# What are people talking about real wars and artificial wars?

## SENTIMENT ANALYSIS – SYRIAN WAR VS AVENGERS INFINITY WAR

**By**

**Nisha Muthukumaran**

### Introduction

This project touches upon various aspects of sentiment analysis and ways to visualize them and understand them. These concepts are applied on live tweets that were extracted in the month of April, 2018. The topic chosen is warfare, real and imaginary. The civil war currently happening in Syria is a topic of discussion for every news channel and so are entertainment news about movies like the Avengers Movie. The #SyrianWar captures sentiments of people across the globe and emotions that people are going through. The #AvengersInfinityWar is a more light-hearted but much awaited topic of discussion amongst movie fanatics across the world. The analysis is done over the following steps:

Step 1: Extraction of 1000 tweets from each "hashtag" which totals up to 2000 live tweets.

Step 2: Data Cleaning

Step 3: Building unigrams, bigrams and trigrams word clouds from the top 200 keywords caught from the tweets

Step 4: Building a TF-IDF word cloud from the TF-IDF matrix

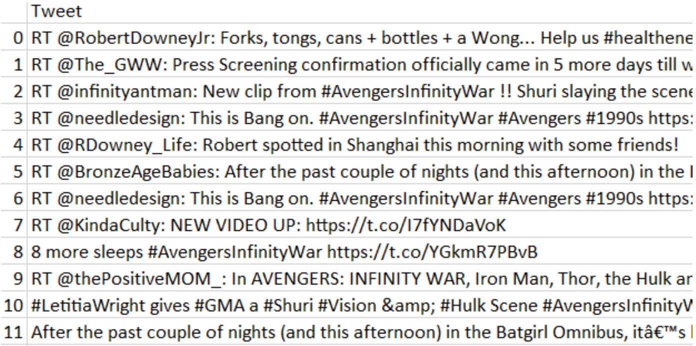
Step 5: Measuring the path of the sentiments across these tweets with the help of lexicons

Step 6: Building a commonality and comparison word cloud from the two topics

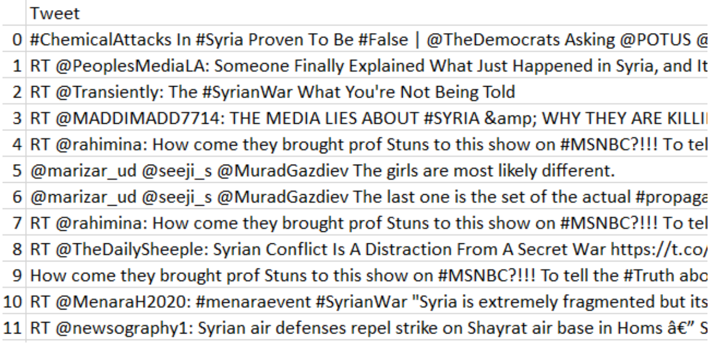
Step 7: Creating an NRC Radar chart to measure the emotions across the tweets

### Step 1: Tweets extraction

Twitter has a developer API platform which enables the common man to extract any publicly visible tweet from the past week from hashtags or '@' publicly visible accounts. It provides 4 unique keys which are required to extract data namely: consumer key, consumer secret, access token key, access token secret.  Using the **tweepy** library exclusive to Python, the tweets were extracted. These tweets are extracted to text files and CSVs and fed as inputs to the next steps and then are merged into 1 whole file.



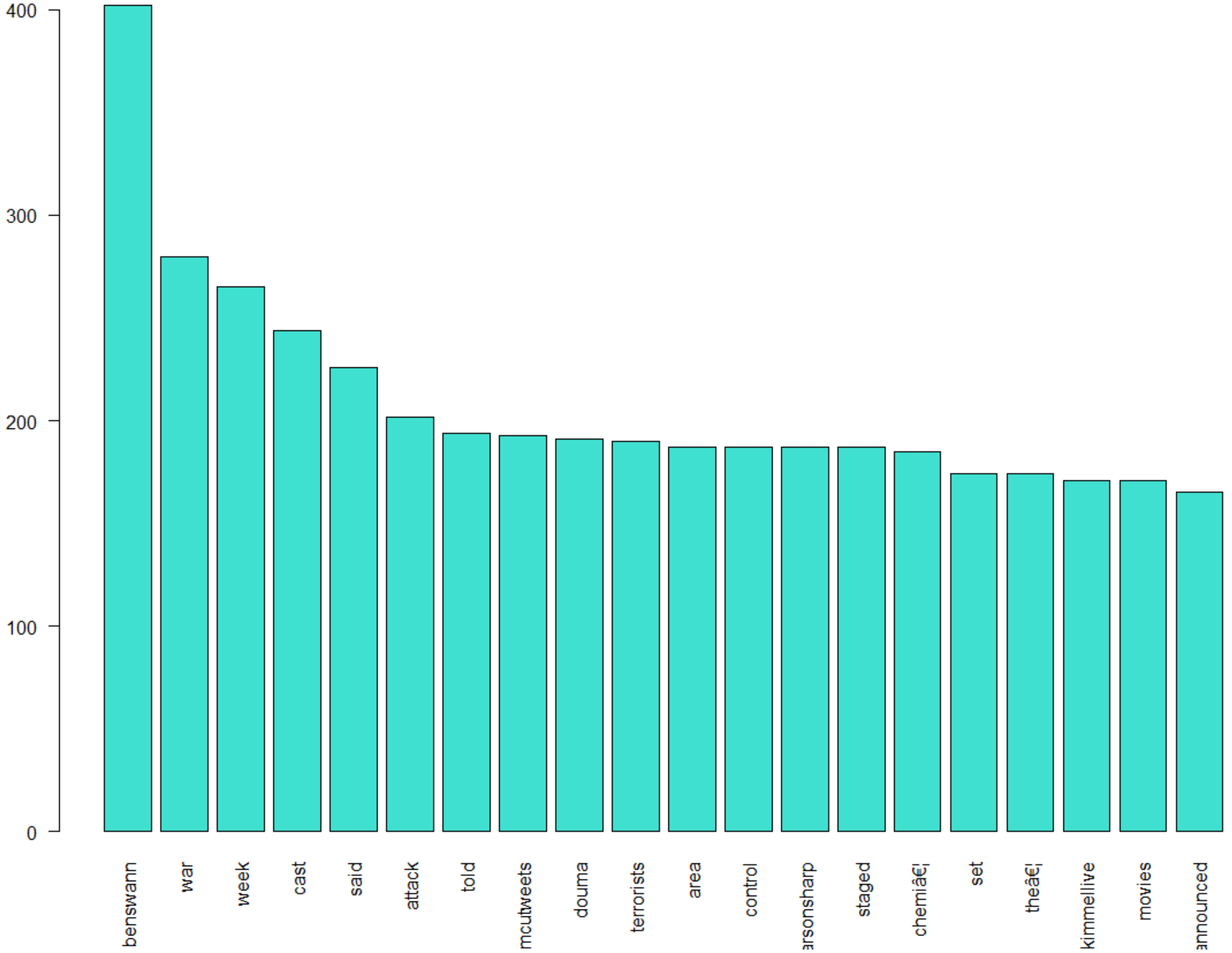
| | Tweet |
|---|---|
| 0 | RT @RobertDowneyJr: Forks, tongs, cans + bottles + a Wong... Help us #healthene |
| 1 | RT @The_GWW: Press Screening confirmation officially came in 5 more days till w |
| 2 | RT @infinityantman: New clip from #AvengersInfinityWar !! Shuri slaying the scene |
| 3 | RT @needledesign: This is Bang on. #AvengersInfinityWar #Avengers #1990s https: |
| 4 | RT @RDowney_Life: Robert spotted in Shanghai this morning with some friends! |
| 5 | RT @BronzeAgeBabies: After the past couple of nights (and this afternoon) in the |
| 6 | RT @needledesign: This is Bang on. #AvengersInfinityWar #Avengers #1990s https: |
| 7 | RT @KindaCulty: NEW VIDEO UP: https://t.co/I7fYNDaVoK |
| 8 | 8 more sleeps #AvengersInfinityWar https://t.co/YGkmR7PBvB |
| 9 | RT @thePositiveMOM_: In AVENGERS: INFINITY WAR, Iron Man, Thor, the Hulk ar |
| 10 | #LetitiaWright gives #GMA a #Shuri #Vision &amp; #Hulk Scene #AvengersInfinityW |
| 11 | After the past couple of nights (and this afternoon) in the Batgirl Omnibus, itâ€™s |



| | Tweet |
|---|---|
| 0 | #ChemicalAttacks In #Syria Proven To Be #False \| @TheDemocrats Asking @POTUS @ |
| 1 | RT @PeoplesMediaLA: Someone Finally Explained What Just Happened in Syria, and It |
| 2 | RT @Transiently: The #SyrianWar What You're Not Being Told |
| 3 | RT @MADDIMADD7714: THE MEDIA LIES ABOUT #SYRIA &amp; WHY THEY ARE KILLI |
| 4 | RT @rahimina: How come they brought prof Stuns to this show on #MSNBC?!!! To tel |
| 5 | @marizar_ud @seeji_s @MuradGazdiev The girls are most likely different. |
| 6 | @marizar_ud @seeji_s @MuradGazdiev The last one is the set of the actual #propaga |
| 7 | RT @rahimina: How come they brought prof Stuns to this show on #MSNBC?!!! To tel |
| 8 | RT @TheDailySheeple: Syrian Conflict Is A Distraction From A Secret War https://t.co/ |
| 9 | How come they brought prof Stuns to this show on #MSNBC?!!! To tell the #Truth abo |
| 10 | RT @MenaraH2020: #menaraevent #SyrianWar "Syria is extremely fragmented but its |
| 11 | RT @newsography1: Syrian air defenses repel strike on Shayrat air base in Homs â€" S |

### Step 2: Data Cleaning

A vector corpus is made from these 2000 tweets. With the help of the **tm** library in R the corpus is cleaned. The tweets mostly end in URLs and this is the first thing that is removed. With the help of regular expressions, tweets that have links beginning with http are removed. Next all the letters are converted to lower case, this helps is removing redundant words in the latter steps, whether they occur in the beginning of sentences or in the middle. Next the punctuations across the tweets are removed. Numbers are removed in the following steps. There are a set of predefined stop-words in the English language and these are removed. Custom stops words related to the topics can be removed so that they do not get accounted for in the Term Document Matrix. Finally, all the white spaces are removed. The graph below shows the top twenty words frequency occurring across the cleaned corpus for all 2000 tweets.



Ben Swann is a tv news anchor for a news channel and is claimed to talk about conspiracy theories and is said to play a role in the Russian air strike on Syria (Source: Wikipedia) which is probably why people are talking about him. War is a term that is most commonly used. MCU tweets is a twitter handle that talks about the Marvel Comic Universe. This graph shows that both topics are covered in terms of word frequencies.

This data is now fit to be converted into a term document matrix. TDM contains columns and rows. Column headers are the unique words found in every tweet across all 2000 tweets. And the row numbers are every document in tweets. The presence of words across every document is recorded in a TDM. The disadvantage of a TDM is that it is very sparse.

**Step 3: Unigram, Bi-gram and Tri-gram word clouds**

Machines do not perceive the English language (or any other language) like humans do. It breaks it down to understanding the meaning of the sentence.

Example: “This is an interesting project”

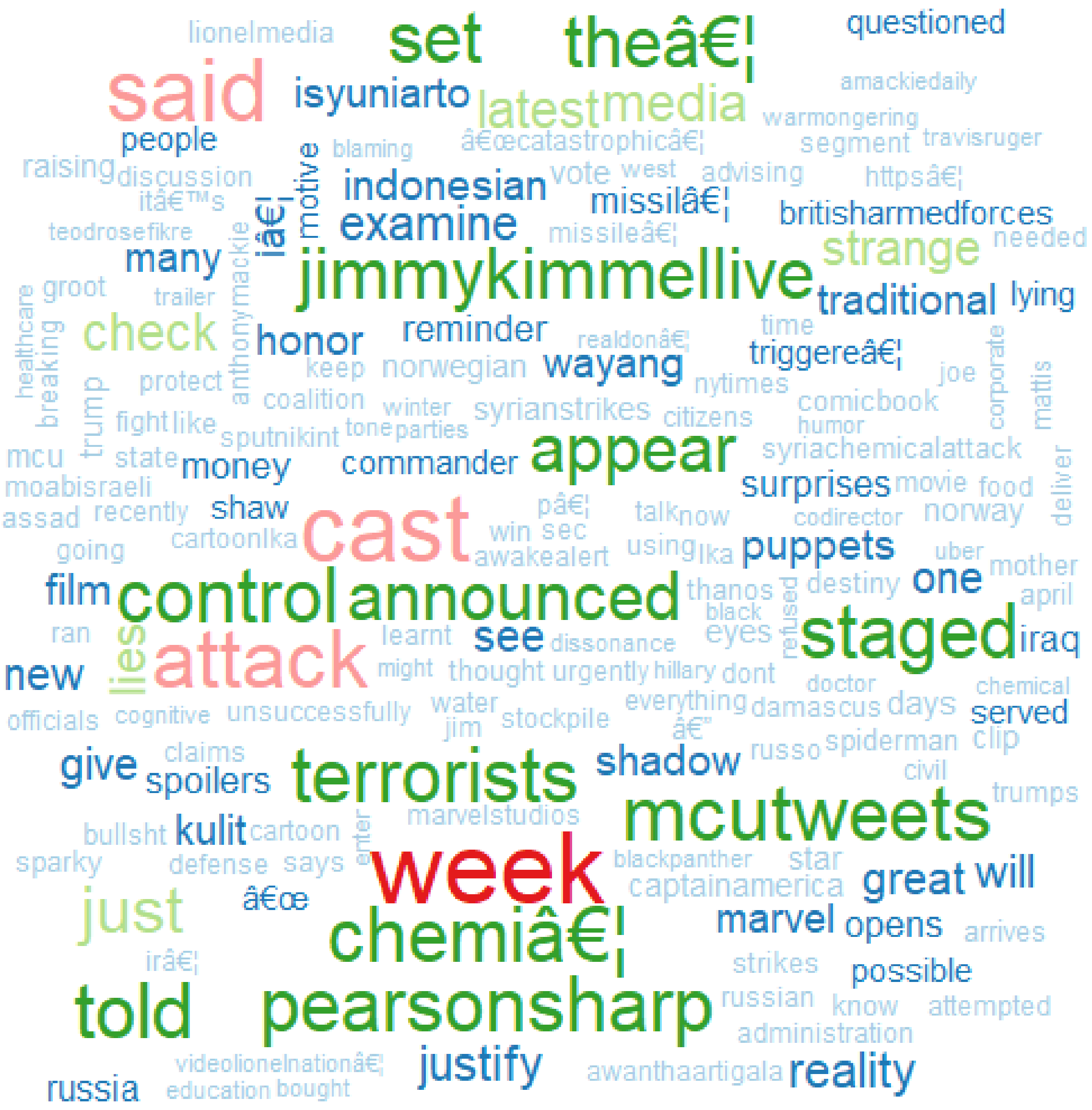
As a unigram each word is a gram and is read as: “This”, “is”, “an”, “interesting”, “project”.

As a bigram it read two adjacent words at a time and this is read as: “This is”, “is an”, “an interesting”, “interesting project”.

As a trigram it reads three adjacent words at a time and reads this as: “This is an”, “is an interesting”, “an interesting project”.

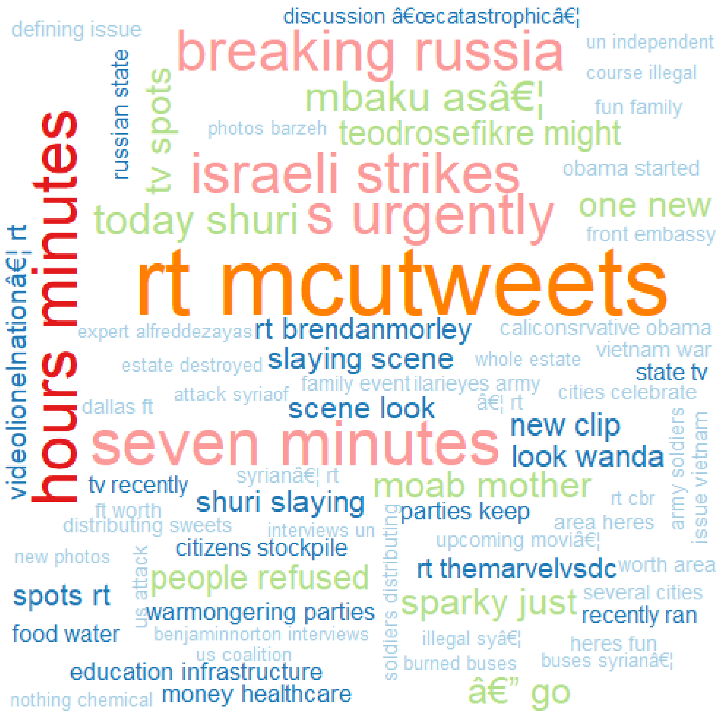
**Unigram word cloud:**

Based on the frequency of the word, the size is determined. The colour palette is defined to be a gradient of a few colours. This is composed of the top 200 unigrams with the minimum frequency set to 10 which means that it accounts for words that have occurred for 10 times or more.



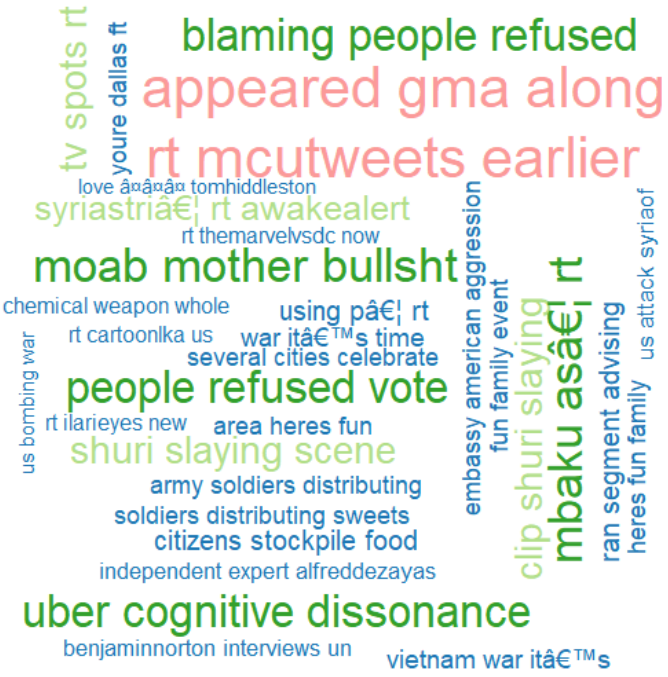
**Bi-gram word cloud:**

This is made from the top 200 bigrams and the size is defined by the frequency. Larger the size, more often it was used. The limitation of the R output cuts off the visual by reducing the number of combinations showed as it did not fit on the R output console. Since most of the retweets are from the Marvel Comic Universe handle, that is why it has been highlighted here. Time measurements spoken here are about the movie run time and how long the war has been going on for. This cloud considers word pairs that occur most commonly across the tweets.



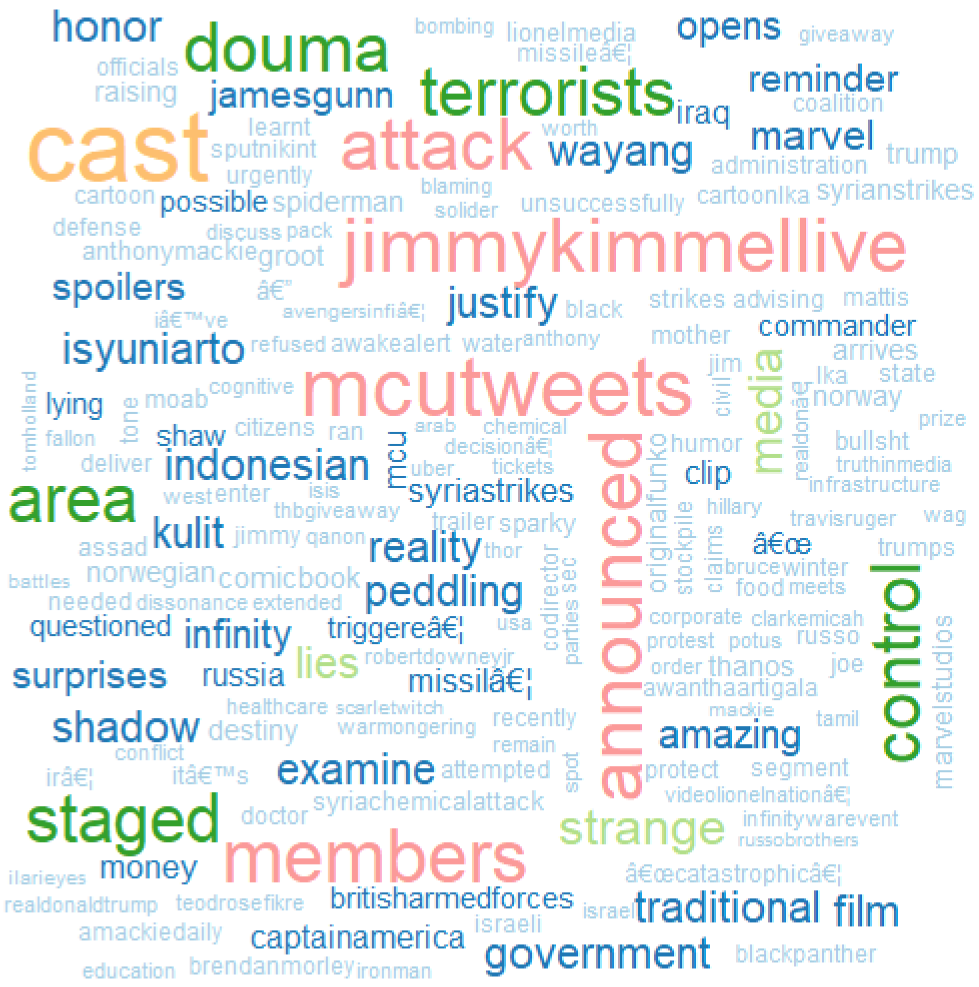
**Tri-gram word cloud:**

This is made from the top 200 tri grams and the size is defined by the frequency. Larger the size, more often it was used. The limitation of the R output cuts off the visual by reducing the number of combinations showed as it did not fit on the R output console.



**Step 4: TF-IDF word cloud**

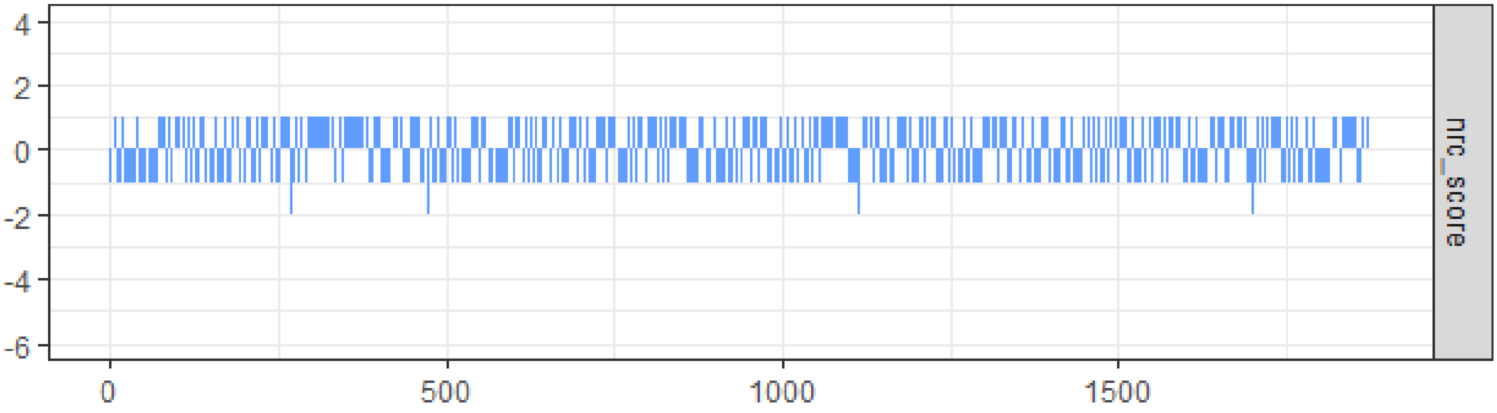
Since the term document matrix is very sparse, it is converted to a term frequency – inverse document frequency matrix which shows how important a word is to a document in a collection or corpus. This is a weighting method where each word's contribution or relatability to a document is given as a score. This helps in identifying how important a word is to a specific document. A word cloud is made from the top 200-word frequencies in a TF-IDF but due to space constraints not all the 200 appear and a lot of it get cut-off.



**Step 5: Sentiments using NRC Lexicon**

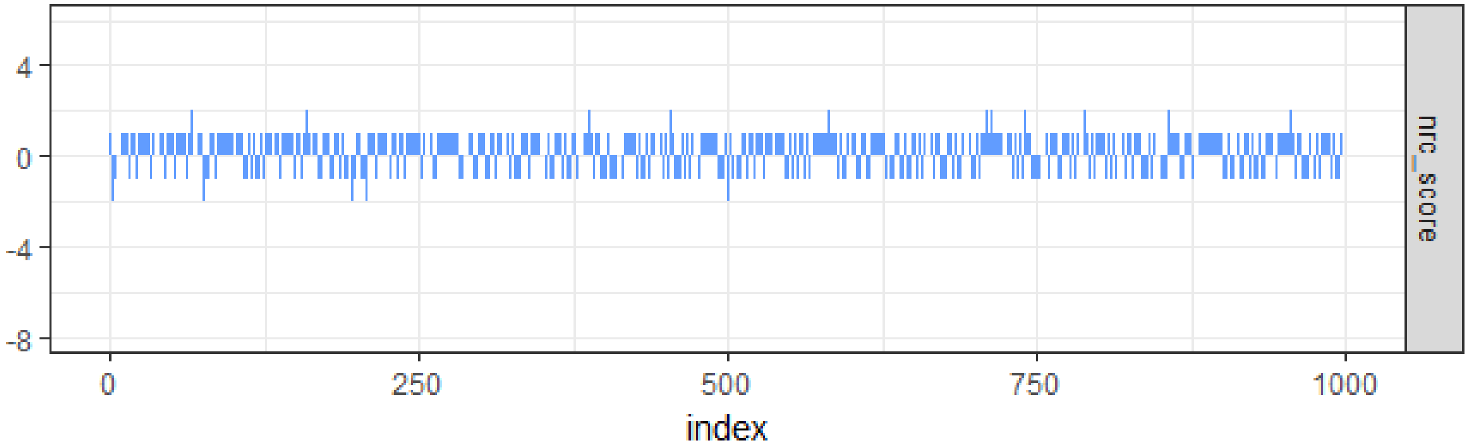
Lexicons are based on unigrams. The NRC lexicon categorizes words in a binary fashion (“yes”/ “no”) into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The BING lexicon categorizes words in a binary fashion into positive and negative categories from the start. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. This is exercise is done using the NRC lexicon.

**FOR SYRIAN WAR**



Although the NRC lexicon has 8 emotions, the concern here is only for positive and negative emotions for plotting purpose. So just the negative emotion is kept as negative and everything else is converted to positive. These scores are then aggregated. When the scores are aggregated the strength of the positive and negative emotions add up and a clear distinction in emotions throughout the tweets can be identified. It is visible that there are negative peaks in the Syrian was NRC lexicon more than the positive and vice versa for the Avengers infinity war. In fact, for the Avengers infinity was lexicon you can see more positive peaks than negative peaks.

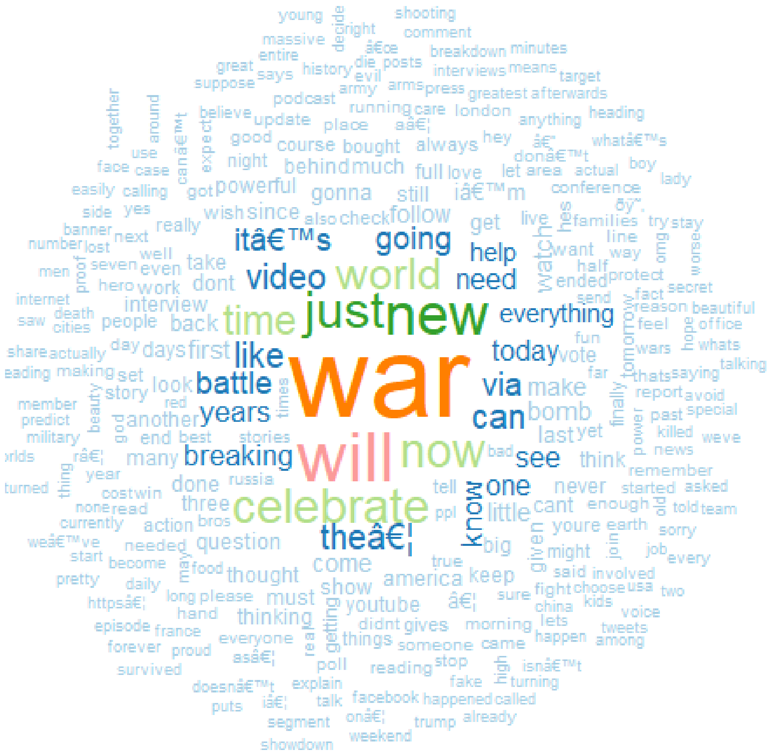
**FOR AVENGERS INFINITY WAR**



**Step 6: Commonality and comparison word cloud**

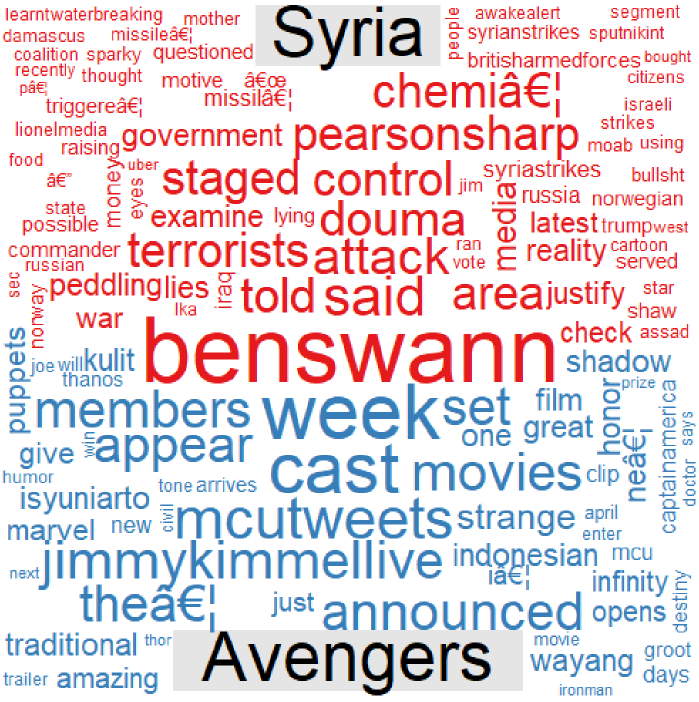
The commonality word cloud shows what are the common keywords talked amongst both topics. Since both are wars, that's the one that was found occurring. It can be understood that most of the topics talked here can be involving the world since that is being highlighted too.

COMMONALITY WORD CLOUD

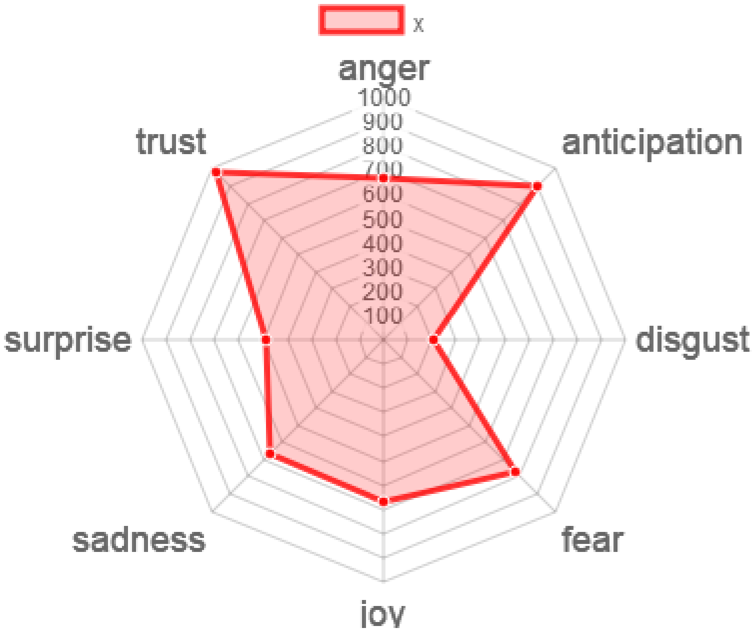


COMPARISON WORD CLOUD

This talks about what is unique across the real war and the movie. It's very interesting to see the highlights for each topic. And without having to go through 2000 tweets, the topics of importance can be understood with this. This was created with the top 200 words only as R cuts off when the size increases.



**7: NRC Radar chart**



Since NRC is spread over 8 emotions, that can be noticed here as the edges of the octagon. This chart talks about the general emotions experienced over all the 2000 tweets. Sadness, fear and anger are the negative emotions that are mostly described because of the Syrian war. And anticipation is high probably for the movie as it has a huge fan base and every Marvel fan will be looking out for this movie as it combines Guardians of the Galaxy and the Avengers including Doctor Strange and Black Panther. Joy as an emotion which might also be because of the movie.