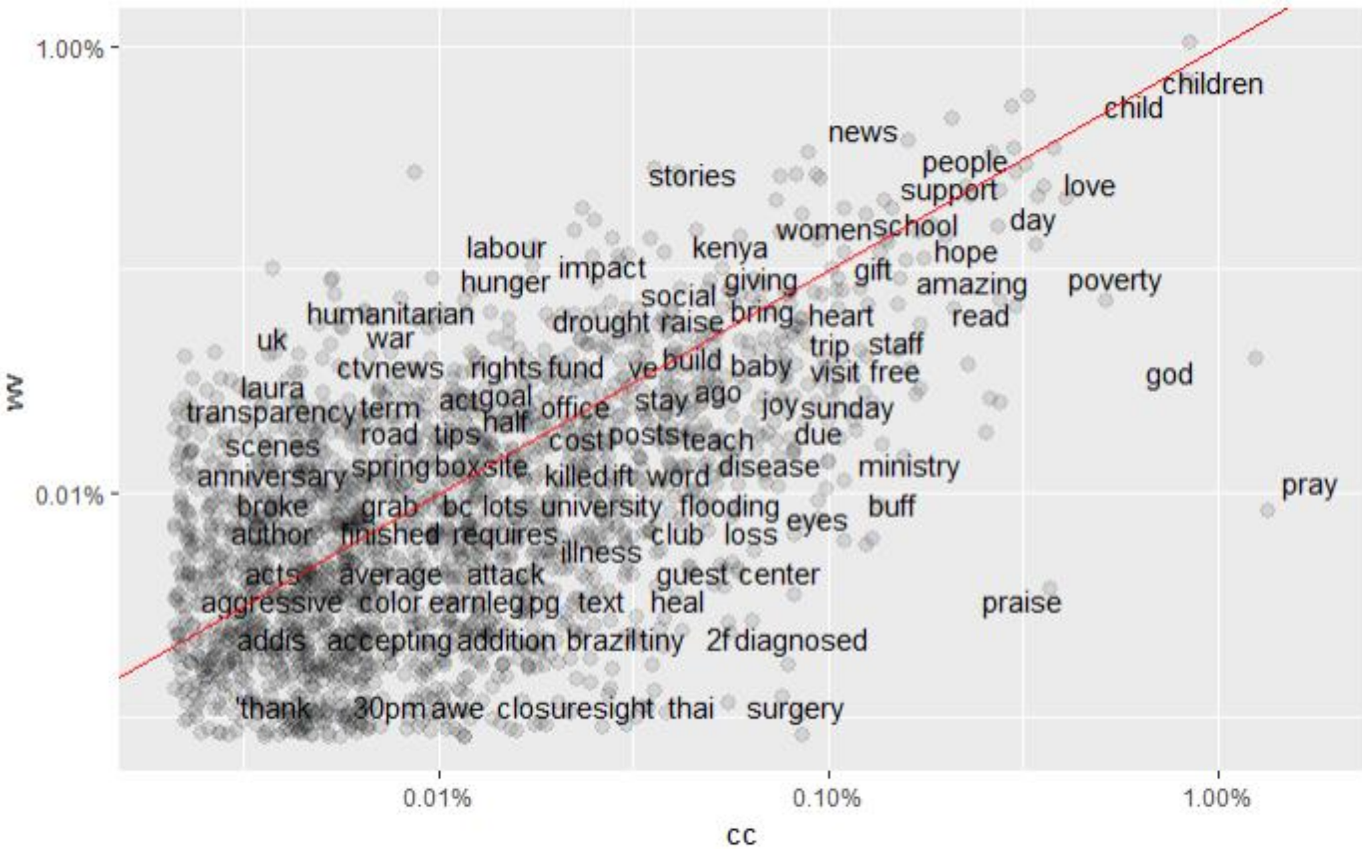
The engagement variables related to each charity (likes, replies, and retweets) are similar in their median values, with World Vision slightly higher in retweets. World Vision and Plan have higher mean values for 'likes' and Plan has a higher mean value for 'replies', but this is likely due to some unusually high values in a small subset of tweets.

Comparing Word Frequencies

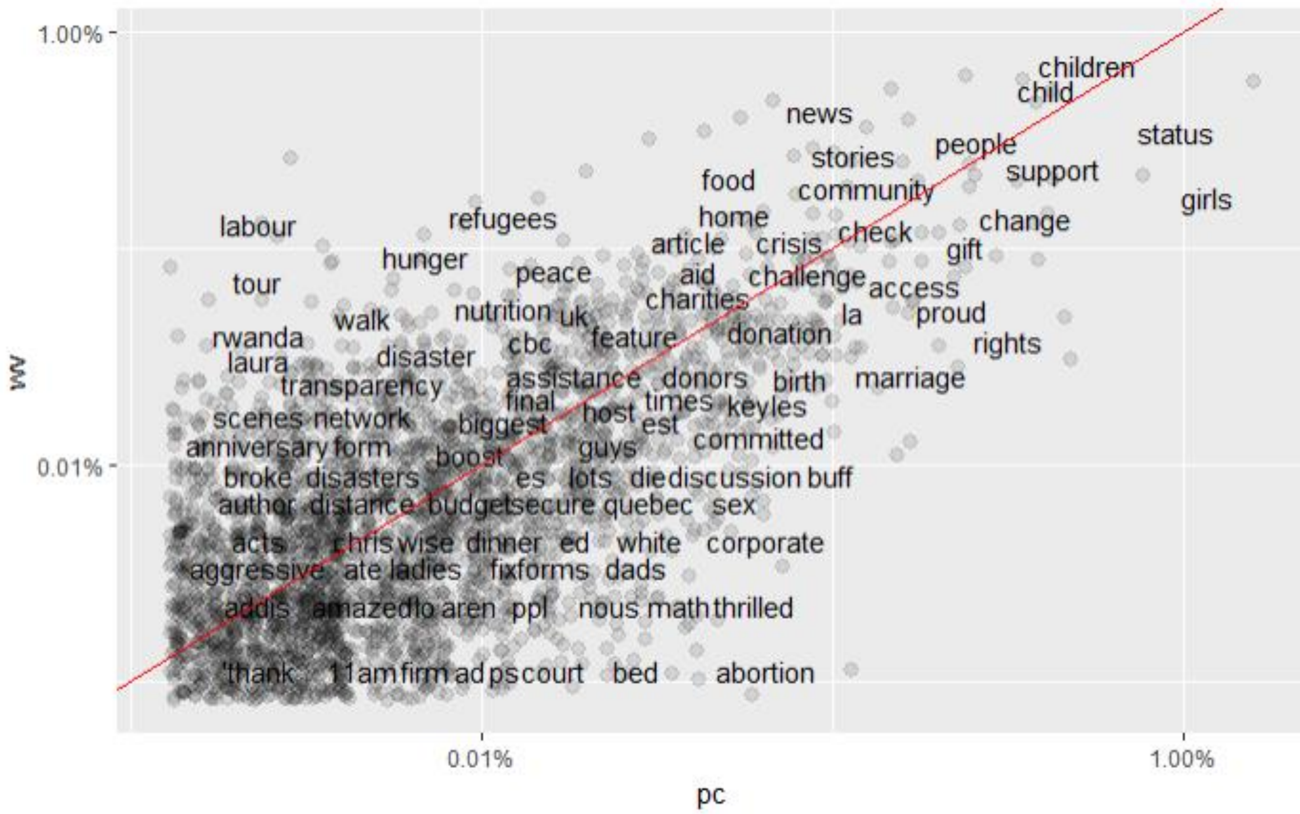


The data spans a 2-year period, and the above plot shows a comparison of the frequency of tweets associated with each charity across that period. No single day exceeded 20 tweets, and there were days in which there were no tweets. World Vision had the highest consistent volume of tweets, whereas Compassion Canada appears to have the shortest breaks between days with associated tweets.

The following analysis looks at the data after cleaning the text by removing: stop words, frequent-but-uninformative words (like the names of the organization, or 'twitter'), URLs, @ mentions, and hashtags. This will provide a better picture of the true differences in the content these charities are communicating or associated with.

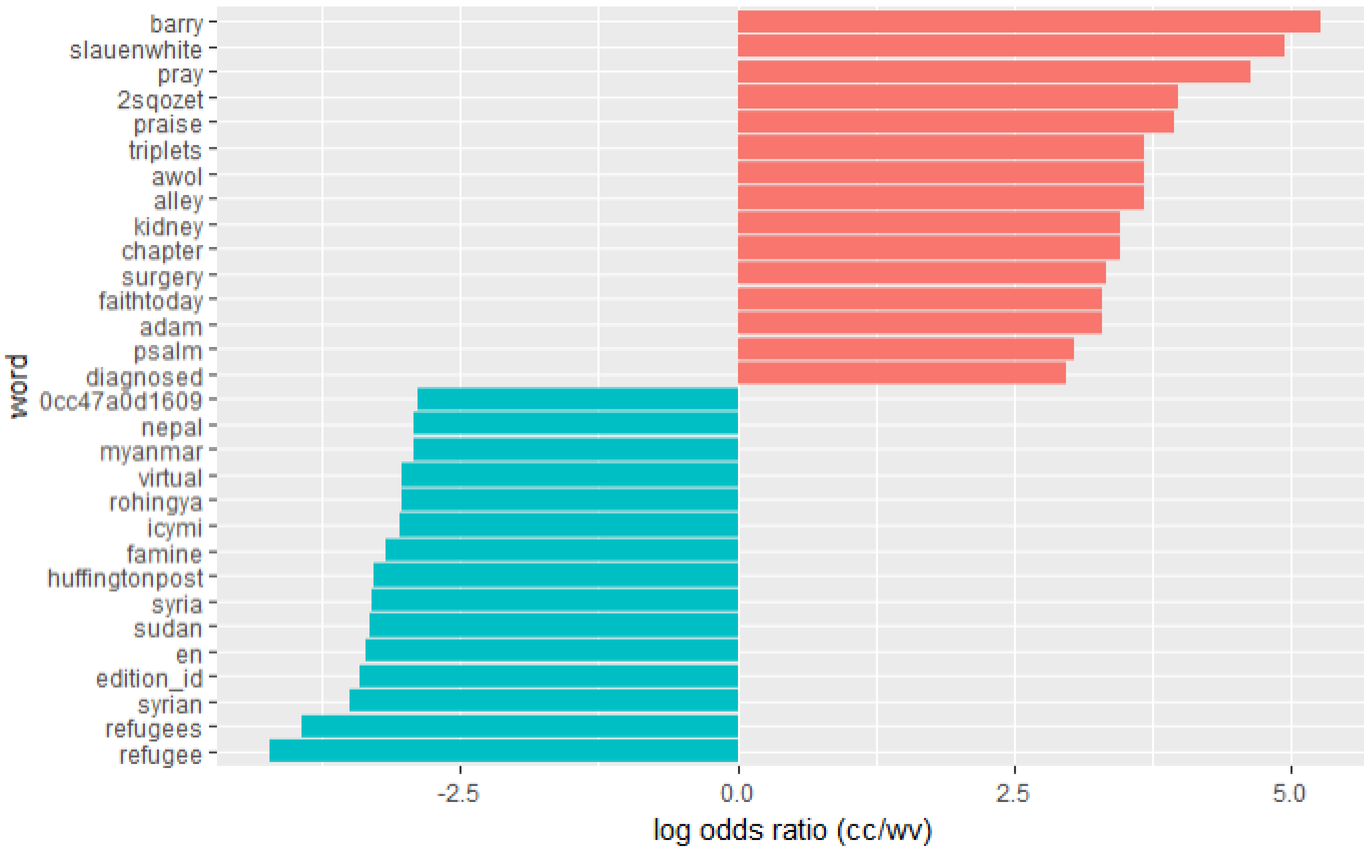


The plot above shows the most frequent words associated with World Vision and Compassion Canada; grouping words with similar frequencies close to the red line, and unique words further from it. Both charities are associated with verbs in similar frequency: ex. ‘finished’, ‘build’, ‘bring’, ‘support’; and nouns: ex. ‘school’, ‘child’, and ‘children’. These are both child-focused charities who would often request support and who report on finishing projects. World Vision is associated with words like ‘humanitarian’, ‘labour’, ‘stories’, and ‘news’ more frequency; and Compassion Canada is associated with words like ‘surgery’, ‘praise’, ‘pray’, and ‘god’ more frequently. While both charities are Christian, Compassion seems to be more overt in its use and association with religious language.

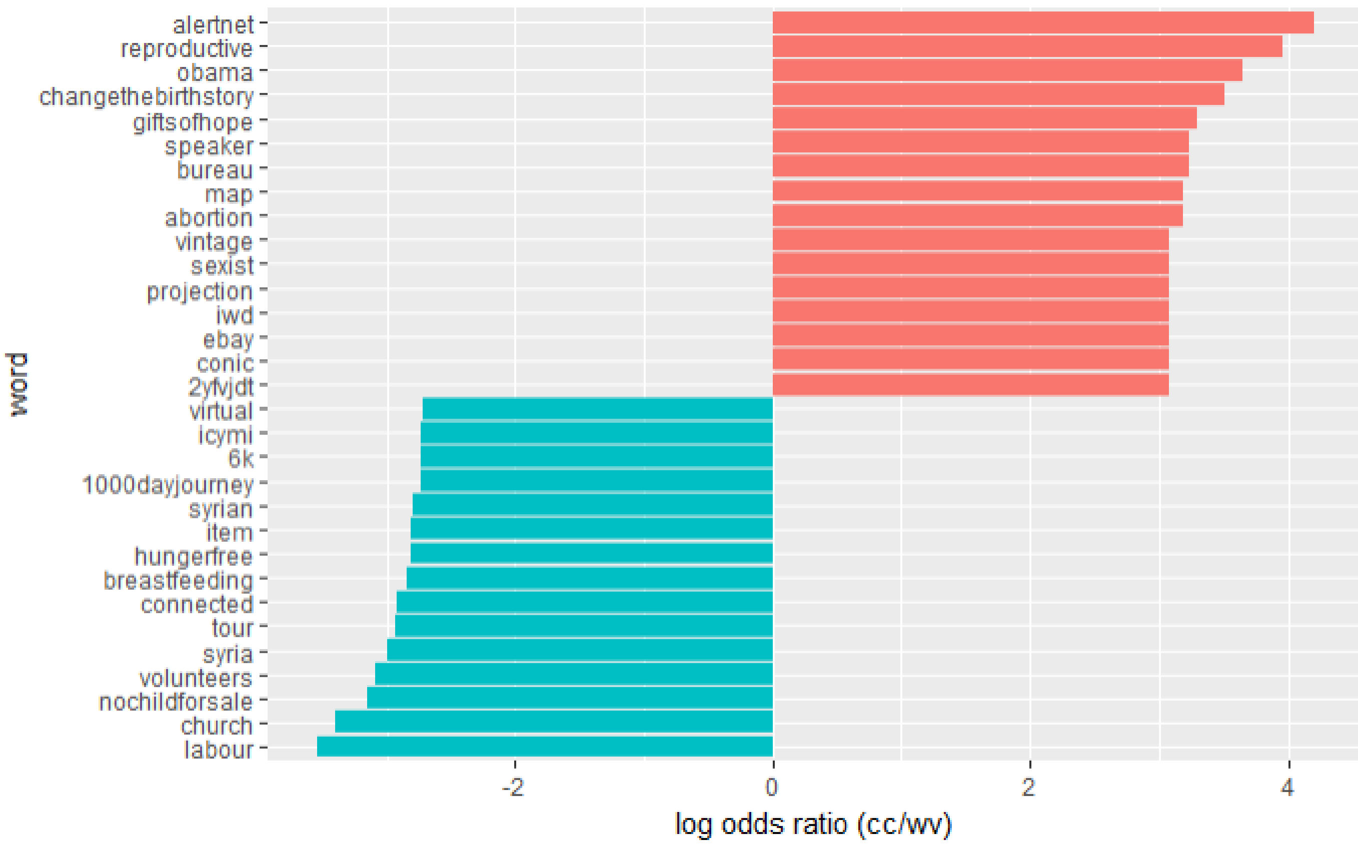


In contrast, the comparison of World Vision and Plan Canada show association with similar frequency words like: ‘biggest’, ‘assistance’, ‘challenge’, ‘check’, ‘people’, ‘child’, and ‘children’. These relate to the fact that both charities work in emergency disaster situations and have a focus on children. Word Vision is associated with words like ‘labour’, ‘hunger’, ‘refugees’, ‘food’, and ‘news’ more frequently; and Plan with words like ‘abortion’, ‘marriage’, ‘rights’, and ‘girls’ more frequently. Plan Canada has a specific focus on the rights of girls and women, which is seen in these particular words.

Looking more statistically at the unique words per organization, we can compare the ‘log odds ratio’ of words occurring to find to top 15 most unique words, when comparing World Vision to the other two charities respectively.

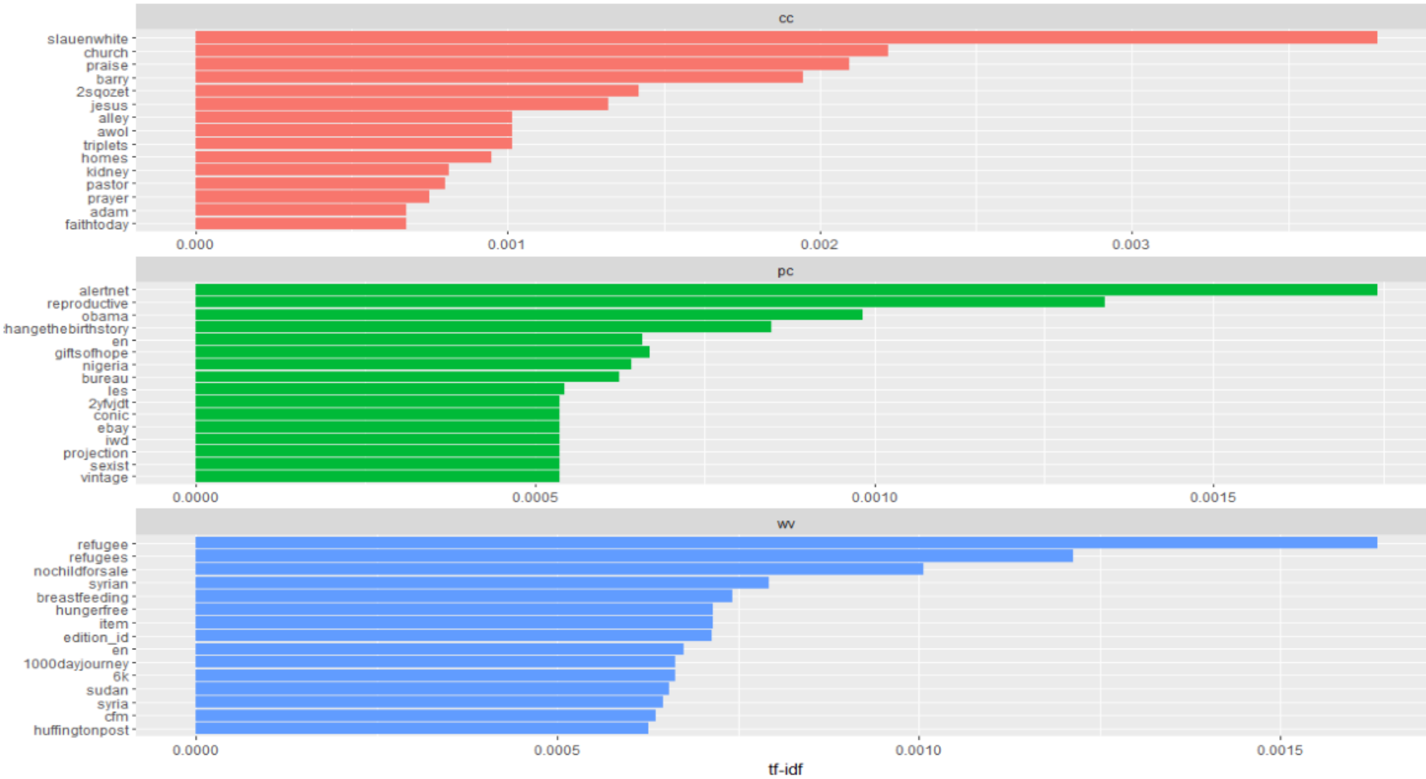


The upper (red) bars represent the most unique words associated with Compassion Canada, as compared to words associated with World Vision. For Compassion, we see the words ‘barry’ and ‘slauenwhite’ – Barry Slauenwhite is the CEO of Compassion Canada, and his name is included frequently in tweets. Again, we see religious words: ‘pray’, ‘praise’, ‘faithtoday’ and ‘psalm’. The top words for World Vision Canada have a predominant emphasis on emergency crisis countries and topics (Nepal experienced an earthquake during this time period, and there were refugee crises in Sudan, Syria, and Myanmar).



World Vision in comparison with Plan Canada (upper red bars) shows some different unique words of interest – including awareness and marketing campaigns like ‘no child for sale’, ‘hunger free’, and ‘1000 day journey’ (a health and nutrition program that promoted ‘breastfeeding’ of infants). Plan Canada’s unique words include ‘alertnet’, a disaster relief information hub. Plan also ran specific campaigns entitled ‘change the birth story’ and ‘gifts of hope’. Interestingly, some political words emerge: ‘obama’ and ‘bureau’.

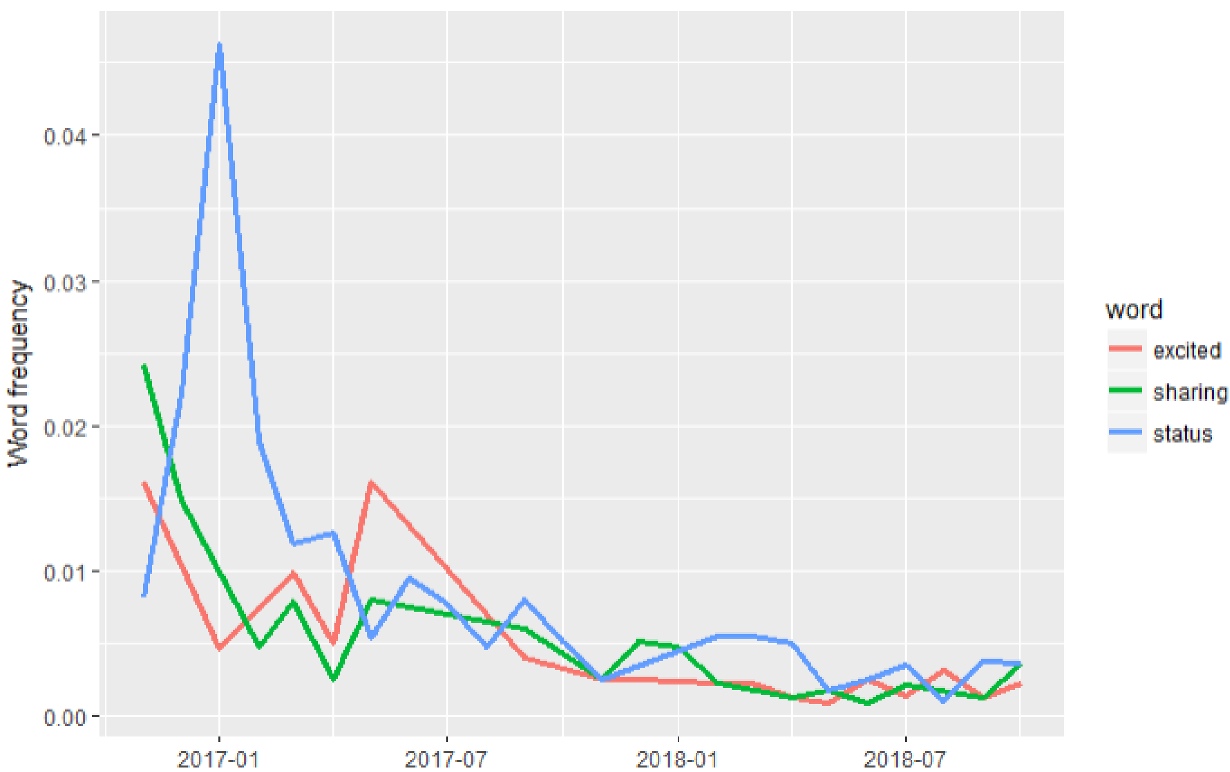
Term-Frequency Inverse Document Frequency (TF-IDF) is another useful technique in finding unique words:



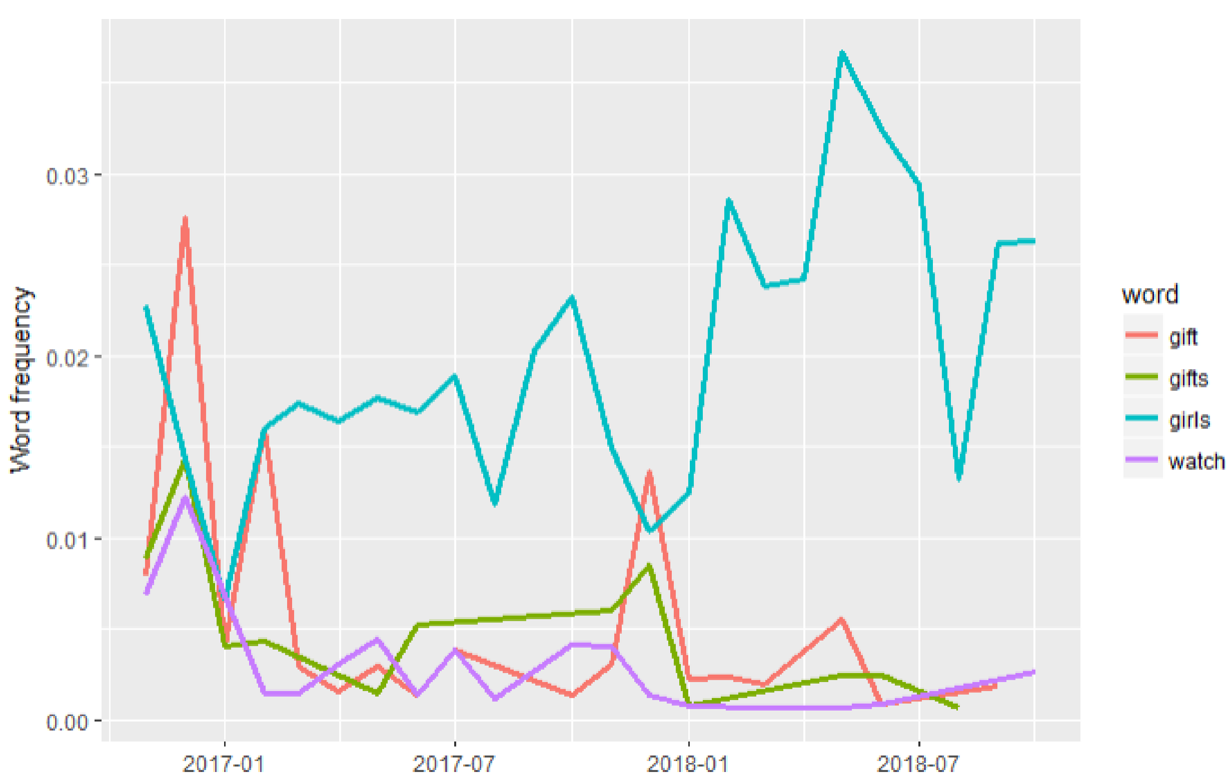
TF-IDF can be a useful technique to filter out frequently used, but also common words. Many of these top words are the same as the previous analysis – with some new words like 'church' and 'jesus' associated with Compassion Canada, and 'huffingtonpost' associated with World Vision.

Word usage may change over time, and this may indicate some interesting trends. The following analysis counts the frequency of words in 1-month time periods and plots their changes over time.

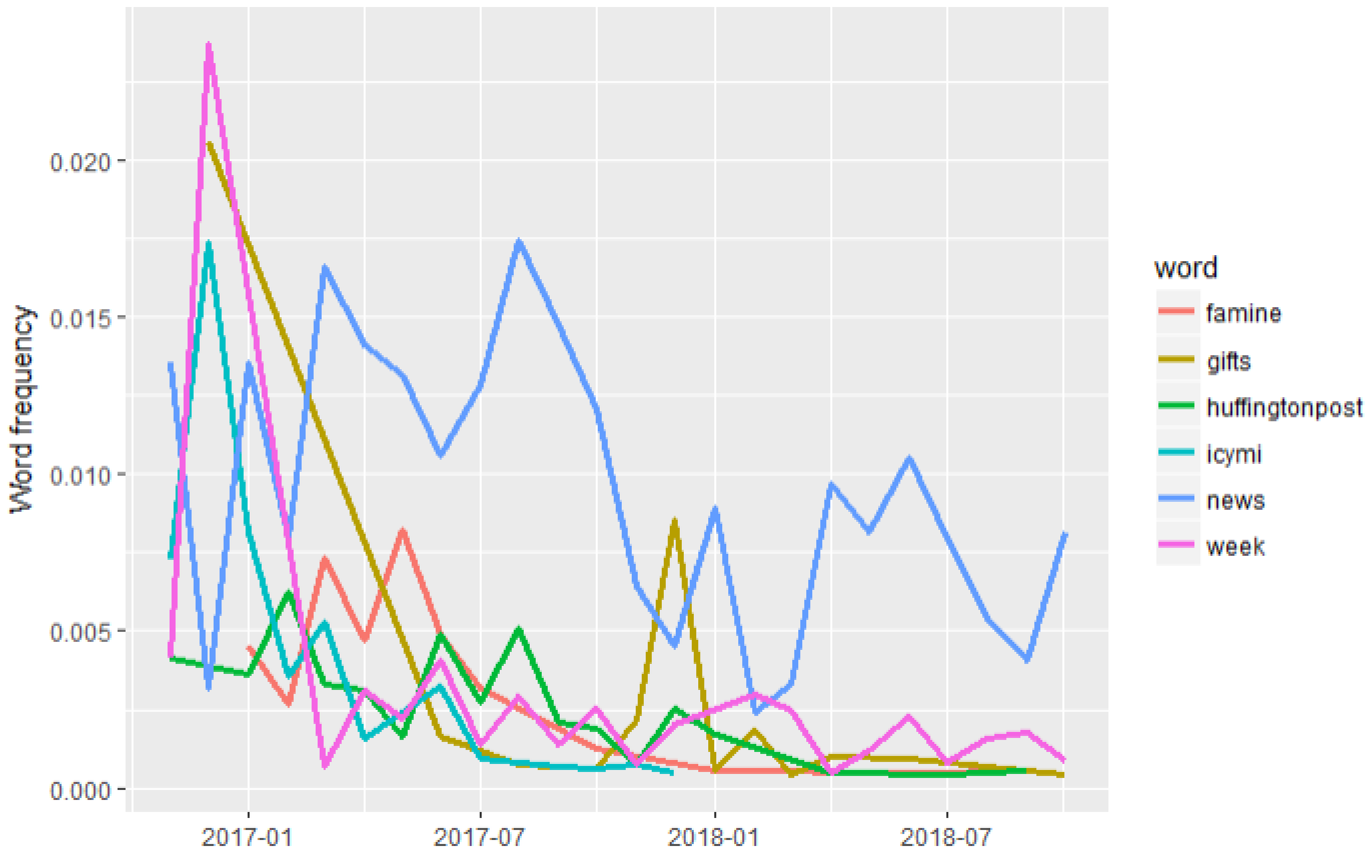
Compassion Canada



Plan Canada



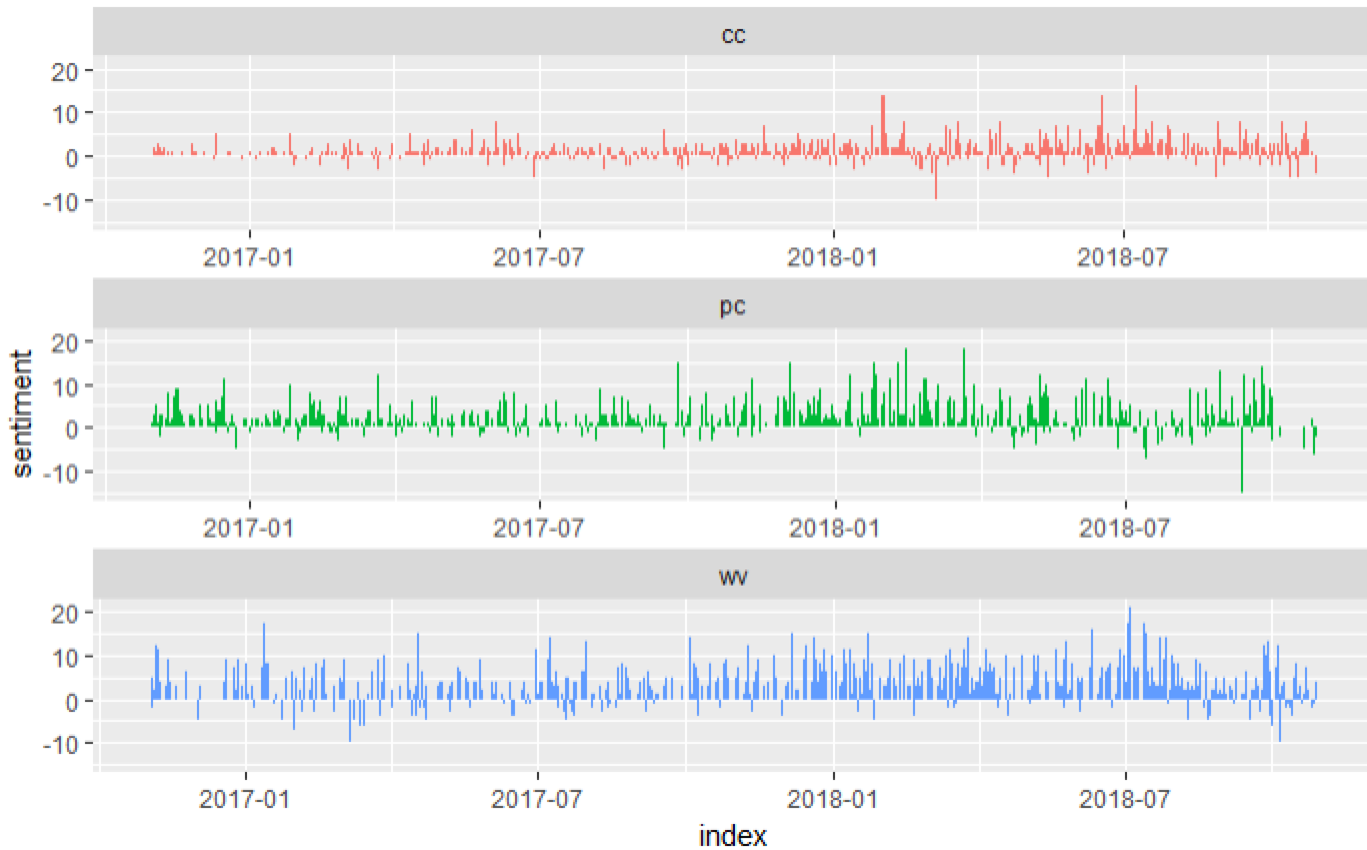
World Vision Canada



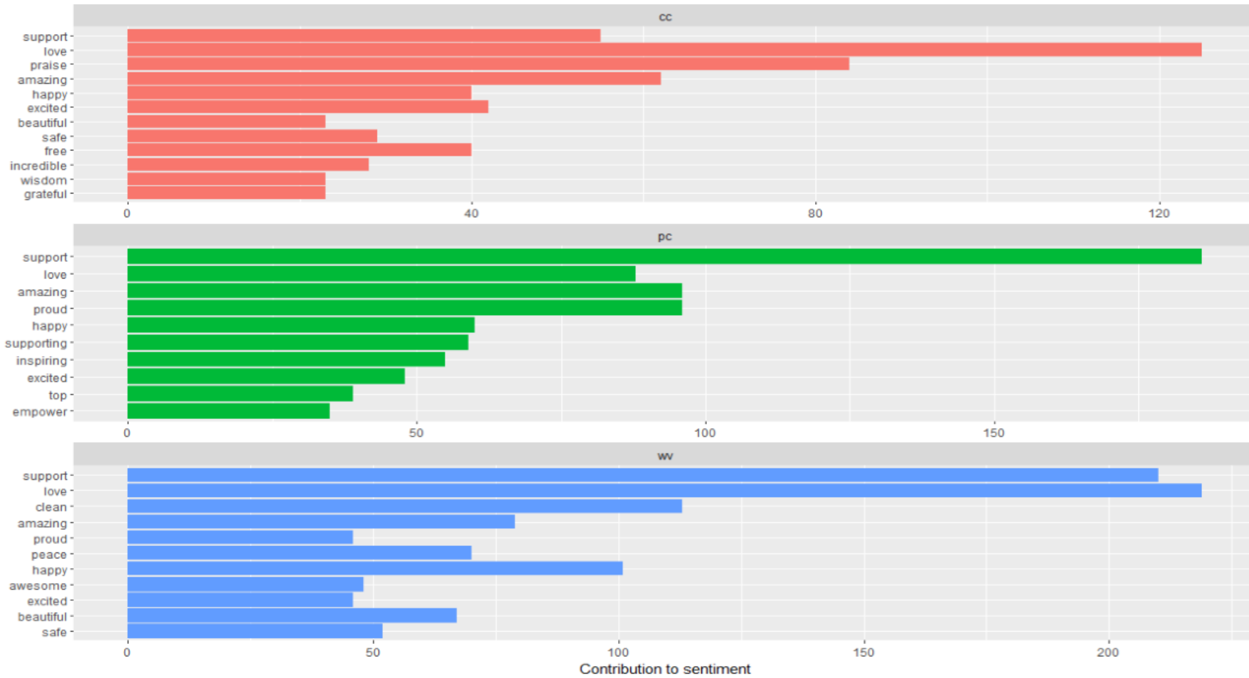
Most frequently used words were more common at the beginning of the data time period, suggesting that all three charities have become more varied in their language and topics. The expectations are ‘girls’ relating to Plan and ‘news’ relating to World Vision. The words ‘gift’ and ‘gifts’ have a rise in the months preceding Jan. 2018, which is certainly due to these charities promoting their gift catalogues – where a person in Canada donates on behalf of a family member as a Christmas gift, donating towards specific items that help programs internationally (ex. medical supplies or livestock).

Sentiment Analysis

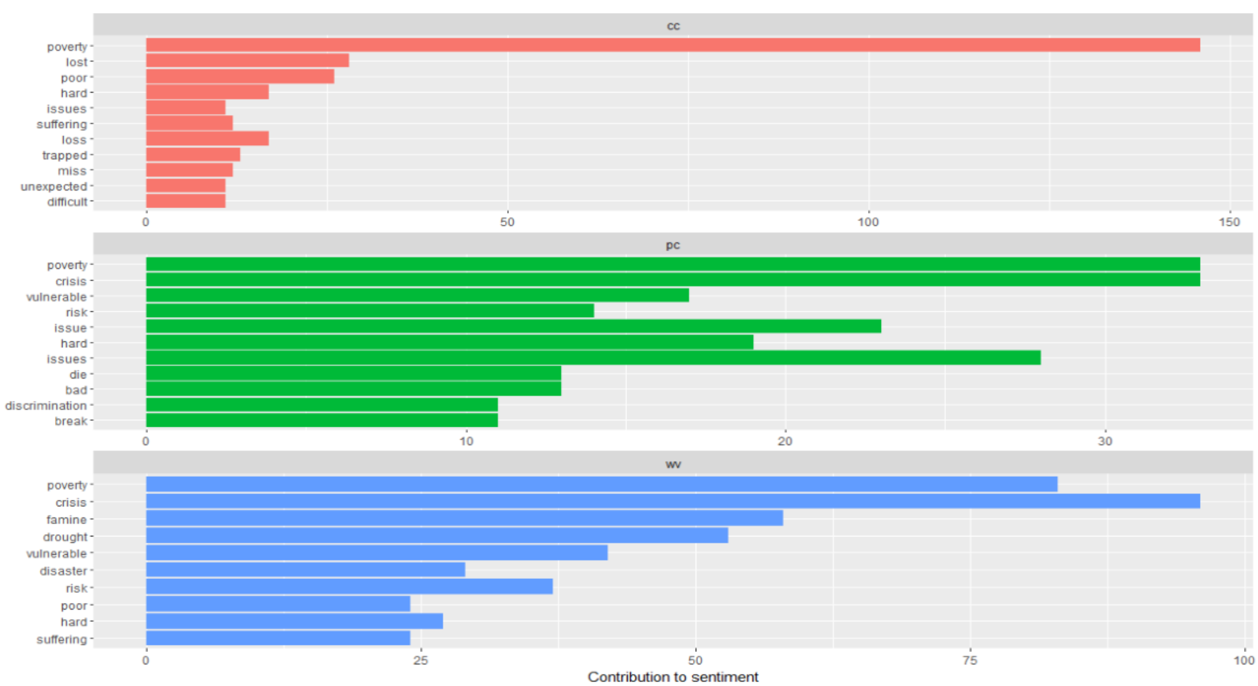
The sentiment used in tweets is also a very important factor in engaging audiences. The following plot shows the frequency of positive and negative words in tweets over time.



The majority of tweets have a positive sentiment. Words relating to all charities have a more varied sentiment (both positive and negative) as time progresses. There are also some significant outliers – particularly a very negative tweet (or tweets) associated with Plan Canada in December 2018.



The most common positive words per charity all show ‘support’ as the most frequent positive word. These charities are often asking the public to ‘support’ their work, or demonstrating their accomplishments as a result of donations from supporters. ‘love’, ‘amazing’, and ‘beautiful’ are also frequently used positive words. Uniquely, ‘wisdom’ is associated with Compassion Canada, ‘empower’ with Plan, and ‘peace’ with World Vision.



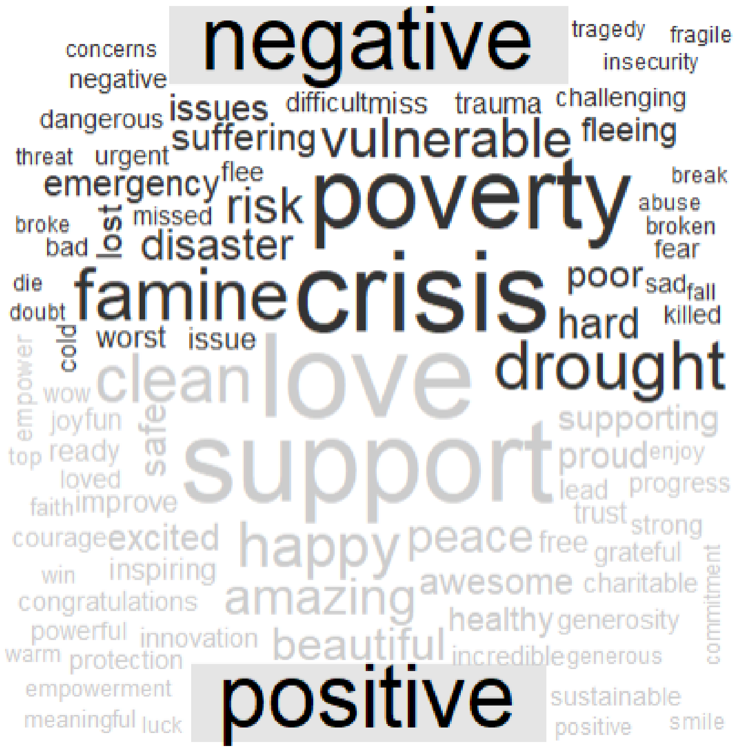
Similarly, the three charities share the top most frequent negative word: ‘poverty’. They also share words like ‘difficult/hard’, and ‘suffering’. Other words are more unique to each charity: ‘lost’ and ‘miss’ associated with Compassion, ‘die’ and ‘discrimination’ associated with Plan, and ‘famine’ and ‘risk’ associated with World Vision.

These same findings can be visualized in the form of wordclouds.

Compassion Canada

Plan Canada

World Vision Canada



It is informative to see the context in which some of these words occur. By finding the mean sentiment score per tweet, here are some examples of the tweets with the lowest (most negative) mean sentiment score:

Compassion Canada

A horrific attack by fire: one year later https://www.compassion.ca/blog/a-horrific-attack-by-fire-one-year-later/Â â€¦ via @CompassionCA

Plan Canada

I swear charities are accelerating this Age of Disillusionment same way the Daily Show degraded political discourse into playground bullying.

World Vision Canada

"Violence affects more than 1.7 billion children every year...The good news is that there's growing evidence of the most effective solutions to end violence against children & mounting public pressure that it will no longer be tolerated

And alternatively, examples of the tweets with the highest (most positive) mean sentiment score:

Compassion Canada

Amazing. Amazing. What a generous, passionate, driven church. So grateful. (@CompassionCA) https://twitter.com/ssundby/status/871570849509261312Â â€¦

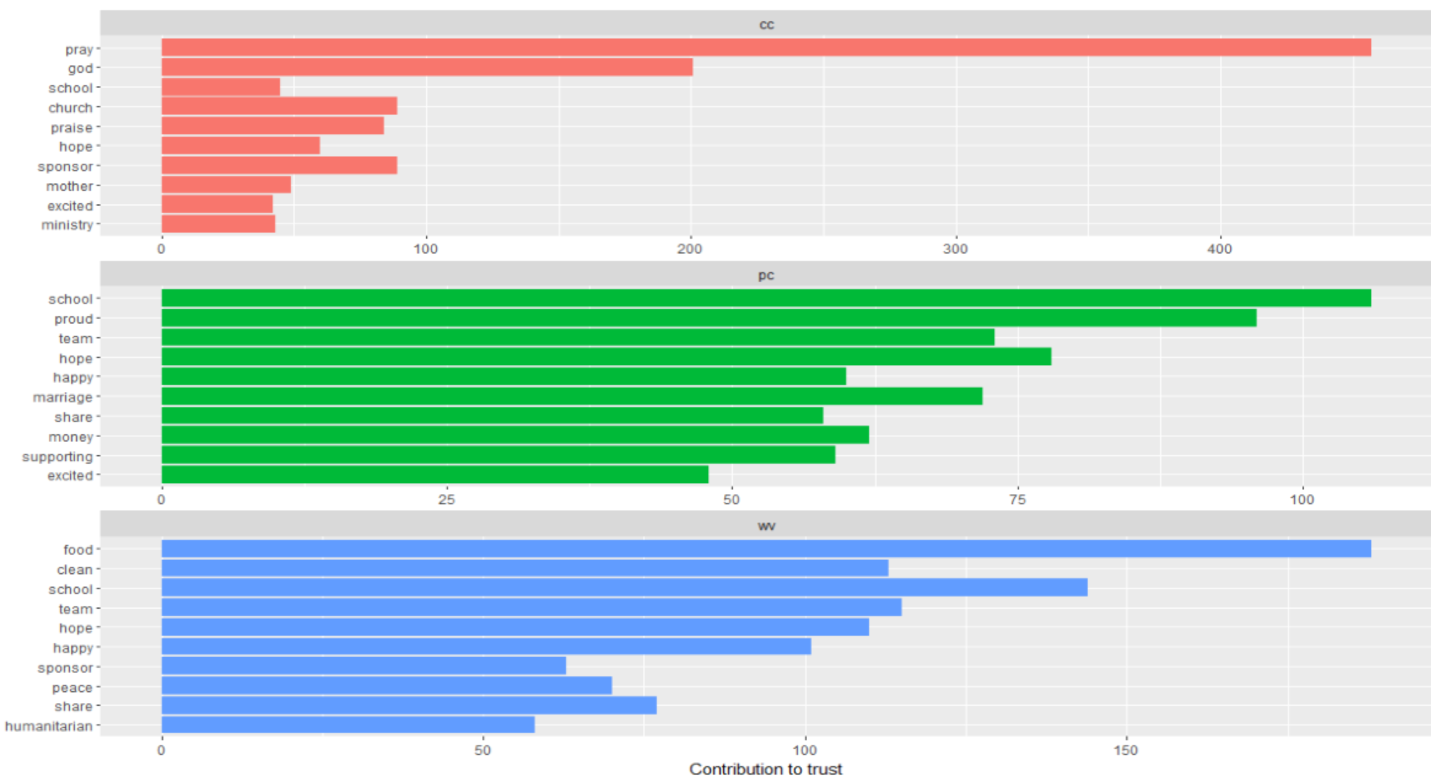
Plan Canada

Pretty amazing to have been part of helping to shape this G7 focus on girls education in my role as @PlanCanada CEO. Even more thrilled to see our PM @JustinTrudeau creating global momentum to make this historic commitment a reality.

World Vision Canada

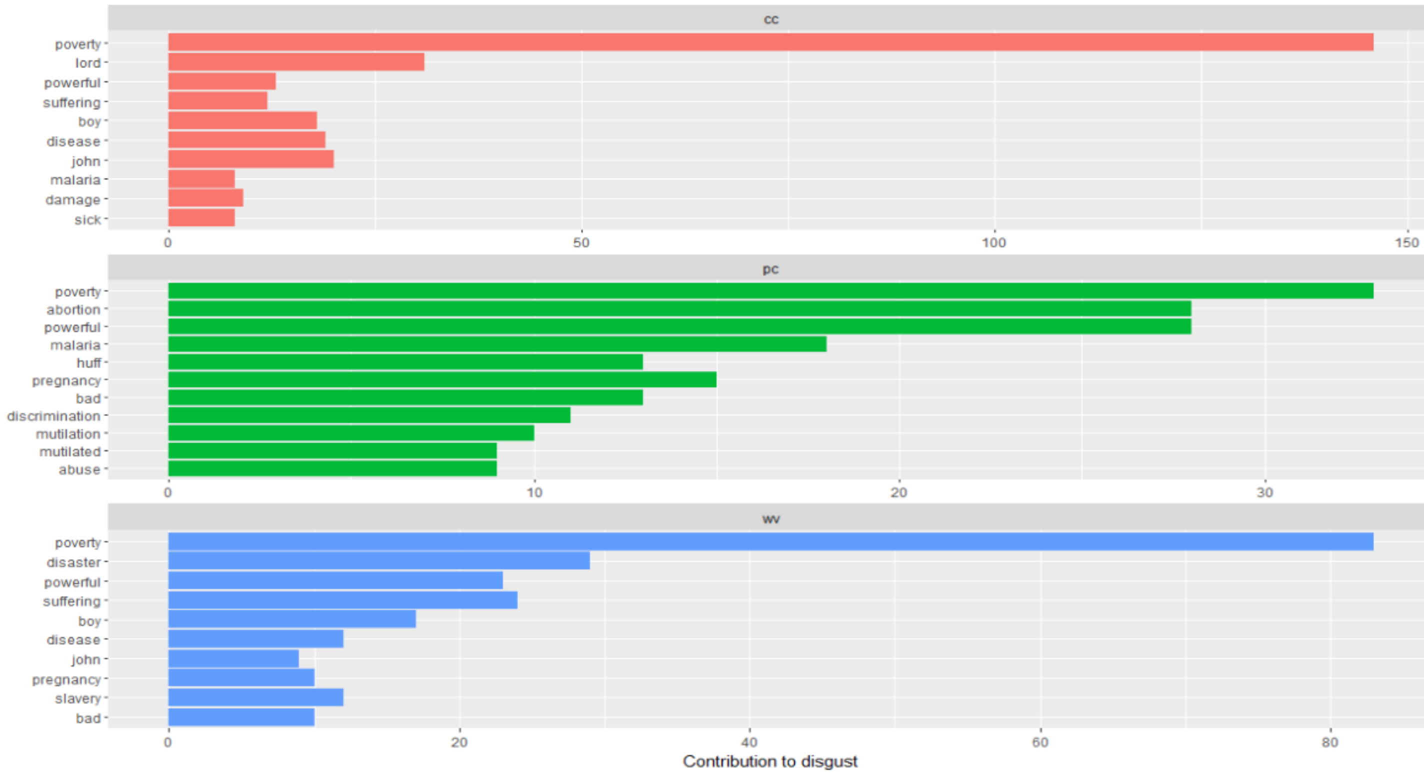
Visit @worldvisioncan to see how they're making a difference in this world and what you can do to help \n#nonprofit #givingtuesday #hungerfreepic.twitter.com/M3My4i5FlA

Using a different tool, we can look beyond 'positive' and 'negative' sentiment to words like 'trust'. Here are the top words associated with 'trust' for each of the three charities:



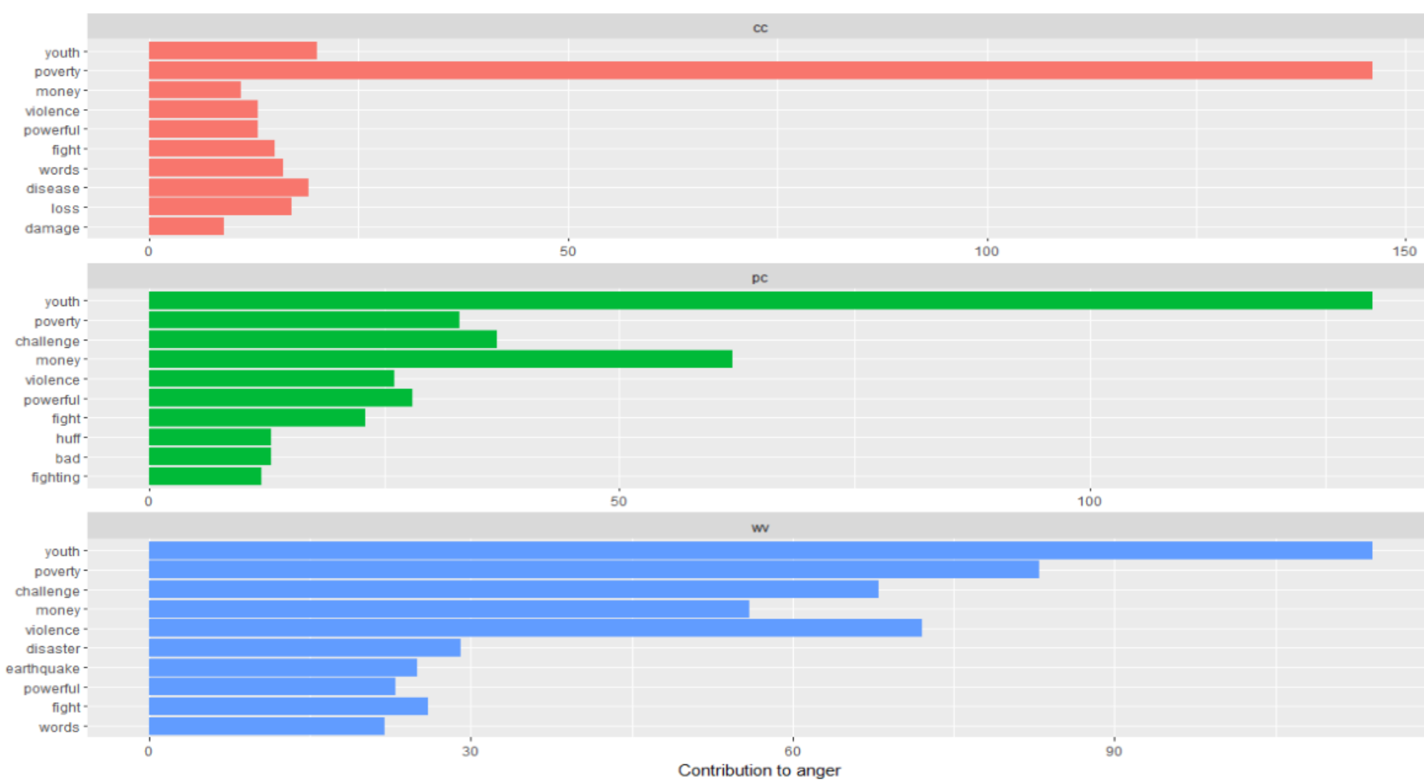
Top words associated with Compassion include 'pray', 'god' and 'sponsor'.
Top words associated with Plan include 'school', 'proud' and 'hope'.
Top words for World Vision include 'food', 'clean' and 'team'.

Words associated with 'disgust' give a more nuanced view of negative sentiment words:



‘Poverty’ and ‘powerful’ are both frequently used word associated with disgust for all charities.
Top words associated with Compassion also include ‘lord’, ‘suffering’ and ‘disease’
Top words associated with Plan also include ‘abortion’, ‘malaria’ and ‘discrimination’
Top words for World Vision also include ‘disaster’, ‘disease’ and ‘slavery’

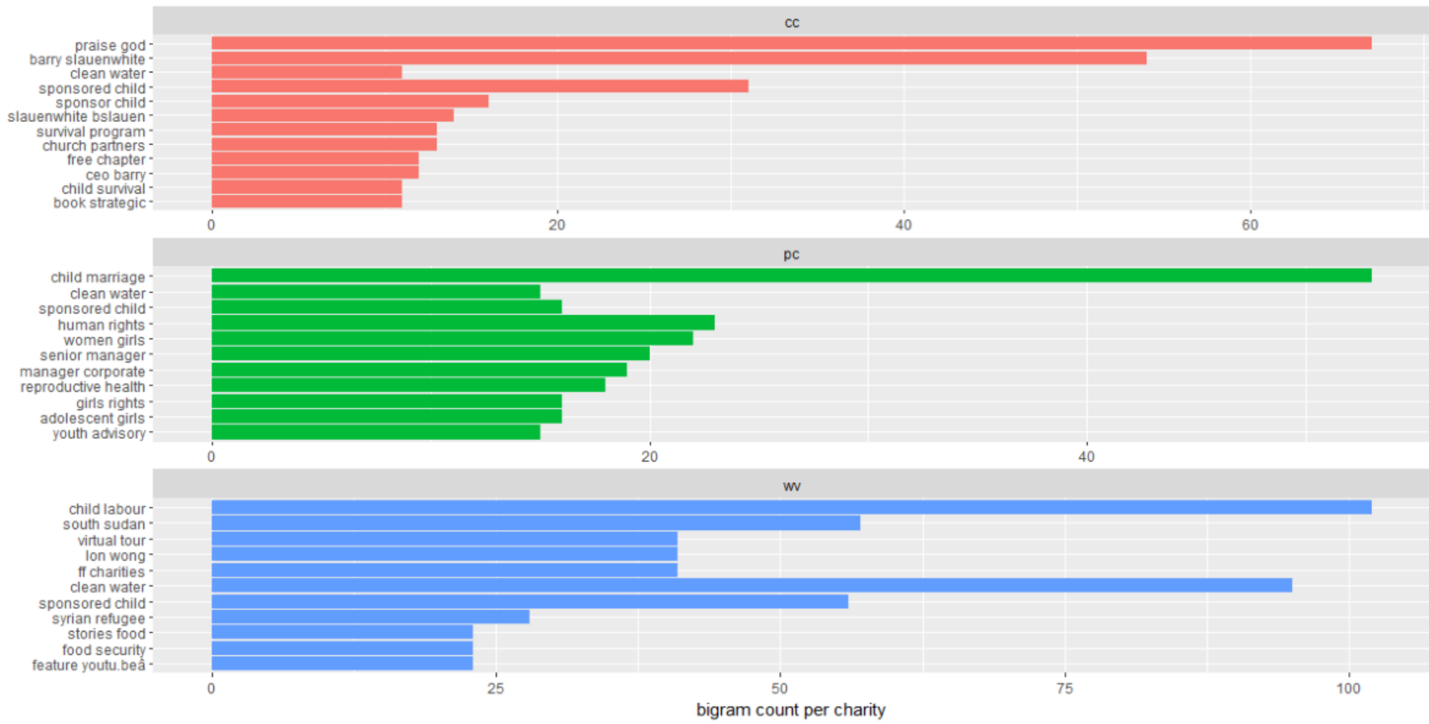
And words associated with ‘anger’ give a different view of association with negative sentiment words:



‘youth’, ‘poverty’, ‘money’, and ‘violence’ are all frequently used word associated with anger for all charities.
Top words associated with Compassion also include ‘disease’ and ‘damage’
Top words associated with Plan also include ‘fight’ and ‘bad’
Top words for World Vision also include ‘earthquake’ and ‘challenge’

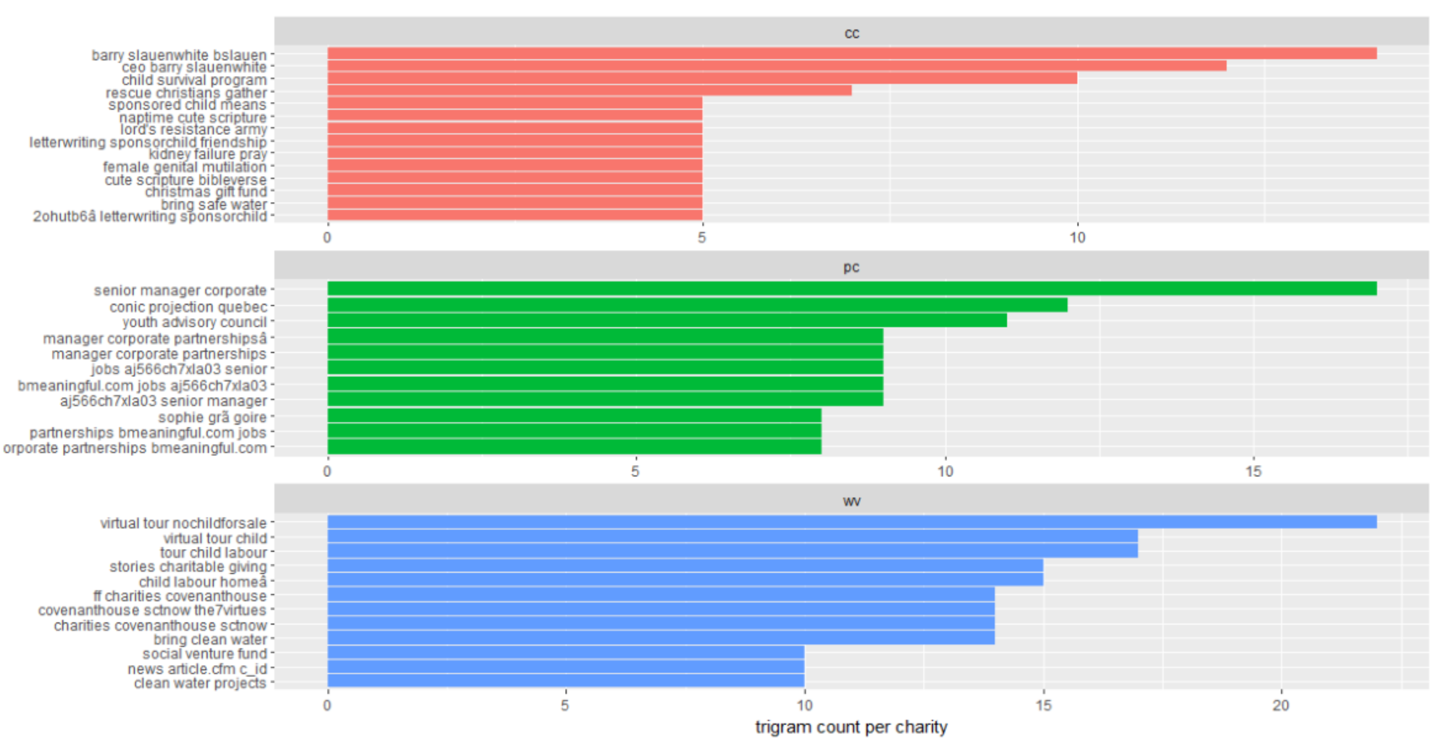
Single words are informative, but looking at the occurrence of two or more words brings more context to the analysis. The most frequent bi-grams in this dataset include:

bigram <chr>	**n** <int>
clean water	121
child labour	106
sponsored child	103
praise god	67
child marriage	65
south sudan	63
barry slauenwhite	54
ff charities	41
lon wong	41
virtual tour	41
child sponsorship	34
human rights	30
syrian refugee	30
feature youtu.beâ	26
food security	25
sponsor child	25
child protection	24
women girls	24
stories food	23



The most frequently-used bi-grams associated with each charity include:
Compassion Canada: ‘praise god’; ‘sponsored child’; ‘barry slauenwhite’; and ‘church partners’
Plan Canada: ‘child marriage’; ‘clean water’; ‘human rights’; and ‘women girls’
World Vision: ‘clean water’; ‘child labour’; ‘south sudan’; and ‘sponsored child’

The unique brand and areas of focus are much more visible when looking at bi-grams. Whereas the most frequent uses for tweets start to emerge when looking at tri-grams.



The most frequently-used tri-grams associated with each charity include:
Compassion Canada: ‘barry slauenwhite bslauen’; ‘child survival program’; and ‘rescue christians gather’
Plan Canada: ‘senior manager corporate’; ‘youth advisory council’; and ‘manager corporate partnerships’
World Vision: ‘virtual tour nochildforsale’; ‘stories charitable giving’; and ‘bring clean water’

From this analysis, one can assume that Compassion tends to use twitter to update about their CEO and children, Plan to advertise job vacancies, and World Vision advocating for causes such as anti-child labour and clean water projects

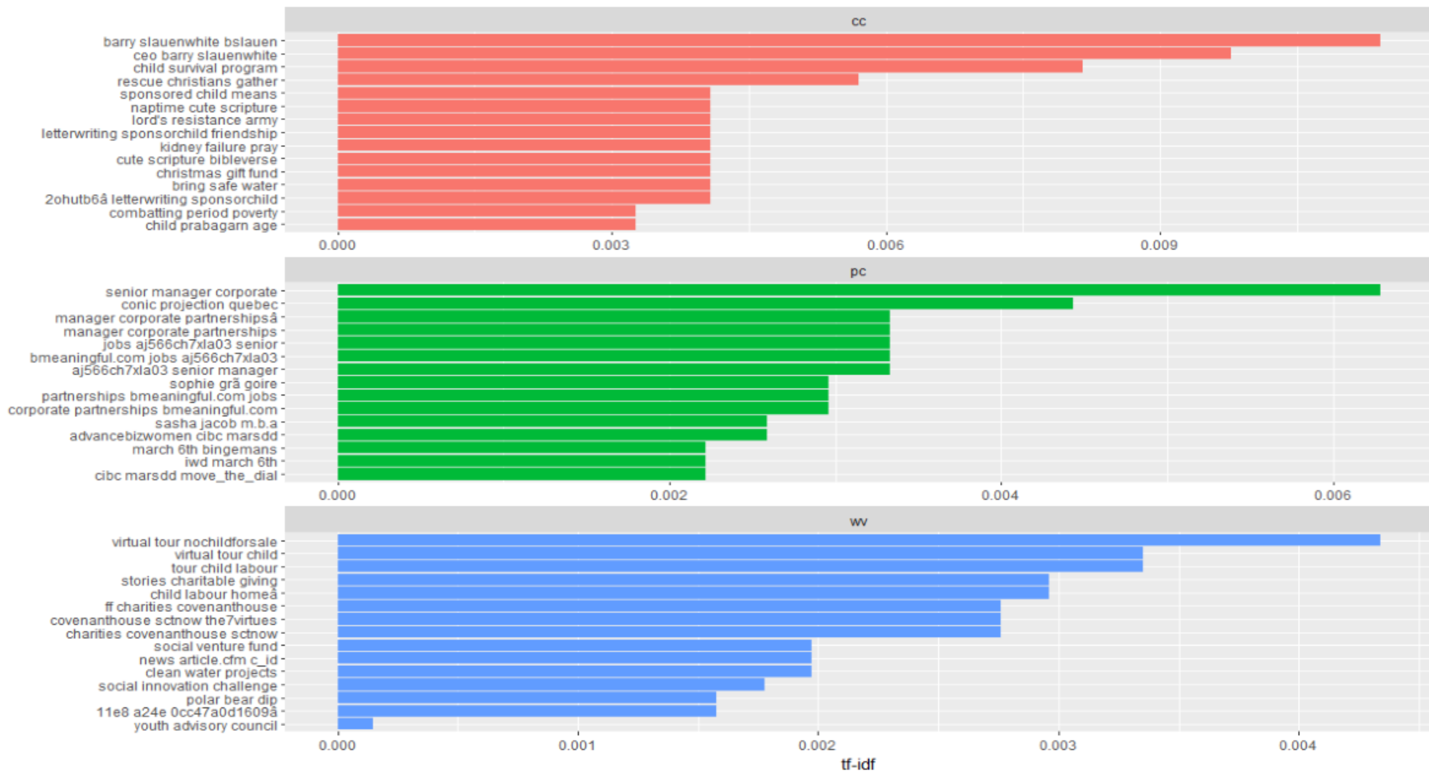
Bi-grams containing specific words of interest can also be investigated, for example, most frequent words that follow the word ‘child’:

tag <fctr>	**word2** <chr>	**n** <int>
wv	labour	102
pc	marriage	53
wv	sponsorship	19
wv	labourers	14
wv	protection	14
wv	health	12
wv	marriage	12
cc	survival	11
wv	refugees	10
wv	soldiers	9

Or most frequent words that preceded the word ‘refugees’:

tag <fctr>	**word1** <chr>	**n** <int>
wv	syrian	18
wv	child	10
wv	rohingya	9
wv	stories	7
wv	messenger	2
wv	myanmar	2
wv	protect	2
wv	sudan	2
wv	sudanese	2

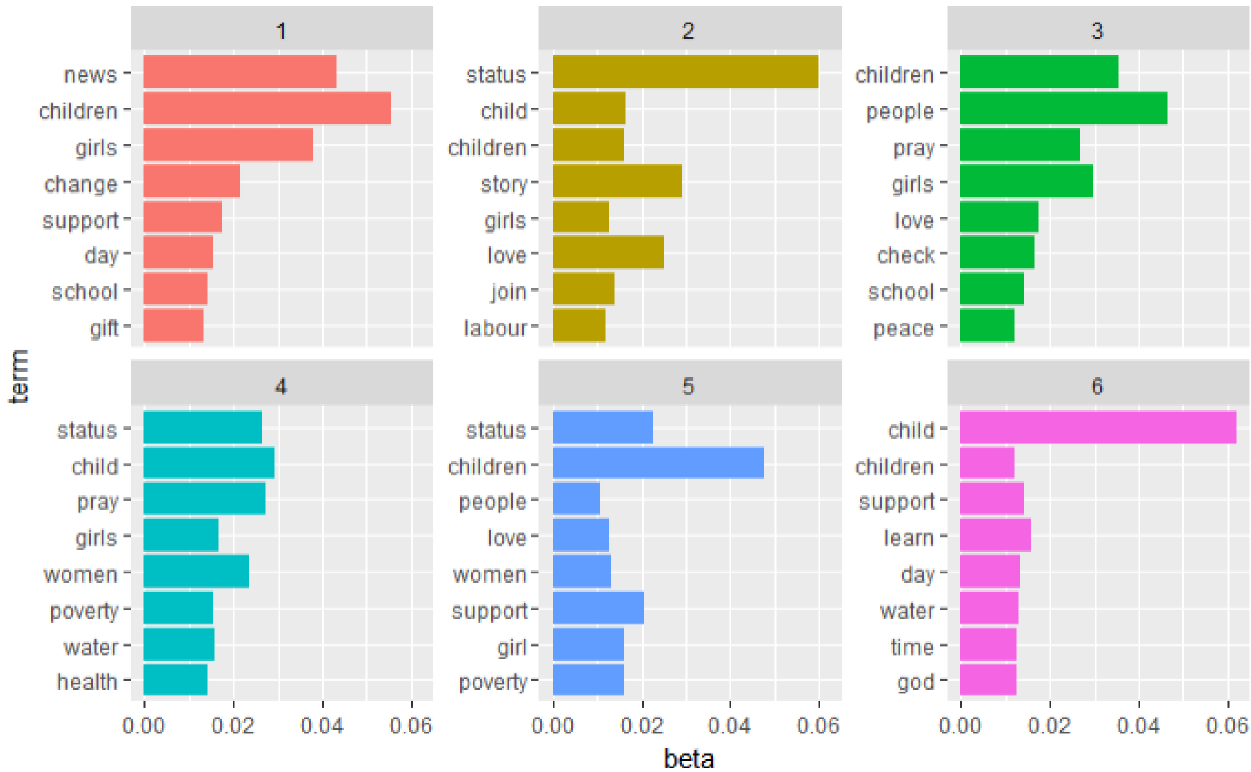
The same TF-IDF technique used above can be used with N-grams to find the most frequent and most unique 2 or 3 word combinations for each charity. In this dataset, since the most frequent words per charity are mostly unique to those charities, there is only a small difference between the results. For example, for trigrams:



Topic Modeling and Engagement Metrics

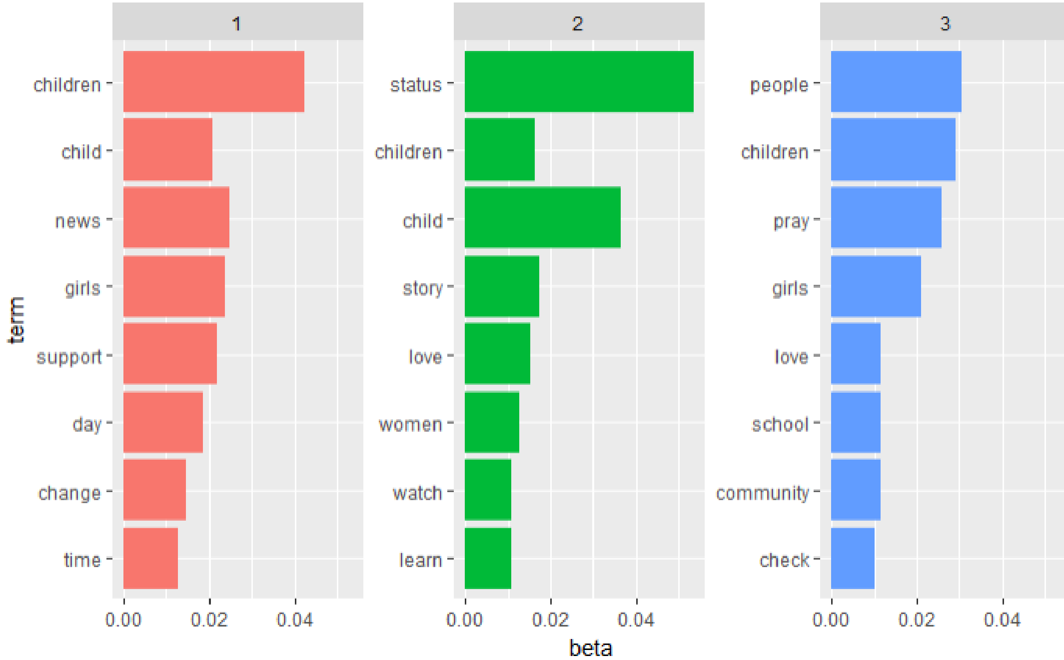
Certain topics also start to emerge from the data when looking at the frequency and association of words. Using the Latent Dirichlet Allocation (LDA) technique, different ‘clusters’ or groups of topics can be investigated.

Specifying 6 groups, the analysis found the following:



Many of the words in each group are similar but give a slightly different emphasis. The last (or second-last) word in each group provides insight that there may be groups of tweets that focus on the ‘health status of girls and women in povery’ (group 4) vs. ‘praying for children and girls who are in school’.

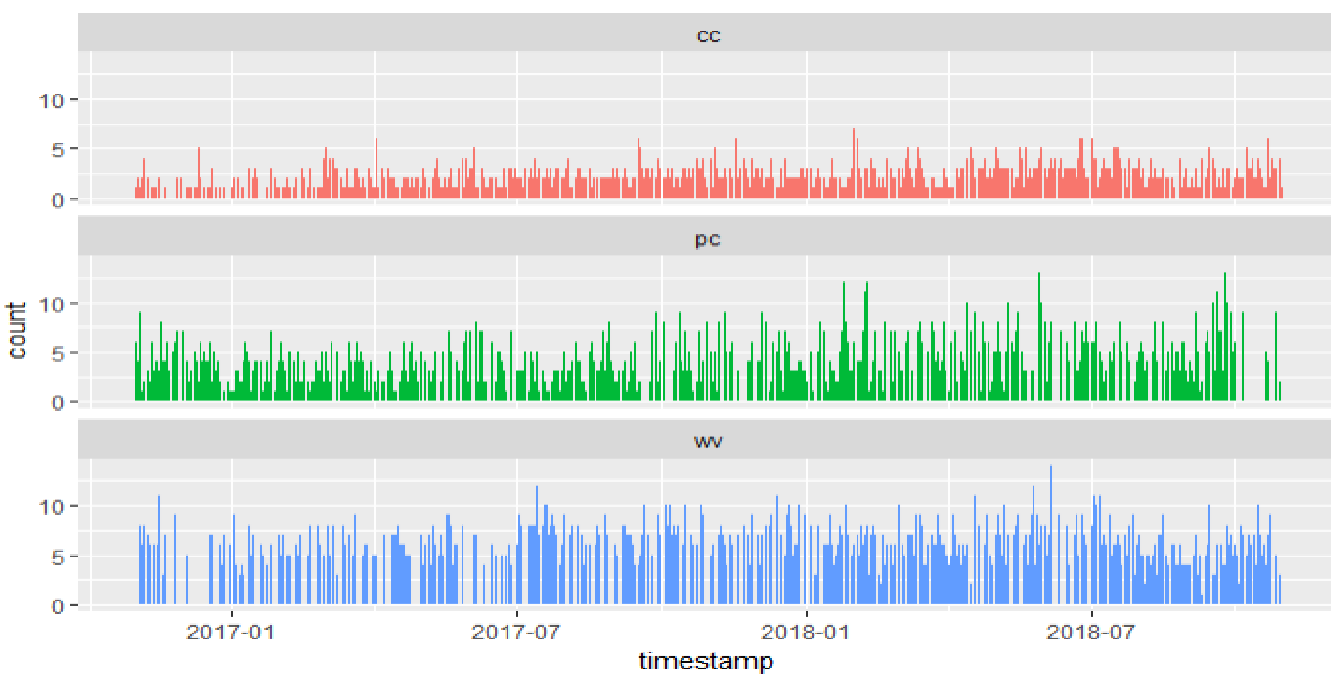
Specifying 3 groups, the analysis found the following:



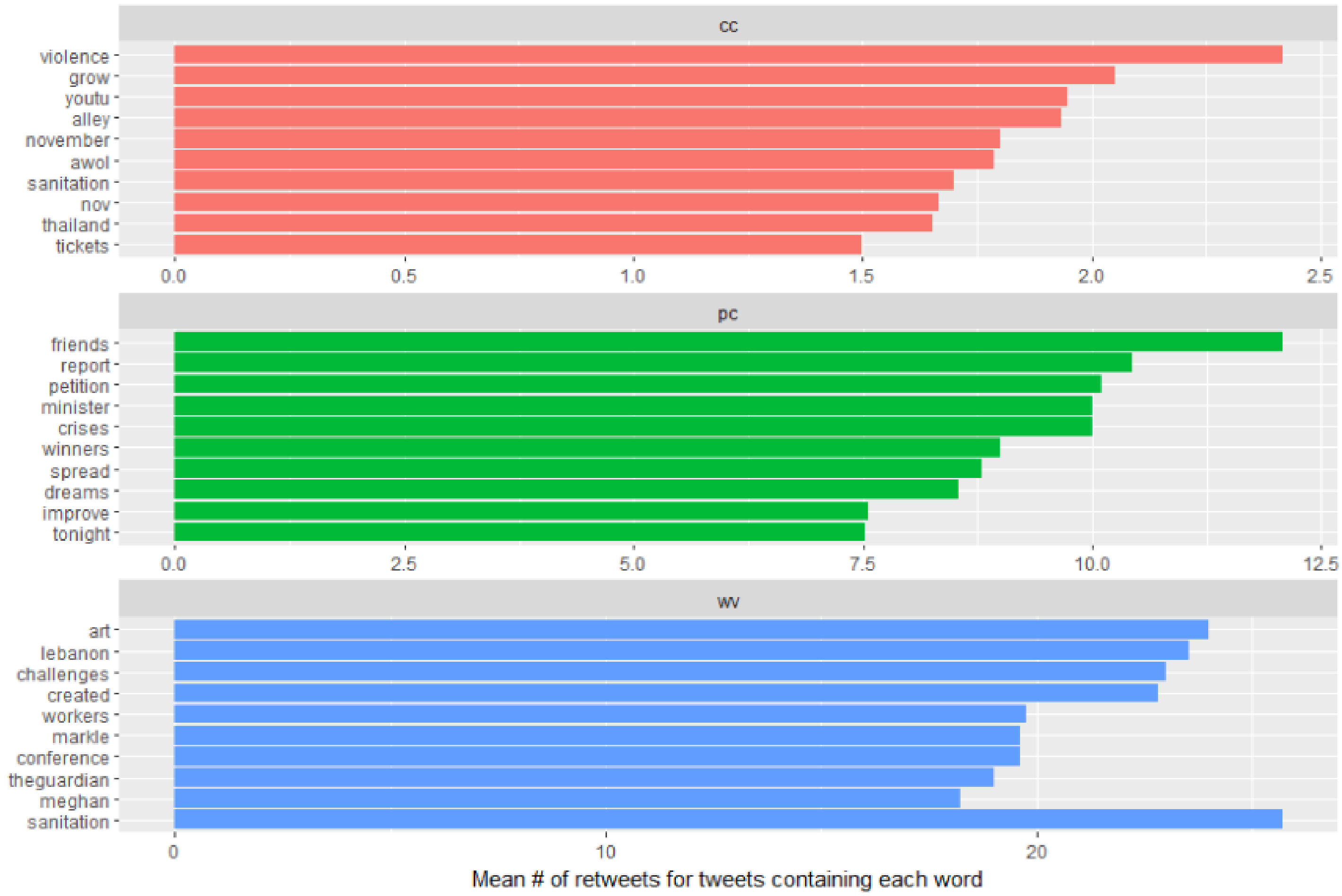
If the three charities were starkly different in their associated topics, this analysis would find 3 distinct groups. However, the words in these groups are still quite similar – suggesting that the topics these charities discussed were relatively similar.

The tweet data also contains numeric data quantifying engagement with the tweets: likes, replies, and retweets. Likes are the most common form of engagement, followed by retweets, with very few replies in the dataset overall.

The frequency of likes per day over time is as follows:

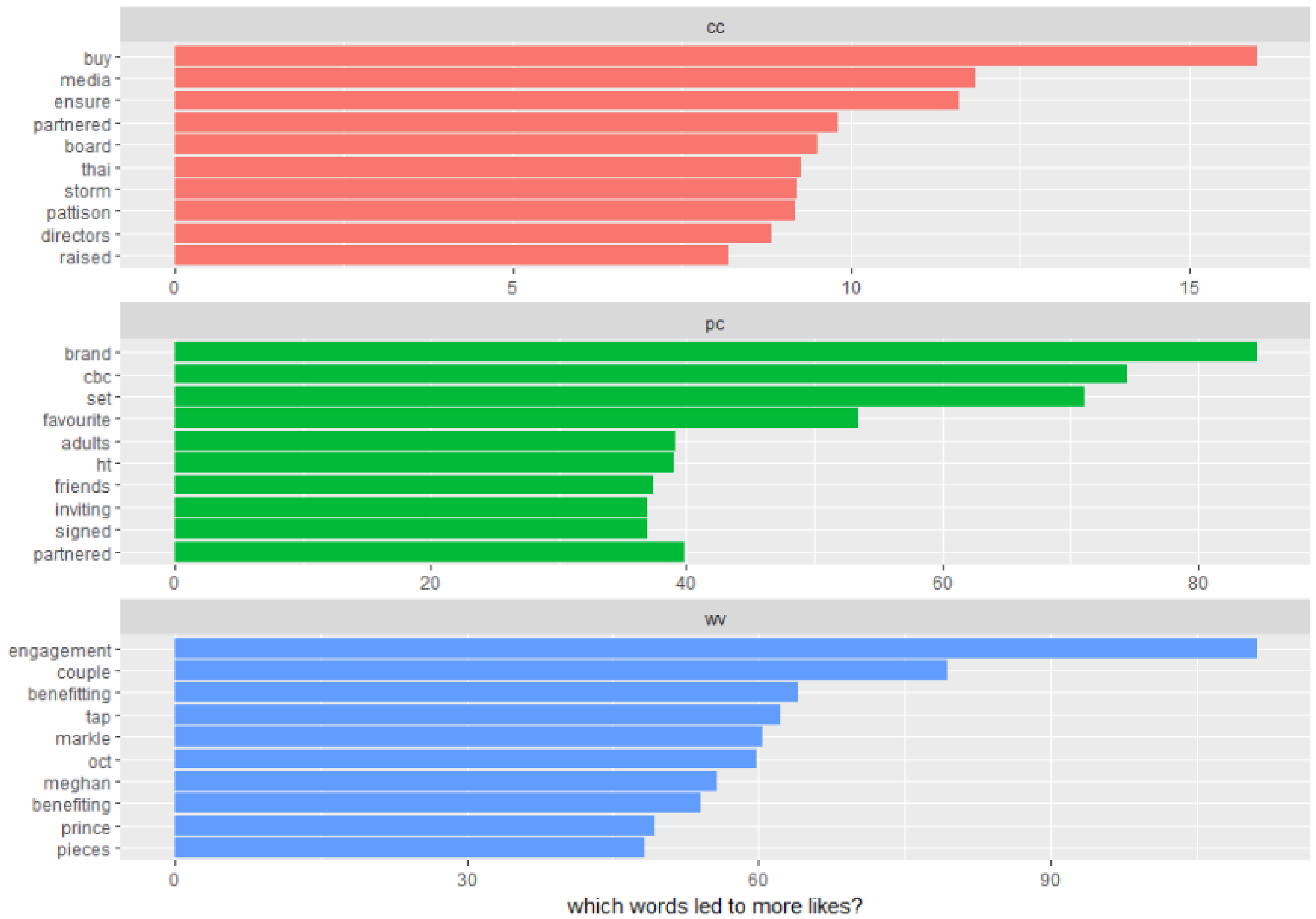


Starting with retweets, this analysis shows a comparison in engagement metrics:



The words associated with retweets with the highest means include:
Compassion Canada: 'violence'; 'grow'; 'youtu' (indicating a youtube link); and 'november'
Plan Canada: 'friends'; 'report'; 'petition'; and 'minister'
World Vision: 'sanitation'; 'art'; 'lebanon'; and 'challenges'

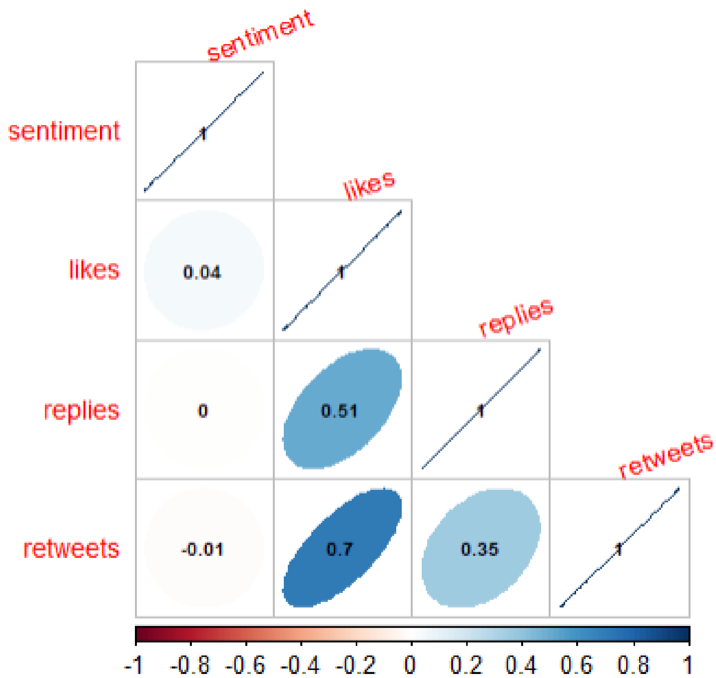
Extrapolating from these themes, it seems that tweets received more retweets when there was a direct element of engagement (a youtube link), reports on engagement (petitions, winners), mentions of specific places (Thailand, Lebanon), or 'sanitation' in particular. A few of World Visions words are related to a popular news event – the marriage of Meghan Markle (a 'World Vision ambassador') to Prince Harry of England.



The words associated with tweets with the highest mean likes include:
Compassion Canada: ‘buy’; ‘media’; ‘partnered’; and ‘board’
Plan Canada: ‘brand’; ‘cbc’; ‘favourite’; and ‘adults’
World Vision: ‘engagement’; ‘couple’; ‘benefitting’; and ‘tap’

Extrapolating from these themes, it seems that tweets received more likes in similar circumstances: engagement (signing petetions, partnering), or when there’s some kind of known-person endorsement: Rick Mercer from CBC (Plan), perhaps announcing the appointment of someone popular to Compassion’s Board, or the engagement of Meghan Markle to Prince Harry (World Vision).

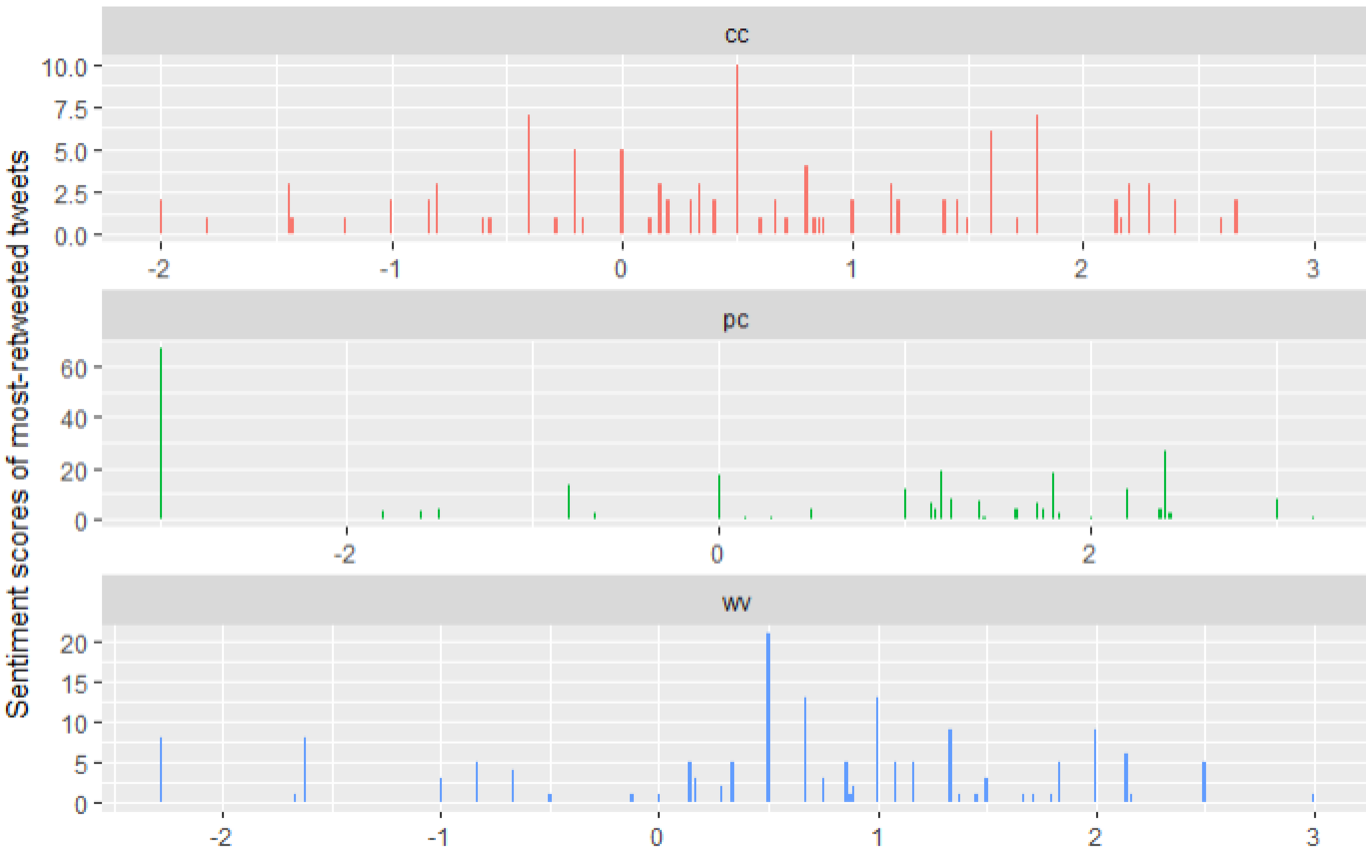
Looking at the sentiment score of tweets, any obvious relationship between sentiment and engagement may be found statistically in a correlation plot:



However, the only significant correlations are between retweets and likes, and between replies and likes – not between sentiment and any engagement metrics.

The sentiment scores of the top most-liked tweets varied between very negative to very positive:

The top most retweeted tweets are correlated to likes, but show a slightly different spread in sentiment:



The NRC sentiment library provides 10 sentiment words total:

```
[1] "trust"        "fear"         "negative"     "sadness"      "anger"        "surprise"
 [7] "positive"     "disgust"      "joy"          "anticipation"
```

These can be mapped to the tweet dataset, to show the relationship between engagement and words associated with these sentiments.

Looking at the most-liked tweets, most contain words associated with ‘trust’, ‘positive’, and ‘joy’:

	tag	user	likes	trust.present	fear.present	negative.present	sadness.present	anger.present	surprise.present
1	wv	worldvisioncan	540	1	0	1	1	1	0
2	pc	rickmercer	421	0	1	1	0	1	0
3	wv	worldvisioncan	280	0	0	0	0	0	0
4	wv	Vandiekins22	247	0	0	0	0	0	0
5	pc	RonaAmbrose	216	1	0	0	0	0	1
6	wv	worldvisioncan	192	0	1	0	0	0	1
7	pc	BrentToderian	189	1	0	0	0	0	0
8	pc	coollike	175	1	1	1	1	0	0
9	pc	sarahgrafferty	156	1	0	0	0	0	1
10	wv	worldvisioncan	154	1	0	0	1	0	1

…cont.

positive.present	disgust.present	joy.present	anticipation.present
1	0	1	1
1	0	1	0
0	0	0	0
1	0	1	0
1	0	1	1
1	0	1	0
1	0	0	1
1	1	0	0
1	0	1	1
1	0	1	1

The content of the top 3 tweets are:

Heartfelt congratulations to Meghan Markle, our ambassador for the past 2 years, on her engagement to Prince Harry. We're grateful for her support of the worldâ€™s most vulnerable children. We wish the couple every happiness together.

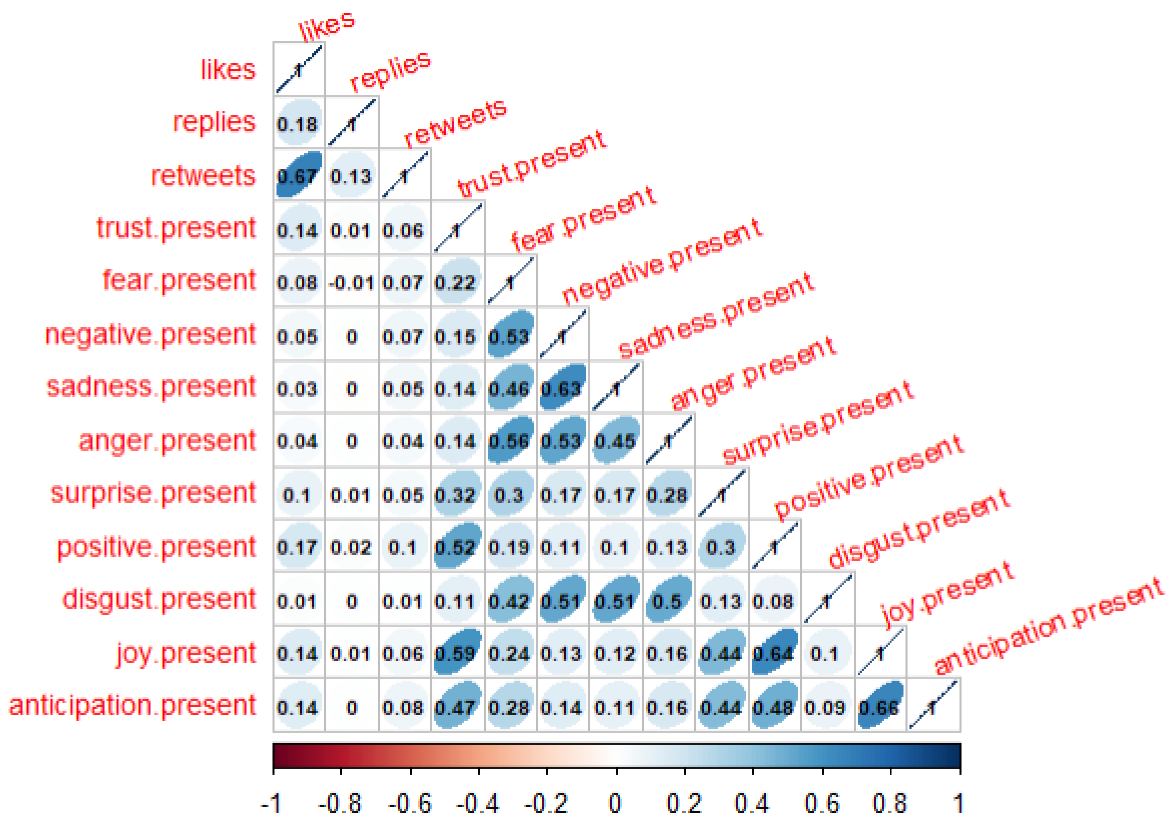
Tonight on a brand new Mercer Report â€“ always my favourite show of the year. We profile & salute the winners of this yearâ€™s Spread the Net Student Challenge. The feel good episode of the decade! 8:00 (8:30 NL) on CBC. Set your PVR Retweet & tell two friends. #PlanCanada

Such neat innovations in refrigeration and sanitation #socent #socinnhttps://www.theguardian.com/global-development/2016/oct/24/a-water-chilled-coolbox-gets-vaccines-on-tap-to-the-worlds-poorest-grand-challenges-conference-london

Of most-retweeted tweets, the majority contain words associated with ‘positive’, but also some with ‘fear’ and ‘sadness’.

Of most-replied to tweets, the majority contain words associated with ‘positive’ sentiment.

The correlation of sentiment words to engagement can be visualized in a correlation plot:



Here we see that the presence of sentiment-related words are not correlated to engagement metrics in any significant way.

Conclusions

This analysis used twitter data to investigate themes, sentiment, and engagement metrics as related to three Canadian charities (Compassion Canada, Plan Canada, and World Vision Canada). The purpose was to better understand how these charities were communicating their work, and to investigate their use of Twitter to engage existing and new groups of donors.

While the data showed much similarity in the content and sentiment associated with all three charities, there was still evidence of unique topics that emerged – suggesting they are increasingly leveraging a unique 'brand' to differentiate themselves. Compassion Canada used overtly religious language to engage its primarily Christian donor base, and emphasized its child-sponsorship program. Compassion was the least active and received the least amount of engagement (likes, retweets). They spoke very frequently about their CEO – perhaps using the Twitter platform as a medium to update donors on its organizational news. This suggest that higher activity and perhaps more cause-related messages might increase their engagement. Plan Canada has a well know brand associated with advocating for the rights of girls, and reproductive health – which was certainly evident in the analysis. Plan also used the Twitter platform to post about job openings – which probably helped them find candidates, but also reduces the engagement with twitter users who are not looking for a job at Plan. They were however associated with a few celebrities – CBC's Rick Mercer in particular; which led to their highest-engagement. In comparison, World Vision Canada had an emphasis on children as well – but also focused on refugees and specific countries where emergency situations are happening. World Vision's high level of activity and association with known celebrities supported the highest level of engagement among the three charities.

The sentiment of tweets didn't have a strong correlation with engagement metrics; though the most-liked tweets were mostly positive and contained words associated with 'trust', 'positive', and 'joy' sentiments. Some of the most retweeted tweets also contained words associated with 'fear' and 'sadness'. This suggests that the subject of tweets and the person they are associated with is more likely to drive engagement than sentiment alone.

While this analysis is informative, it is limited in scope. Single-word and even bi-gram and tri-gram associations don't consider the entire context and contents of each tweet. Further analysis will help to illuminate more actionable insights for Canadian charities to engage with a younger audience in new ways, and suggested next-steps are as follows:

- Train a classifier to more accurately predict sentiment of tweets – and/or use deep learning (a recurrent neural network) to consider entire context and the order of words in each tweet
- Incorporate more data into the analysis, including location of posts (geographic differences) and other social media data (from Facebook, Snapchat, etc.)

The ultimate question for these charities is this: how does engagement on social media drive revenue for their programs. To answer this difficult question of attribution, it is recommended these charities begin or continue work to incorporate specific digital, campaign, and financial data into future analyses.