Sentiment analysis and Text classification (Deep learning) of climate research studies

By
Yemi O ~ Michael P~Ryan P

Background: Socioeconomic impact of climate change

- Public health
  - climate change is a significant threat to the health of the people
  - Climate change can affect human health in two main ways:
    - by changing the severity or frequency of health problems that are already affected by climate or weather factors;
    - by creating unprecedented or unanticipated health problems or health threats in places or times of the year where they have not previously occurred
- Agriculture
  - Threat to food security
  - Species extinction
- Nutrient cycle
  - Low crop yield (depletion of nitrate)
  - Air pollution/greenhouse effect
  - Increase in natural diseases such as floods, hurricanes, desert encroachment
  - Slow decomposition process
- Migration
  - Excessive migration due to unfavorable weather conditions and climatic factors

Concept: Sentiment analysis

- Sentiment = feelings
  - Attitudes
  - Emotions
  - Opinions
- Subjective impressions, not facts
- Generally, a binary opposition (polarity) in opinions is assumed
- Using NLP, statistics, or machine learning methods to extract, identify, characterize the sentiment content of a text unit
- Sometimes referred to as opinion mining,
- P : positive, N:Negative, O: other words

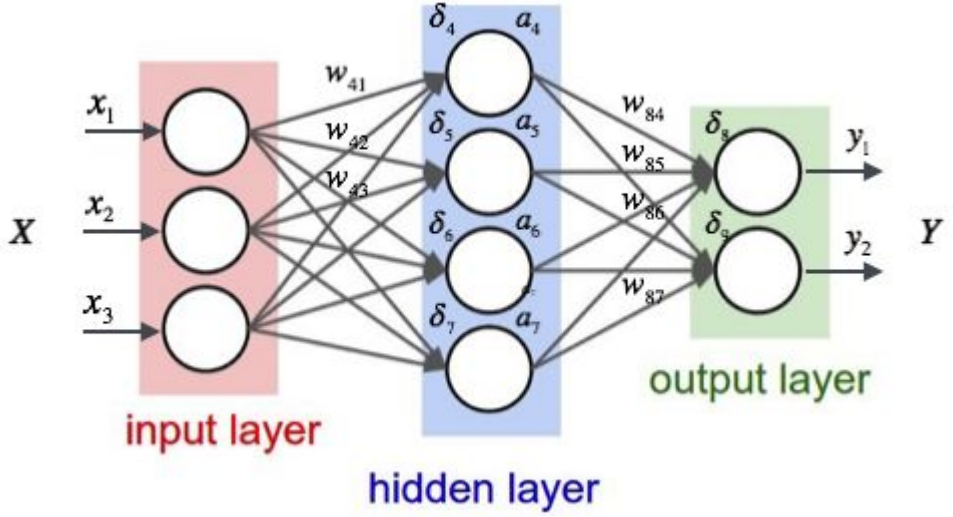
$$Sentiment = \frac{(P - N)}{(P + N + O)}$$

$$Sentiment = \frac{(P - N)}{(P + N)}$$

$$Sentiment = log(P+0.5) - log(N+0.5)$$

Concept: Deep neural network learning for text classification

- ➔ Text classification is an example of Machine Learning (ML) in the form of Natural Language Processing (NLP).
- ➔ The goal of text classification is to automatically classify the text documents into one or more predefined categories.
- ➔ Text Classification Using **Recurrent Neural Network (RNN) :**
  - ◆ A (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence.
  - ◆ This allows it to exhibit dynamic temporal behavior for a time sequence
- ➔ Word embedding - **word2vec** representation
  - ◆ Similarity metric : cosine similarity



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

About the data by the numbers

- Data source **National Science Fellowship**
- timeline: **1993 -2017**
- Year window **10**
  - **1993-2002**
  - **2003-2011**
  - **2012 -2017**
- Number of records **26424**
- Number of features/variables **7**
- Variable of interest **Abstract**

Methodology

Methodology : Preprocessing

- Abstracts were filtered using the climaterealityproject.org keywords and IPCC glossary of terms
- Terms such as **'climate','global warming','fossil','methane','Greenhouse','emission','carbon dioxide'** were used
- The result filtered records were bound in a dataframe
- Data was aggregated to remove duplicates that overlapped during the filtration

- Climate Reality Project

```
# https://www.climaterealityproject.org/blog/key-terms-you-need-understand-climate-change

climate_nsf<-filter(nsf,grepl(c('climate'),abstract))
global_warming_nsf <-filter(nsf,grepl(c('global warming'),abstract))
fossil_nsf <- filter(nsf,grepl(c('fossil'),abstract))
methane_nsf<-filter(nsf,grepl(c('methane'),abstract))
ppm_nsf <- filter(nsf,grepl(c('PPM'),abstract))

ipcc_nsf <-filter(nsf,grepl(c('IPCC'),abstract))
ocean_nsf <-filter(nsf,grepl(c('ocean'),abstract))
renewable_nsf <- filter(nsf,grepl(c('renewable'),abstract))
unfccc_nsf <-filter(nsf,grepl(c('UNFCCC'),abstract))
weather_nsf <-filter(nsf,grepl(c('weather'),abstract))
temperature_nsf <- filter(nsf,grepl(c('average temperature'),abstract))
```

Methodology : Preprocessing

10 year Window

Window one : 1993 -2002

Window two : 2003 -2012

Window three : 2013- 2017

```
# we want to look at 10 year window 1993 -2002, 2003 - 2012, 2013-2017
#1993 -2002
nsf9302<-nsf %>% filter(between(year,1993,2002))

#2003 -2012
nsf0312 <-nsf %>% filter(between(year,2003,2012))

#2013 - 2017
nsf1317 <-nsf %>% filter(between(year,2013,2017))
```

Methodology : general housekeeping

```
library(tm)


doc <- Corpus(VectorSource(health_impact_nsf_9302$abstract))
doc = tm_map(doc,tolower)
doc = tm_map(doc, removePunctuation)
doc = tm_map(doc, removeWords, stopwords("english"))
doc= tm_map(doc,removeNumbers)
doc <- tm_map(doc, stripWhitespace)
```

Preliminary Results

Sentiment Analysis

Methodology: Preprocessing

- Data was filtered based on the terms affiliated with socio-economic impact of climate change per time window
- For example
  - Public Health
    - Public health,diseases
  - Agriculture
    - Animal husbandry, crops, hunger
  - Nutrient cycle
    - Methane,CO2,Emission

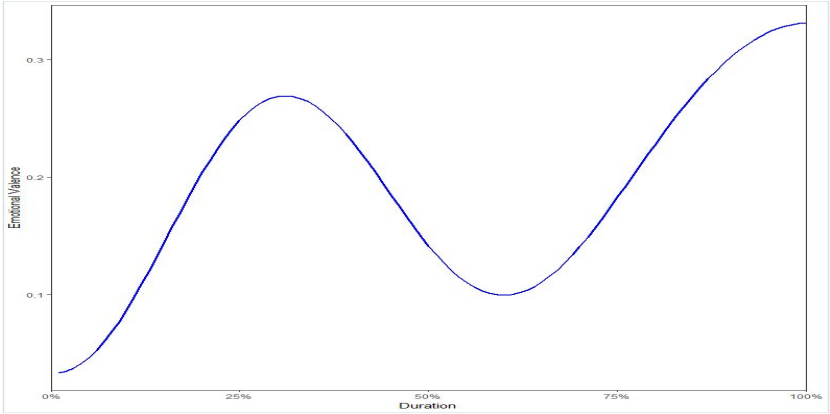
```
###################key impact public health,

health_impact_nsf_9302<-filter(nsf9302,grepl(c('public health','diseases'),abstract))
```

Result: Impact of Climate Change on Public health

- There are 11 article/abstracts with abstracts 3 and 8 indicating a **negative** polarity based on frequency of words such as fire, malaria, diseases

The curve shows the emotional valence modulation between polarity

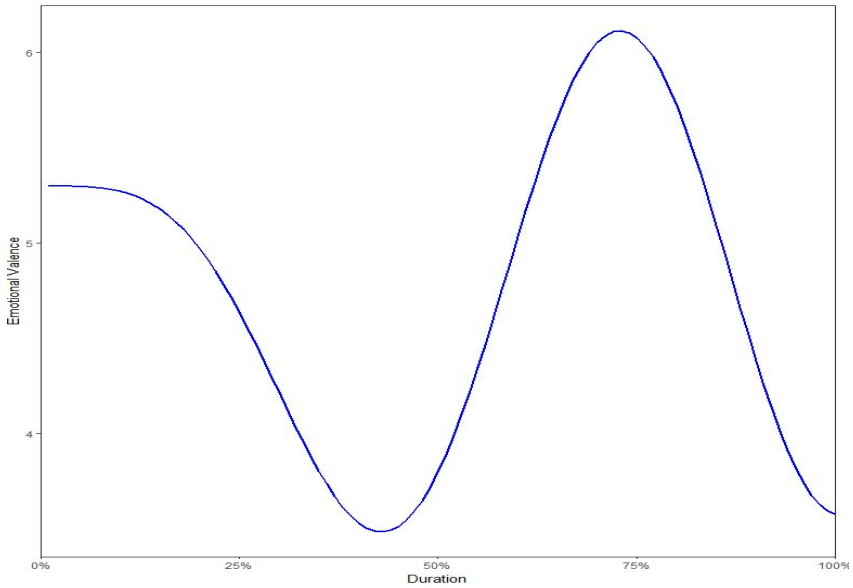
```
> as.data.frame(sentiment_by(health_impact_nsf_9302$abstract))
   element_id word_count        sd ave_sentiment
1           1        339 0.1959596    0.09745500
2           2        418 0.2816347    0.12431274
3           3        568 0.4032787   -0.02908096
4           4        630 0.3389112    0.22608584
5           5        519 0.1790369    0.08736418
6           6        255 0.2244205    0.19413235
7           7        631 0.1951083    0.05486518
8           8        260 0.3571415   -0.06595470
9           9        249 0.1902841    0.12306615
10         10        627 0.2628934    0.22179165
11         11        508 0.3326434    0.13301875
```



Result : Impact of climate change on agriculture

We can see that abstract 16 out of the first 20 agric affiliated abstracts had an average sentiment value of -0.003(0.295 stdev))

```
> as.data.frame(sentiment_by(agric_combined$abstract))
   element_id word_count        sd ave_sentiment
1           1        266 0.14996469   0.1756210677
2           2        390 0.19119993   0.0254537329
3           3        243 0.30367560   0.2360015382
4           4        215 0.18867804   0.2518463506
5           5        294 0.09969936   0.0876625927
6           6        255 0.14040486   0.0572297495
7           7        318 0.16229740   0.2352123846
8           8        246 0.26058789   0.1511790285
9           9        285 0.15411083   0.1643704116
10         10        311 0.20544707   0.1534377520
11         11        295 0.23942468   0.1850856796
12         12        466 0.36051443   0.2525667462
13         13        487 0.28817142   0.4702750797
14         14        346 0.19088951   0.3187317862
15         15        343 0.37844673   0.3113493694
16         16        158 0.29508540  -0.0003110204
17         17        162 0.28555285   0.1600145397
18         18        269 0.16488527   0.0658883354
19         19        269 0.16488527   0.0658883354
20         20        211 0.21473371   0.1850538430
```



Next steps……

- ➔ Finish the sentiment analysis of the socioeconomic factors
  - ◆ 10 year window per socioeconomic factors
- ➔ Text classification of the abstracts based on their polarity
  - ◆ Comparative study
    - **Support Vector Machine**
    - **Naive Bayes** ***** baseline algorithm
    - **Deep Neural Network (deep learning)**
  - ◆ Model evaluation using **Receiver operating characteristic -area under the curve AUC**