Final Project

Paloma Cartwright, Allie Cole, Wylie Hampson, Ben Moscona-Remnitz, Julia Parish

2022-06-01

Sentiment of Sustainable Agriculture over 30 year period

This text sentiment analysis was completed as an assignment for the course, Environmental Data Science 231: Text and Sentiment Analysis for Environmental Problems. The data was sourced from Nexis Uni database and the Food and Agriculture Organization (FAO) of the United Nations.

Original assignment instructions can be found here

Load Libraries

```r
#install packages as necessary, then load libraries
if (!require(librarian)){
  install.packages("librarian")
  library(librarian)
}

librarian::shelf(here,
                 igraph,
                 janitor,
                 kableExtra,
                 ldatuning,
                 LDAvis,
                 LexisNexisTools,
                 lubridate,
                 pdftools,
                 quanteda,
                 quanteda.textplots,
                 quanteda.textstats,
                 readr,
                 reshape2,
                 sentimentr,
                 tidyr,
                 tidytext,
                 tidyverse,
                 tm,
                 topicmodels,
                 tsne)
```

Nexis Uni Data - Sustainable Agriculture Sentiment past 30 years

Data consists of news articles from the Nexis Uni database, and published between the years 1990-2021 using the search terms “Integrated Pest Management” and “sustainable farming.” Due to Nexis Uni limiting article returns to 100 documents per search, we collected the 100 most relevant articles from each year. When a year returned less than 100 articles for the search terms, all returned articles were collected. Articles were not geographically restricted, but they were all published in English.

```
my_files <- list.files(pattern = ".docx", path = here("data", "text_data"),
                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)
dat <- lnt_read(my_files) #Object of class 'LNT output'
```

```
meta_df <- dat@meta
articles_df <- dat@articles
paragraphs_df <- dat@paragraphs
```

```
dat2<- data_frame(element_id = seq(1:length(meta_df$Headline)),
                  Date = meta_df$Date,
                  Headline = meta_df$Headline) %>%
  clean_names() %>%
  drop_na()

dat2$date[dat2$date == "0008-05-13"] <- "2008-05-13"

paragraphs_dat <- data_frame(element_id = paragraphs_df$Art_ID,
                             Text  = paragraphs_df$Paragraph)
```

```
# join the headlines with the paragraphs
dat3 <- inner_join(dat2, paragraphs_dat, by = "element_id") %>%
  clean_names() %>%
  drop_na()

dat3$date[dat3$date == "0008-05-13"] <- "2008-05-13"
```

```
dat3 <- subset(dat3, text != " ")
dat3 <- dat3[!duplicated(dat3$text),]
dat3 <- dat3[!grepl("http", dat3$text),]
dat3 <- dat3[!grepl("www.", dat3$text),]
dat3 <- dat3[!grepl("Article Rating", dat3$text),]
dat3 <- dat3[!grepl("ZRL: Zero Risk Level", dat3$text),]
dat3 <- dat3[!grepl("ZOI: Zone Of Incorporation", dat3$text),]
dat3 <- dat3[!grepl("Contact", dat3$text),]
dat3 <- dat3[!grepl("--", dat3$text),]
dat3 <- dat3[!grepl("By", dat3$text),]
dat3 <- dat3[!grepl("Language:", dat3$text),]
dat3 <- dat3[!grepl("Share on", dat3$text),]
dat3 <- dat3[!grepl("Mar ", dat3$text),]
dat3 <- dat3[!grepl("Apr ", dat3$text),]
dat3 <- dat3[!grepl("zones", dat3$text),]
dat3 <- dat3[!grepl("-", dat3$text),]
dat3 <- dat3[!grepl("WASHINGTON, D.C.", dat3$text),]
dat3 <- dat3[!grepl("No19. Pest Controllers", dat3$text),]
dat3 <- dat3[!grepl("YOUR TOWN", dat3$text),]
dat3 <- dat3[!grepl("NO HEADLINE", dat3$text),]

dat3 <- dat3 %>%
```

```r
  drop_na()
```

Calculate the Polarity of the Headlines

```r
mytext <- get_sentences(dat2$headline)
sent <- sentiment(mytext)

sent_df <- inner_join(dat2, sent, by = "element_id")

sentiment <- sentiment_by(sent_df$headline)

sent_df <- sent_df %>%
  arrange(sentiment)

sent_df$polarity <- ifelse(sent_df$sentiment < 0, -1, ifelse(sent_df$sentiment > 0, 1, 0))

sent_df <- sent_df %>%
  mutate(polarity = factor(polarity, levels = c("1", "0", "-1"))) %>%
  clean_names()

sent_df_group <- sent_df %>%
  mutate(year = lubridate::year(date)) %>%
  count(polarity, year) %>%
  drop_na()
```

Create a Plot of Headline Sentiment Polarity

```r
headline_sent <- ggplot(data = sent_df_group, aes(x = year, y = n, color = polarity)) +
  geom_line(size = 1.5) +
  labs(y = "Number of Headlines",
       x = "Year",
       color = "Sentiment",
       title = "Sustainable Farming News Headlines Sentiment Analysis",
       subtitle = "Integrated Pest Management & Sustainable Farming",
       caption = "Data Source: Nexis Uni") +
  scale_color_manual(values = c("darkgreen", "gray", "red"),
                     labels = c("Positive", "Neutral", "Negative")) +
  theme_classic()

headline_sent
```

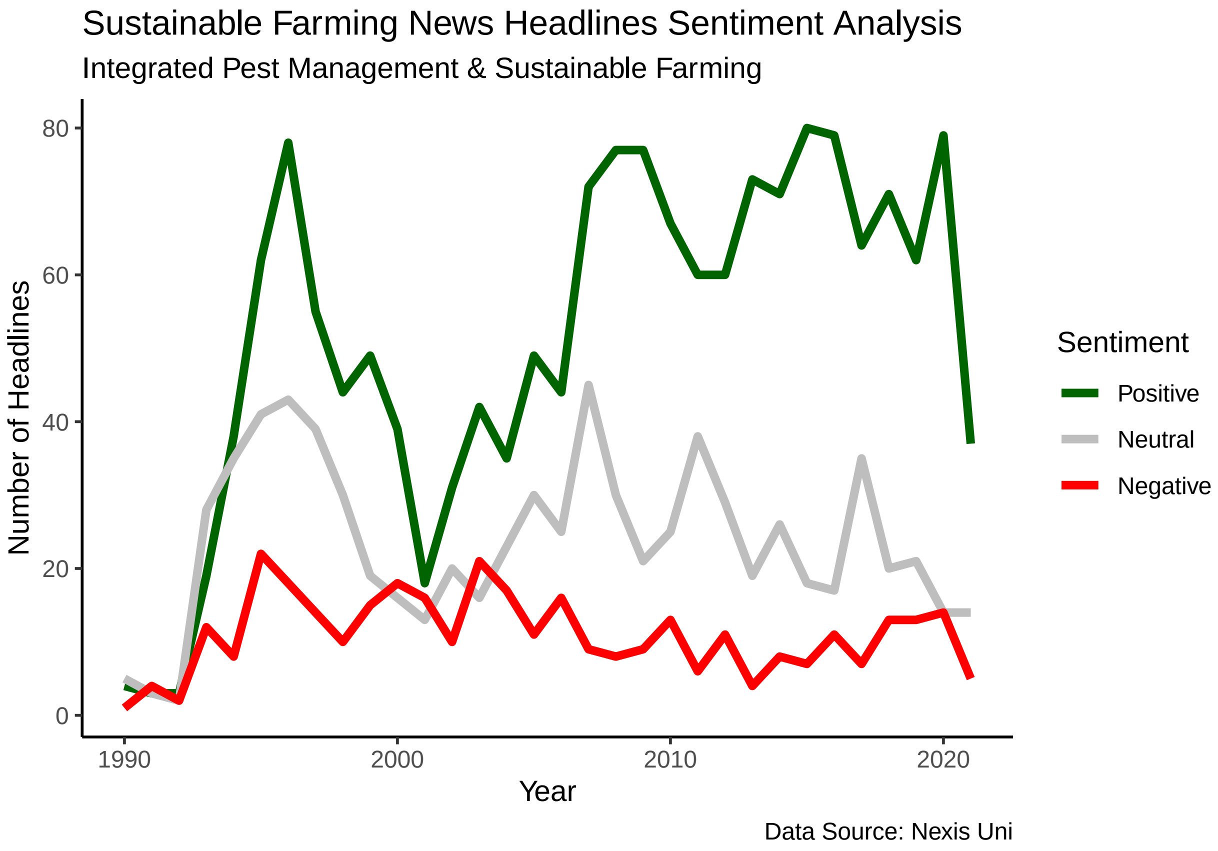

Figure 1: News media sentiment from 1990 to 2021 based on headlines with positive (green), negative (red), and neutral (grey) sentiment. Data Source: Nexis Uni.

Unnest Data into Words

```
bing_sent <- get_sentiments('bing')

#unnest to word-level tokens, remove stop words, and join sentiment words
text_words <- dat3  %>%
  unnest_tokens(output = word, input = text, token = 'words') %>%
  anti_join(stop_words, by = 'word') %>%
  inner_join(bing_sent, by = "word")
```

Calculate the Sentiment Score of Text

This calculation was done by counting the number of sentiment words occurring per day.

```
sent_score <- text_words %>%
  mutate(year = lubridate::year(date)) %>%
  count(sentiment, year) %>%
  spread(sentiment, n)

sent_score[is.na(sent_score)] <- 0

sent_score <- sent_score %>%
  mutate(raw_score = positive - negative,
         offset = mean(positive - negative),
         offset_score = (positive - negative) - offset) %>%
  arrange(desc(raw_score))
```

Create a Plot of Sentiment Scores

```
text_sent <- ggplot(sent_score, aes(x = year)) +
  geom_bar(aes(y = raw_score), stat = 'identity', fill = 'slateblue3') +
  geom_bar(aes(y = offset_score), stat = 'identity', fill = 'red4') +
  geom_hline(yintercept = sent_score$offset[1], linetype = 'dashed', size = .5) +
  theme_classic() +
  labs(title = 'News Article Text Sentiment Analysis',
       subtitle = "Integrated Pest Management & Sustainable Farming",
       caption = "Data Source: Nexis Uni",
       y = 'Sentiment Score',
       x = "Date")

text_sent
```

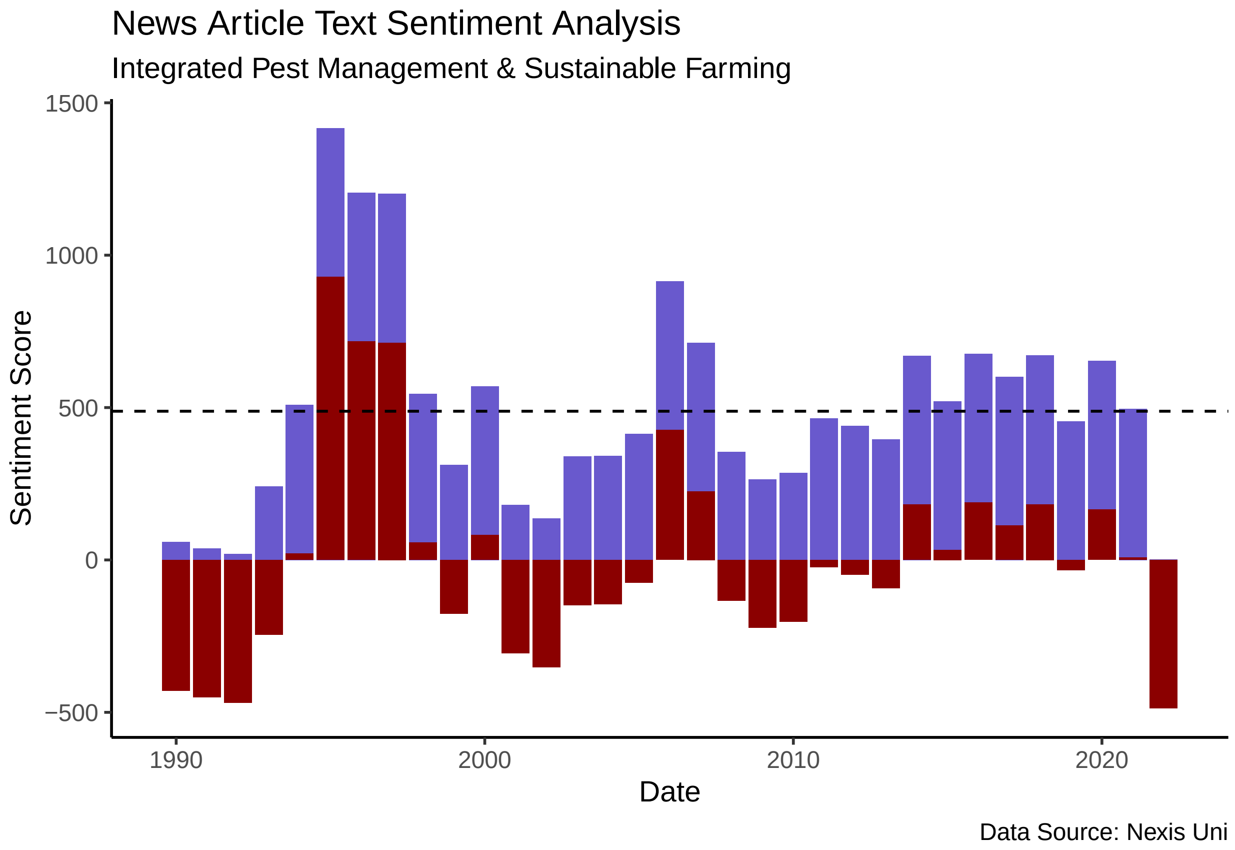

Figure 2: News article text sentiment analysis from 1990 to 2021. The raw sentiment score (purple) and offset sentiment score (red) ... Data Source: Nexis Uni.

Perform Sentiment Analysis using the NRC Sentiment Words

```
nrc_sent <- get_sentiments('nrc')

# restart with the initial data
# unnest into words again before inner joining with nrc sentiment words
slr_nrc_sent <- dat3  %>%
  mutate(year = lubridate::year(date)) %>%
  unnest_tokens(output = word, input = text, token = 'words') %>%
  inner_join(nrc_sent) %>%
  count(word, sentiment, year, sort = TRUE) %>%
  ungroup()

# remove the positive and negative sentiments from the list
slr_nrc_sent <- subset(slr_nrc_sent, !(sentiment %in% c("positive", "negative")))
```

```
slr_nrc_sent_graph <- slr_nrc_sent %>%
  group_by(sentiment) %>%
  slice_max(n, n = 20) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to Sentiment",
       y = NULL,
       caption = "Data Source: Nexis Uni, NRC Emotion Lexicon") +
  theme_minimal()

slr_nrc_sent_graph
```

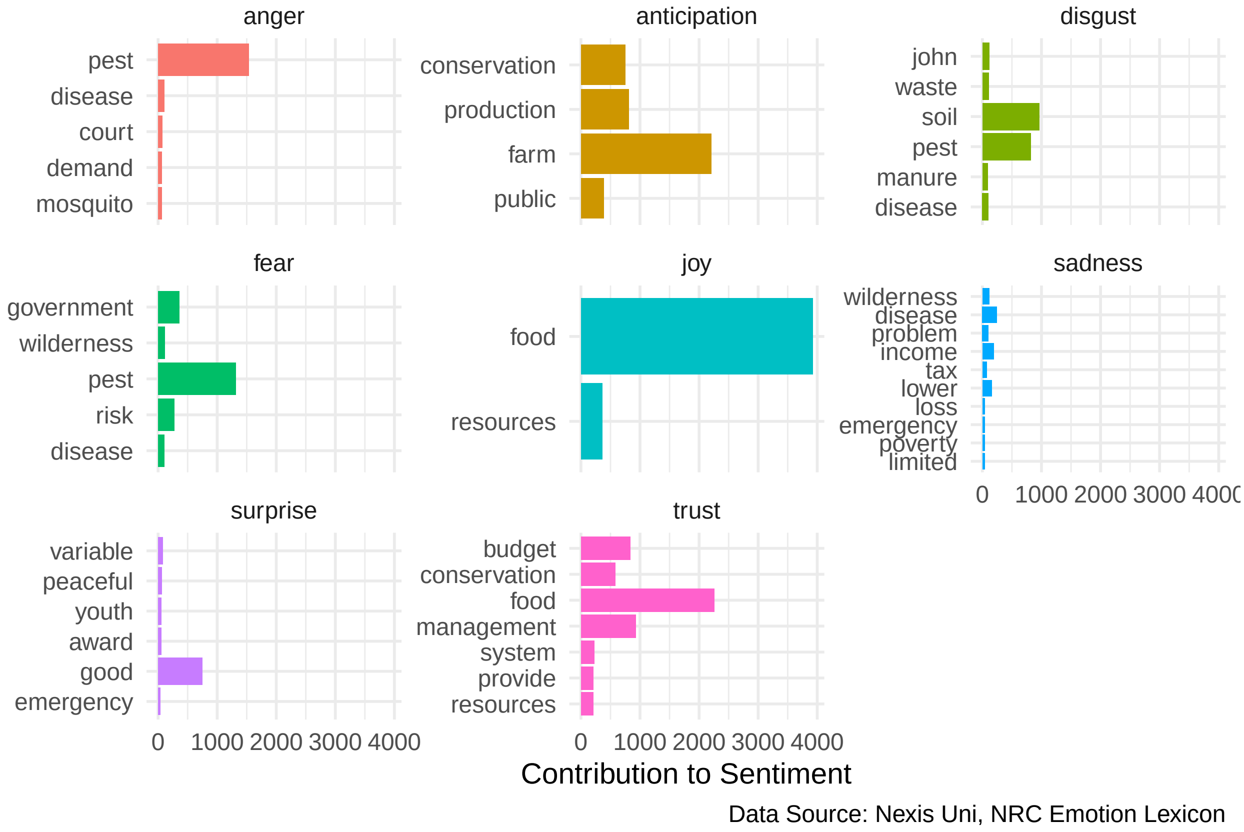

Figure 3: Data Source: Nexis Uni

Calculate the Distribution of Sentiment

```
slr_nrc_sent2 <- slr_nrc_sent %>%
  group_by(year, sentiment) %>%
  summarise(total_n = sum(n)) %>%
  spread(sentiment, total_n) %>%
  ungroup()

slr_nrc_sent2[is.na(slr_nrc_sent2)] = 0

slr_nrc_sent2 <- slr_nrc_sent2 %>%
  mutate(totals = anger + anticipation + disgust + fear + joy + sadness + surprise + trust)

slr_nrc_sent2 <- slr_nrc_sent2 %>%
  pivot_longer(cols = !c("year", "totals"), names_to = "sentiment", values_to = "n")
```

```
slr_nrc_sent2 <- slr_nrc_sent2 %>%
  mutate(percentage = n / totals)
```

Daily Percentage of Sentiment Words

```
sent_words <- ggplot(data = slr_nrc_sent2, aes(x = year, y = percentage, color = sentiment)) +
  geom_smooth(method = lm, se = FALSE) +
  theme_classic() +
  labs(x = "Date",
       y = "Daily Percentage of Sentiment",
       color = "Sentiment",
       title = "Sustainable Farming News Article Sentiment Analysis",
       subtitle = "NRC Emotion Lexicon",
       caption = "Data Source: Nexis Uni, NRC Emotion Lexicon")

sent_words
```

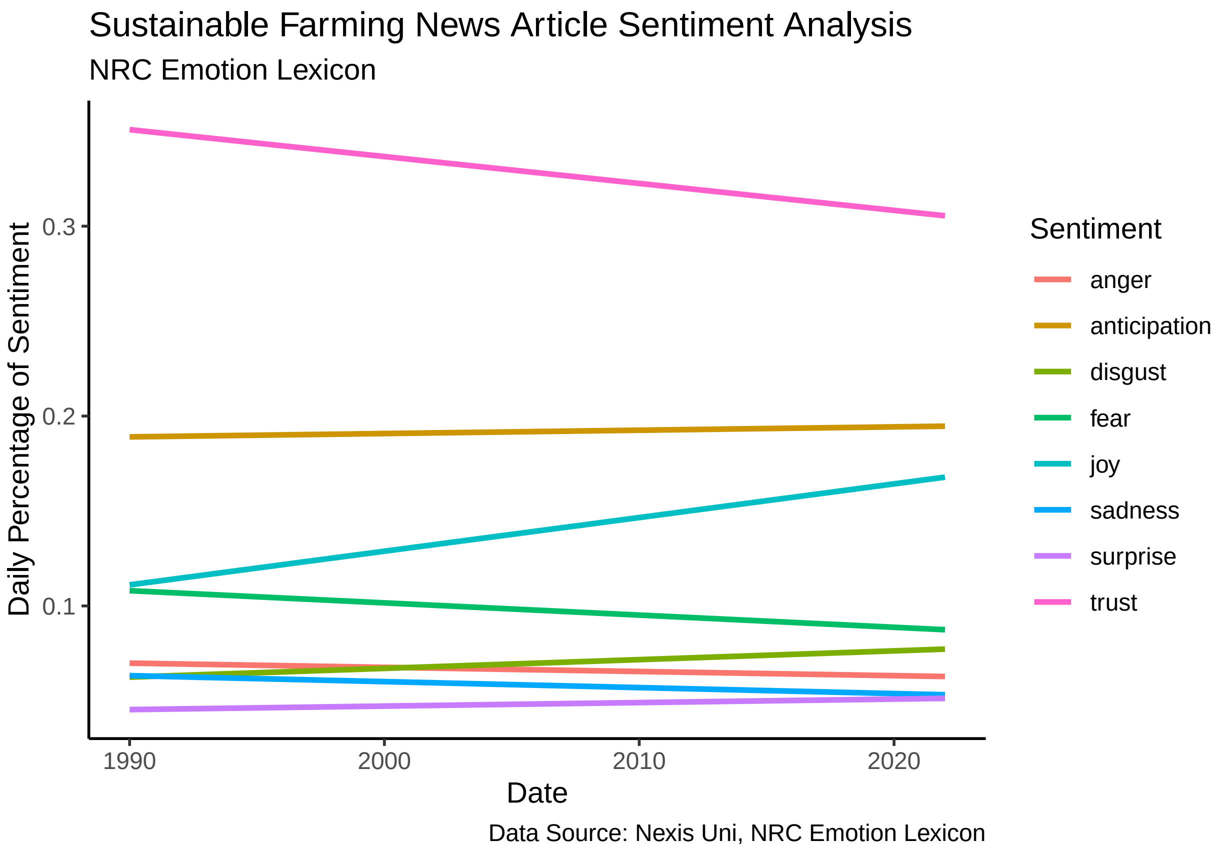

Figure 4: Data Source: Nexis Uni, NRC Emotion Lexicon

Number of Sustainable Farming News Articles: 1990 - 2021

Here we are looking at the number of articles that were published from 1990-2021. These articles were found using the Nexis Uni database using the search terms "Integrated Pest Management" and "sustainable farming".

```r
# Create a table that contains the total number of published articles for each year.
articles_by_year <- data.frame(year = 1990:2021, n_articles = 1)

# Add number of publications for each year.
articles_by_year[1,2] <- 9
articles_by_year[2,2] <- 10
articles_by_year[3,2] <- 7
articles_by_year[4,2] <- 36
articles_by_year[5,2] <- 57
articles_by_year[6,2] <- 76
articles_by_year[7,2] <- 89
articles_by_year[8,2] <- 76
articles_by_year[9,2] <- 67
articles_by_year[10,2] <- 70
articles_by_year[11,2] <- 52
articles_by_year[12,2] <- 47
articles_by_year[13,2] <- 56
articles_by_year[14,2] <- 69
articles_by_year[15,2] <- 70
articles_by_year[16,2] <- 78
articles_by_year[17,2] <- 82
articles_by_year[18,2] <- 100
articles_by_year[19,2] <- 176
articles_by_year[20,2] <- 207
articles_by_year[21,2] <- 227
articles_by_year[22,2] <- 234
articles_by_year[23,2] <- 386
articles_by_year[24,2] <- 381
articles_by_year[25,2] <- 411
articles_by_year[26,2] <- 485
articles_by_year[27,2] <- 696
articles_by_year[28,2] <- 606
articles_by_year[29,2] <- 710
articles_by_year[30,2] <- 905
articles_by_year[31,2] <- 1000
articles_by_year[32,2] <- 1118
```

```r
# Now plot the above data.
articles_plot <- ggplot(articles_by_year, aes(x = year, y = n_articles)) +
  geom_line(size = 1.5, col = "darkgreen") +
  labs(title = "Number of Articles Published by Year",
       subtitle = 'Search terms "Integrated Pest Management" and "Sustainable Farming".',
       caption = "Data: Nexis Uni",
       x = "Year",
       y = "Number of Articles Published") +
  theme_classic()

articles_plot
```

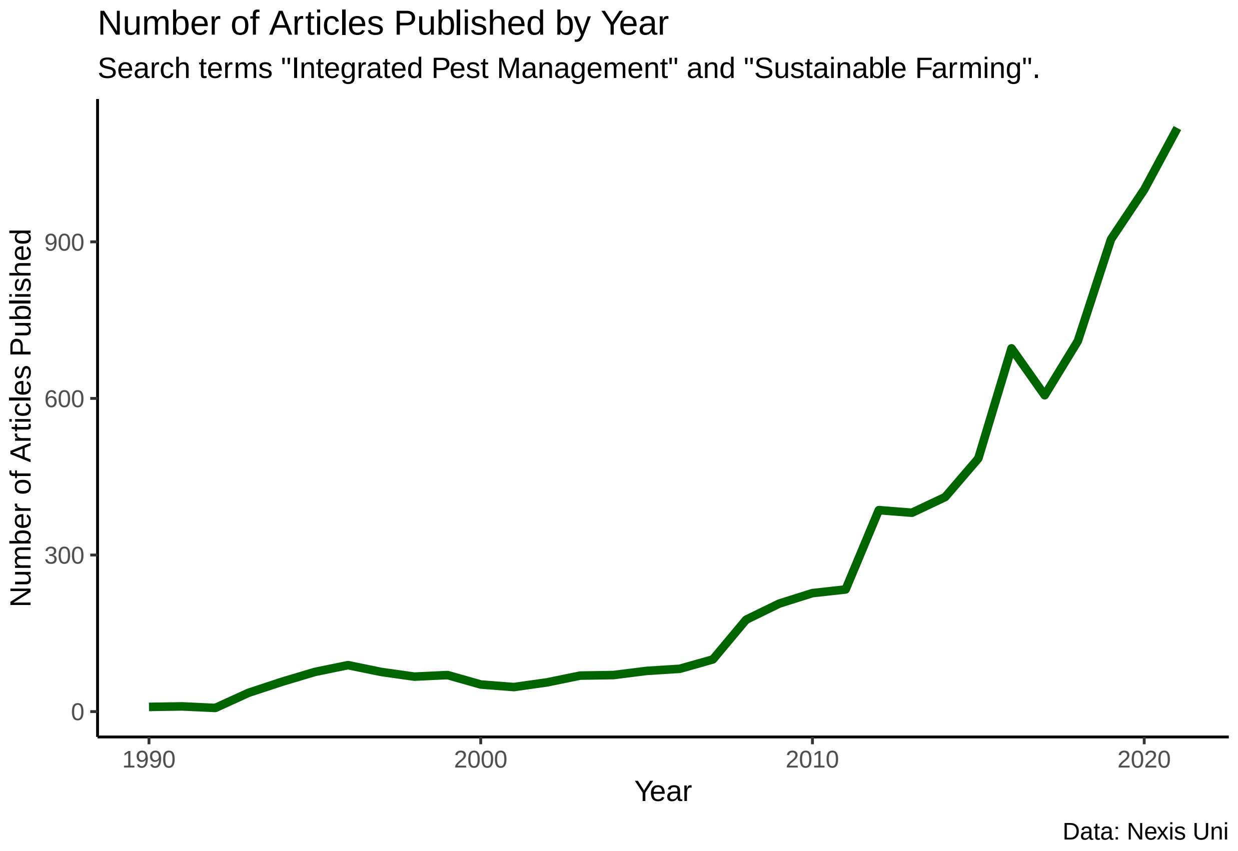

Figure 5: Trend of news articles that include the search terms ‘Intergrated Pest Managment’ and ‘sustainable farming’ over time from 1990 to 2021. Data Source: Nexis Uni

Global FAO data on farming practices over time

Data was collected from the Food and Agriculture Organization (FAO) of the United Nations on global farming practices since 1960. The data was collected from 1990 to 2021 from the FAO data service website, and combined into one csv file.

```
farming_practices_full <- list.files(path = "data/farming_practices/", full.names = T)

farming_practices <- list.files(path = "data/farming_practices/", full.names = F)

practice_names <- str_replace(farming_practices, "fao_", "") %>% str_replace(".csv", "")

practices_df <- map(farming_practices_full, read_csv) %>%
        map(~ select(., Area, Element, Item, Year, Unit, Value)) %>%
        reduce(rbind)
```

```
summary_practices <- practices_df %>%
        group_by(Item, Year, Unit) %>%
        filter(Unit != "%") %>%
        summarize(value = sum(Value))
```

```
graph_practices <- function(topic) {

filtered_practice <- summary_practices %>%
        filter(Item == topic)

filtered_practice %>%
        ggplot(aes(x = Year, y = value)) +
        geom_line(size = 1.5, color = "darkgreen") +
        expand_limits(y = 0) +
        ylab(paste(filtered_practice$Unit[1], topic)) +
        theme_classic()
}

topics <- summary_practices %>%
        ungroup() %>%
        select(Item) %>%
        distinct() %>%
        pull()

practices_graphs <- map(topics, graph_practices)
```

```
filenames_practices_graphs <- paste0(topics, ".pdf")

ggsave_med <- partial(ggsave, device = "pdf", width = 10, height = 6, units = "in")

map2(filenames_practices_graphs, practices_graphs, ggsave_med)
```

```
## [[1]]
## [1] "Cropland area certified organic.pdf"
##
## [[2]]
## [1] "Cropland area under organic agric..pdf"
##
## [[3]]
## [1] "Nutrient nitrogen N (total).pdf"
```

```
##
## [[4]]
## [1] "Nutrient phosphate P2O5 (total).pdf"
##
## [[5]]
## [1] "Nutrient potash K2O (total).pdf"
##
## [[6]]
## [1] "Pesticides (total).pdf"
```

Let's adjust the summary we made earlier by adding a country grouping

```
summary_practices <- practices_df %>%
        group_by(Item, Year, Unit, Area) %>%
        filter(Unit != "%") %>%
        summarize(value = sum(Value))
```

Now, let's filter our data before we graph to only include the U.S.

```
graph_practices <- function(topic, country) {

filtered_practice <- summary_practices %>%
        filter(Item == topic, Area == country)

filtered_practice %>%
        ggplot(aes(x = Year, y = value)) +
        geom_line(size = 1.5, color = "darkgreen") +
        expand_limits(y = 0) +
        ylab(paste(filtered_practice$Unit[1], topic)) +
        theme_classic()
}

topics <- summary_practices %>%
        ungroup() %>%
        select(Item) %>%
        distinct() %>%
        pull()

practices_graphs <- map2(topics, "United States of America", graph_practices)
```

```
filenames_practices_graphs <- paste0(topics, "us_only", ".pdf")

ggsave_med <- partial(ggsave, device = "pdf", width = 10, height = 6, units = "in")

map2(filenames_practices_graphs, practices_graphs, ggsave_med)
```

```
## [[1]]
## [1] "Cropland area certified organicus_only.pdf"
##
## [[2]]
## [1] "Cropland area under organic agric.us_only.pdf"
##
## [[3]]
## [1] "Nutrient nitrogen N (total)us_only.pdf"
##
## [[4]]
## [1] "Nutrient phosphate P2O5 (total)us_only.pdf"
##
## [[5]]
## [1] "Nutrient potash K2O (total)us_only.pdf"
##
## [[6]]
## [1] "Pesticides (total)us_only.pdf"
```