Latent Semantic Analysis and Sentiment Analysis

Nino Miljkovic

June 22, 2025

OBJECTIVE
To explore a large text-based dataset to uncover latent themes and analyze sentiments within the data.

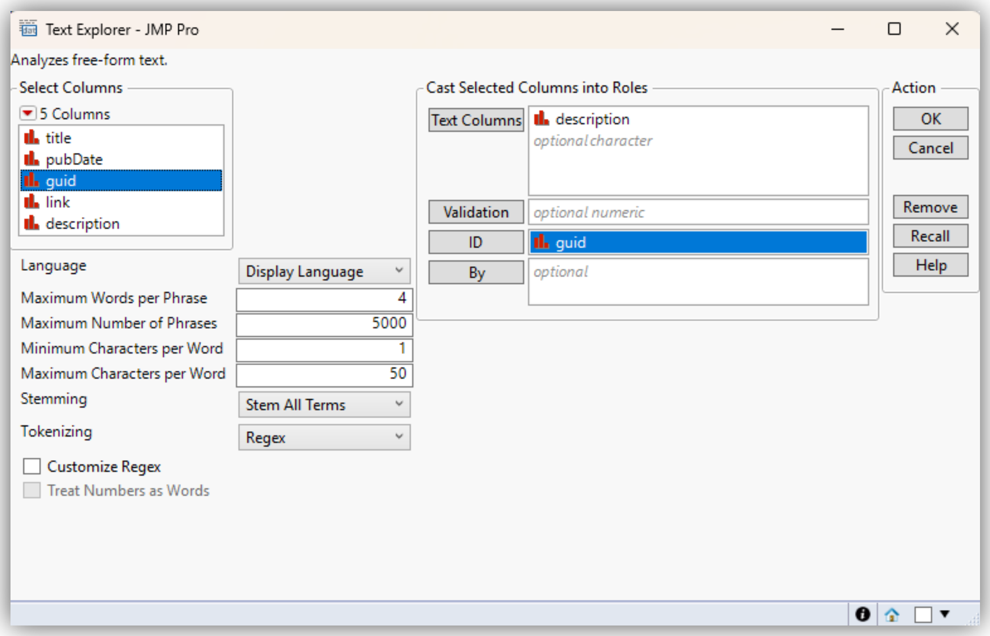
DATASET
Source:
Public Domain dataset found at https://www.kaggle.com/datasets/gpreda/bbc-news

Dataset name:
bbc_news.csv

Description:
The BBC News dataset, created by a data scientist Gabriel Preda, is a publicly accessible and regularly updated collection of BBC news articles collected through their RSS feeds by utilizing requests_html and Beautiful Soup.
The dataset is provided as a CSV file containing around 39,700 values, each representing a news article. Each of those entries is represented by a title, publication date, a unique identifier, the article link, and a brief description. By possessing rich unstructured textual data, this dataset is ideal for text mining projects, including LSA and sentiment analysis.

The copy of the dataset is provided with this submission.

ANALYSIS
The first step was to download the CSV file and verify its content before loading it into JMP Pro. After the verification, the file was loaded into JMP Pro and the Preprocessing stage of the analysis could begin. The first step was to import the textual data into JMP Pro's Text Explorer.

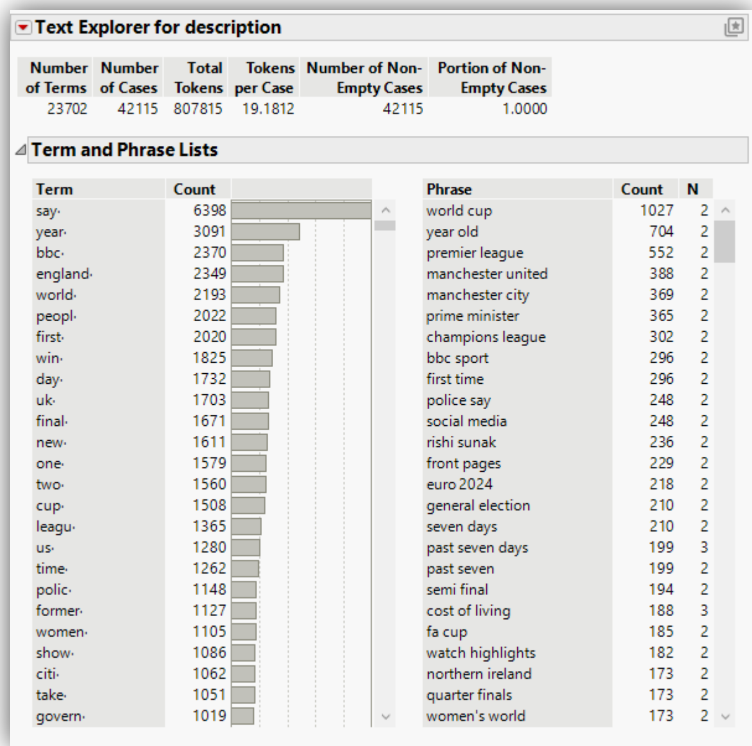


The *description* column containing the unstructured textual data was selected for “Text Columns”, while the unique identifier *guid* was selected for “ID”. *Stem All Terms* was selected as the stemming option.

After clicking OK, the following preprocessing steps were applied:

- **Tokenization:** Text was automatically tokenized by JMP into individual terms and phrases.
- **Stop Words Removal:** Common English stop words were automatically filtered out by JMP’s built-in stop word removal tool.
- **Stemming:** By choosing the *Stem All Terms* option, the words were reduced to their root forms (running→ run) which helped consolidate similar terms.
- **Punctuation and Special Character Removal:** These were removed during the tokenization process.

Lemmatization was not performed, as stemming provided sufficient dimensionality reduction for Latent Semantic Analysis (LSA) and sentiment classification tasks.



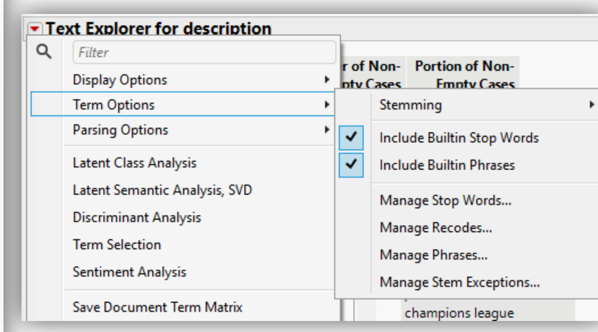
Text Explorer for description

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
23702	42115	807815	19.1812	42115	1.0000

Term and Phrase Lists

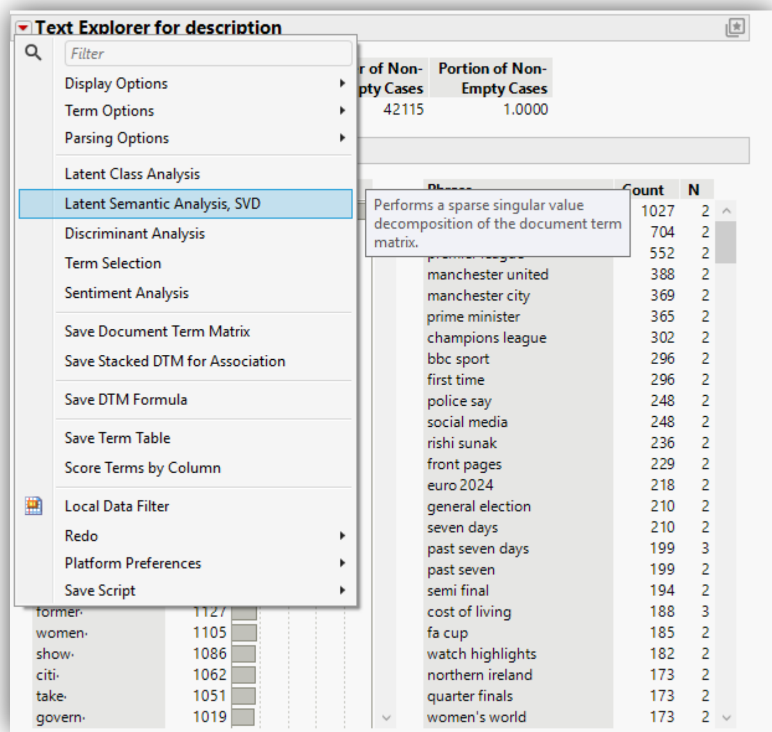
Term	Count
say·	6398
year·	3091
bbc·	2370
england·	2349
world·	2193
peopl·	2022
first·	2020
win·	1825
day·	1732
uk·	1703
final·	1671
new·	1611
one·	1579
two·	1560
cup·	1508
leagu·	1365
us·	1280
time·	1262
polic·	1148
former·	1127
women·	1105
show·	1086
citi·	1062
take·	1051
govern·	1019

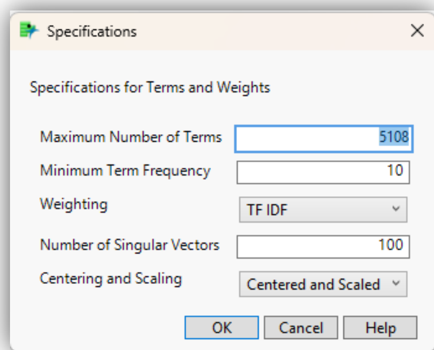
Phrase	Count	N
world cup	1027	2
year old	704	2
premier league	552	2
manchester united	388	2
manchester city	369	2
prime minister	365	2
champions league	302	2
bbc sport	296	2
first time	296	2
police say	248	2
social media	248	2
rishi sunak	236	2
front pages	229	2
euro 2024	218	2
general election	210	2
seven days	210	2
past seven days	199	3
past seven	199	2
semi final	194	2
cost of living	188	3
fa cup	185	2
watch highlights	182	2
northern ireland	173	2
quarter finals	173	2
women's world	173	2



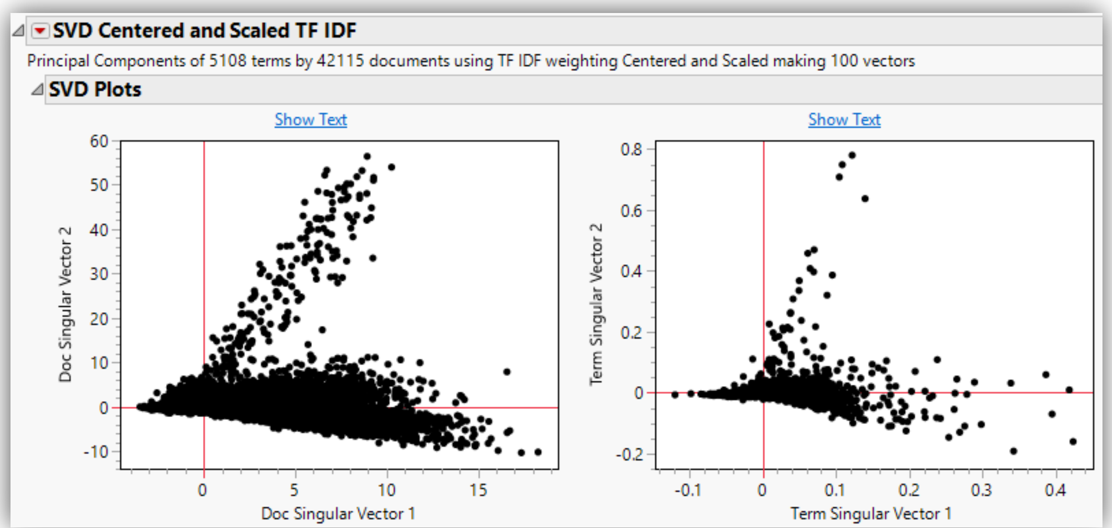
At this stage we are also provided with the List of Terms and Phrases that are most represented as indicated by their count. Terms like “say”, “year”, “bbc”, “england”, “world”, “peopl” and their variations are most common across documents, while “world cup”, “year old”, “premier league”, “manchester united”, “manchester city”, and “prime minister” dominate phrases, possibly indicating that the majority of articles cover themes of politics and sport.

Upon the completion of the preprocessing of the textual data, the next step was to conduct the **Latent Semantic Analysis (LSA)** in order to uncover the underlying themes in the textual data by using **Singular Value Decomposition (SVD)** on a term document matrix with TF-IDF weighting.





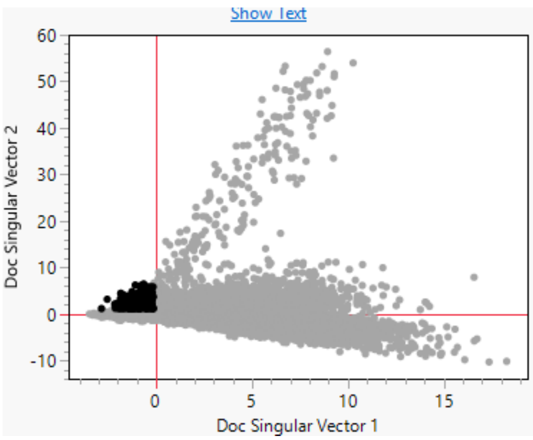
JMP was configured to use 100 singular vectors and the data was centered and scaled to highlight latent patterns. This resulted in two plots shown below.



The left plot (Doc Singular Vector 1 and 2) depicts documents clustered around similar themes, while the right plot (Term Singular Vector 1 and 2) visualizes distribution of terms alongside those same dimensions where densely clustered terms indicate shared context.

By reducing dimensionality SVD plots highlight distinct topic groups. These can be revealed by highlighting quadrants and using *Show Text* to examine the textual data.

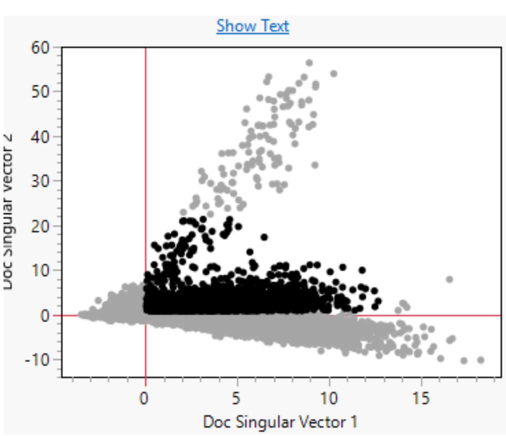
Top Left Quadrant of the Document Plot:




```
Boris Johnson is to meet the Canadian and Dutch PMs, as MPs debate new laws targeting oligarchs. [6]

Boris Johnson is to meet the Canadian and Dutch PMs, as the UK's refugee policy comes under scrutiny.
[70]

Boris Johnson is to meet the Canadian and Dutch PMs, as the UK's refugee policy comes under scrutiny.
[76]

The Killing Eve star reveals why she jumped at the chance to join Pixar's latest film Turning Red.
[147]

The convoy has now largely dispersed, with artillery set up in firing positions, a US company says.
[243]

Reputational risk and practical difficulties are making it harder to do business in the country. [245]

Reputational risk and practical difficulties means the exodus of firms is growing but some remain.
[307]

Families desperate to host Ukrainians in need call for the process to be quicker, with less red tape.
[779]

The former F1 boss killed himself at home after exhausting treatment options, a coroner concludes.
[1042]

Yorkshire chair Lord Patel hails an "overwhelming vote for positive change" as structural reforms at
the
club are approved on Thursday. [1153]

Personal finance expert Martin Lewis calls for government intervention to tackle the cost of living.
[1196]
```

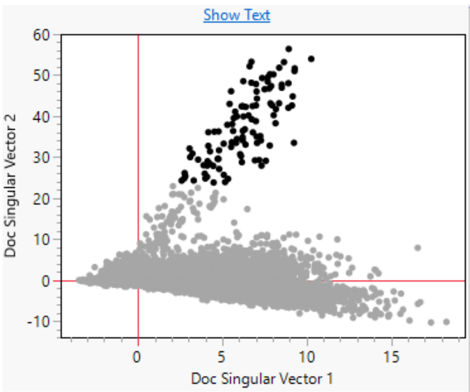
By analyzing this quadrant through *Show Text* function, we can observe that the articles that are clustered together in this section cover themes of global and local politics.

Top Right Quadrant in the Document Plot:




```
Joe Perry beats Judd Trump 9-5 to win the Welsh Open title for the first time at Celtic Manor Resort in
Newport. [49]

Britain's Menna Fitzpatrick and Neil Simpson both win their second medals of the Beijing Winter
Paralympics
with super combined bronze. [50]

Red Bull team principal Christian Horner accuses rivals Mercedes of "bullying" behaviour resulting in
the removal of race director Michael Masi. [126]

Louis Lynagh withdraws from the England squad before Saturday's Six Nations meeting with Ireland
following
a positive test for Covid-19. [128]

World number three Alexander Zverev faces an eight-week ban from tennis if he repeats the behaviour
which
led to his expulsion from the Mexican Open. [157]

Novak Djokovic withdraws from Indian Wells and the Miami Open - the first two Masters Series events of
the year - because of US coronavirus rules. [217]

Mercedes are at the centre of a row over the legality of their car after introducing a radical new
design
at the second pre-season test. [281]
```

The articles that are present in this quadrant all cover different sports, like tennis, formula 1, paralympics, rugby and similar. However, if we shift the focus only on the upper cluster of data points like in the graph below:



2021 Formula 1 world champion Max Verstappen talks to BBC Sport before the 2022 season, which starts with
the Bahrain Grand Prix this weekend. [550]

Red Bull's Max Verstappen edges Ferrari's Charles Leclerc in second practice at the season-opening Bahrain
Grand Prix. [640]

Red Bull's Sergio Perez beats Ferrari's Charles Leclerc to take his first Formula 1 pole position at the
Saudi Arabian Grand Prix. [911]

Max Verstappen wins an intense race-long battle with Charles Leclerc to take his first victory of the season in the Saudi Arabian Grand Prix. [951]

Max Verstappen beats Charles Leclerc to pole position for the Emilia Romagna Grand Prix sprint race at Imola. [2054]

Charles Leclerc leads Carlos Sainz in a Ferrari front-row lock-out at the inaugural Miami Grand Prix. [2671]

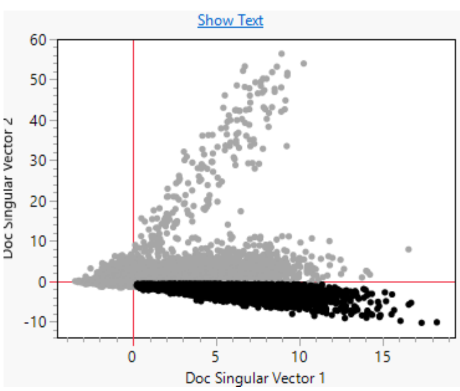
Red Bull's Max Verstappen survives a late-race assault from title rival Charles Leclerc's Ferrari to win
the inaugural Miami Grand Prix. [2719]

Charles Leclerc's Spanish GP misery handed Max Verstappen the title lead, but with resurgent Mercedes in the mix this season is anything but predictable, writes Andrew Benson. [3280]

Ferrari driver Charles Leclerc says Formula 1 dropping the Monaco Grand Prix would be a "bad move". [3456]

We will notice that those articles deal exclusively with Formula One racing.

Bottom Right Quadrant in the Document Plot:



Manchester United legends savage the club's performance in their 4-1 derby defeat by Manchester City. [46]

Who is a candidate for manager of the season? Who could play until he's 40? And who is back to their best?
Find out in Garth Crooks' latest Team of the Week. [48]

MOTD2 pundit Danny Murphy feels Manchester City played like they had a point to prove against Manchester
United and their second-half performance was an emphatic statement to anyone who doubted them. [51]

Match of the Day 2's Danny Murphy and Troy Deeney analyse how the combination play between Phil Foden and Kevin de Bruyne helped Manchester City beat Manchester United 4-1. [73]

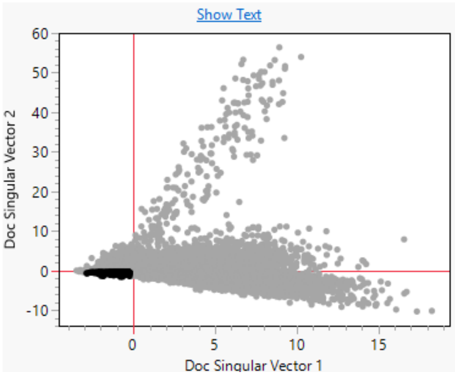
Vivianne Miedema becomes the first player to reach 100 Women's Super League goal contributions as leaders
Arsenal see off a late Birmingham charge to win 4-2 at Meadow Park. [84]

Former West Indies captain Sir Viv Richards, one of the greatest batters of all time, speaks to BBC Sport
on his 70th birthday. [104]

Manchester United's players are becoming increasingly concerned at the managerial void at Old Trafford as their season continues to unravel. [127]

This quadrant contains articles that all share the same topic, which is football, predominantly the Premier League.

Bottom Left Quadrant in the Document Plot:




```
Maxim, a 27-year-old city resident, tells the BBC what happened after the first ceasefire collapsed.
[17]

Civilian attempts to flee Ukrainian cities being targeted by Russian forces lead Monday's papers. [28]

Russia answers resistance with big guns and sieges. Ukrainians pray that will not happen to them,
writes
the BBC's Jeremy Bowen in Kyiv. [40]

At the central train station in Lviv, many thousands arrive after fleeing bombed out cities further
east.
[80]

A 19-year-old Irish woman living in Sumy is making an attempt to leave the city amid Russian attacks.
[82]

At the central train station in Lviv, many thousands arrive after fleeing bombed out cities further
east.
[95]

Russia is still bombarding cities despite offering escape routes. Here's your guide to day 12 of the
war.
[118]

A top official says Russia may close its gas lines to Germany if the West halts oil imports. [130]

The BBC's Fergal Keane travels on a bus with disabled children fleeing the bombed city of Kharkiv.
[142]
```

The articles in this portion of the plot are clustered together due to their shared relationship of topics covering the war in Ukraine.

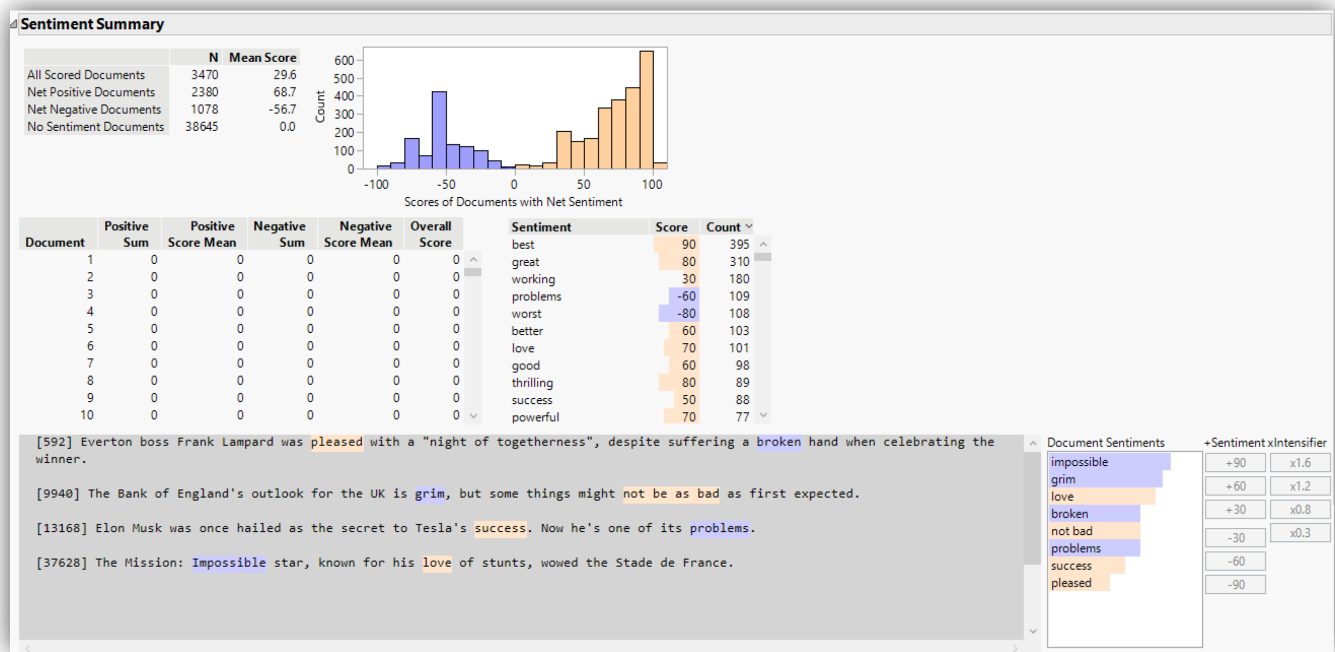
After examining the four quadrants in the SVD Document Plot, we can establish the **four major themes** in the BBC articles:

1. **Politics** (global and local)
2. **Sports** (all sports with a subgroup of articles sharing the theme of Formula One)
3. **Premier League** and other football news
4. **Conflict in Ukraine** and news related to it

The SVD Term Plot quadrants further depict these four themes by having terms related to those four themes clustered together in the corresponding quadrants.

By clearly revealing notable theme groups, Latent Semantic Analysis has helped reduce dimensionality and improve interpretability, thus setting the stage for the Sentiment Analysis.

Sentiment Analysis was conducted using JMP's built-in system that calculates net sentiment score for documents.



The number of all scored documents was 3,470, and of those 2,380 were classified as having net positive sentiment, while 1,078 were classified as having net negative sentiment. The rest either had a neutral or no detectable sentiment. The net positive mean was 68.7, while the negative mean was -56.7.

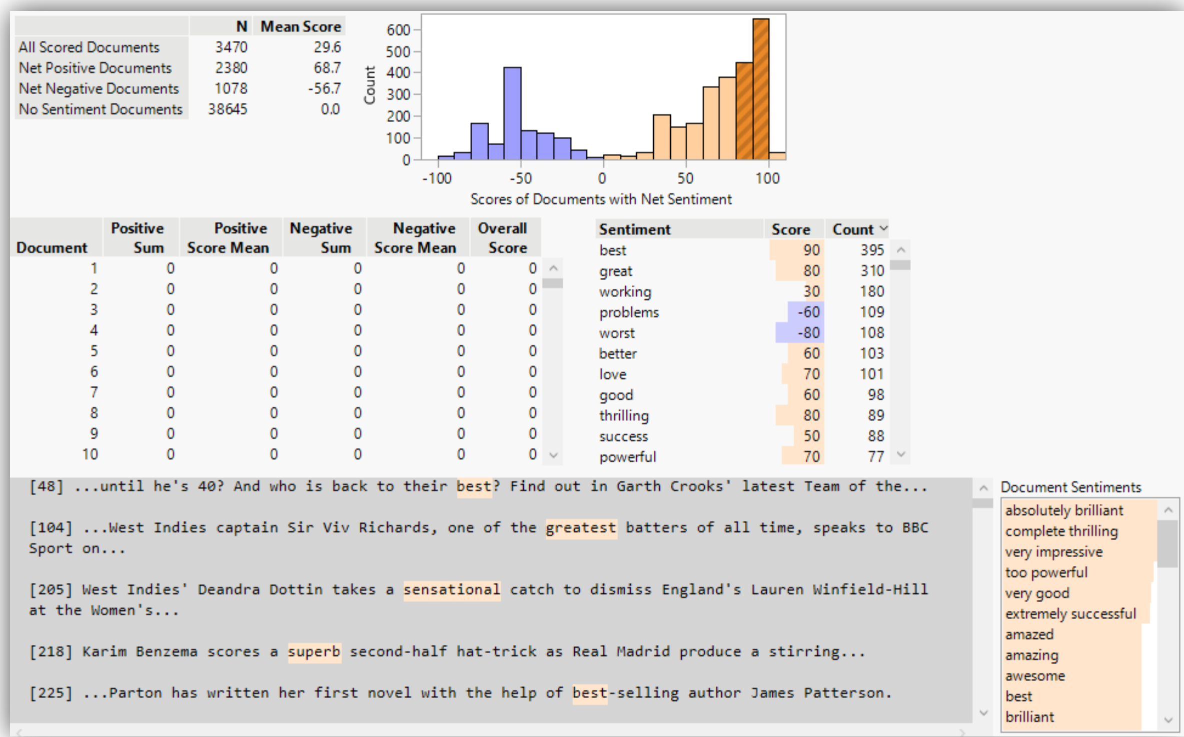
This is visually represented in the histogram that is showing a slight skew towards the net positive sentiment with the peak in +60 to +90 range. The negative sentiment peak was at -60.

The frequently occurring words reveal patterns of what kind of terms are most often related to positive and negative sentiments, as seen in this table:

Sentiment	Score	Count
best	90	395
great	80	310
working	30	180
problems	-60	109
worst	-80	108
better	60	103
love	70	101
good	60	98
thrilling	80	89
success	50	88
powerful	70	77

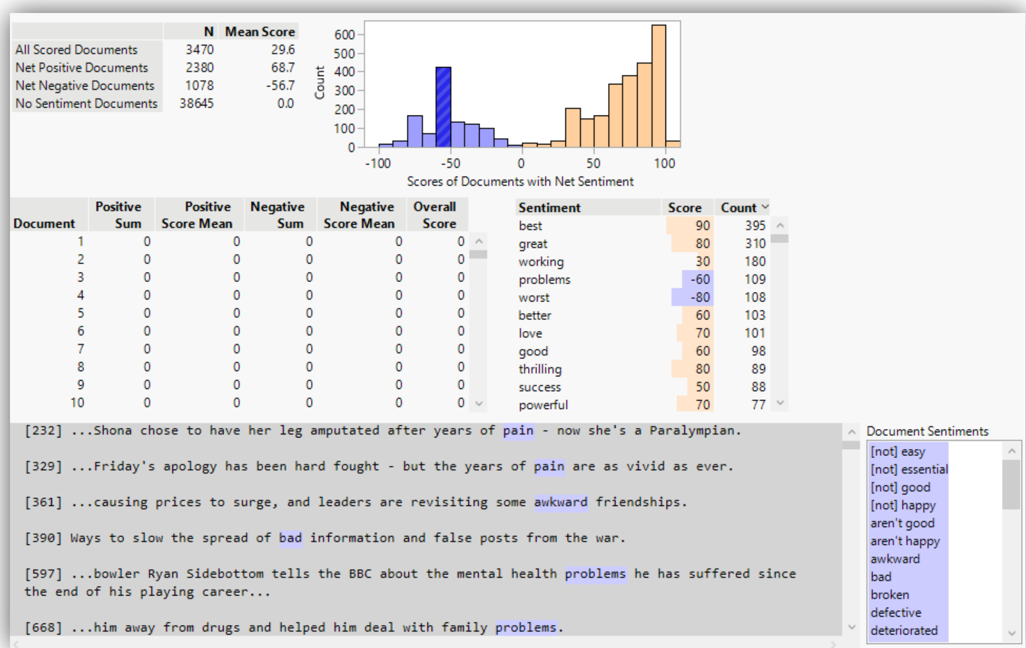
Words like “best”, “great”, “better”, “love” are the most frequently occurring positive terms, while the negative ones are “problems” and “worst”.

By selecting bins of the histogram, we can further investigate different portions of the plot by reviewing the terms and documents associated with them. First, the net positive peaks were chosen as shown in the plot below.



By reviewing the Document Sentiments, we can see the list of words associated with the positive context. Words like "absolutely brilliant", "complete thrilling", "very impressive", "too powerful", and so on. We can also see those words highlighted in the documents where they add a positive sentiment to the topic.

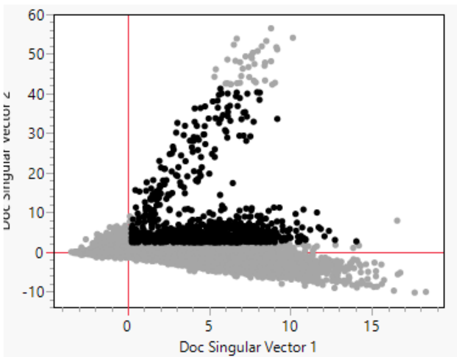
In the case of the net negative sentiment, we can see words that involve negation ("not") and words generally associated with unpleasant feelings and emotions.

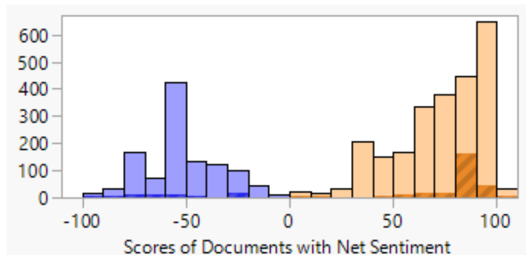


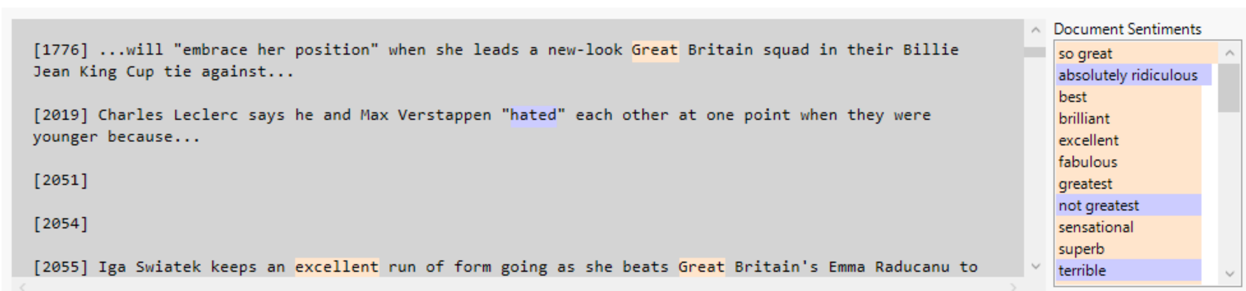
The documents shown highlight the words like “pain”, “awkward”, “bad”, and “problems”, which are adding a negative sentiment to the theme of the document.

Furthermore, by selecting clusters in the SVD plots we can observe the shared sentiment of those particular themes. I will examine two clusters to reveal their overall sentiment.

Cluster 1 (Sports):

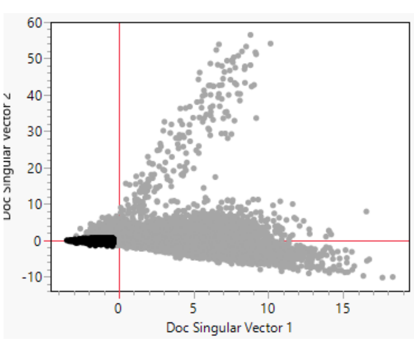


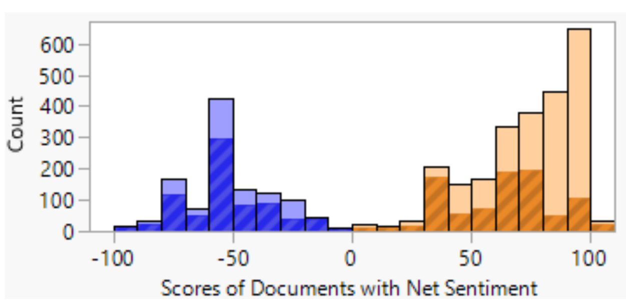




The overall sentiment for news articles covering the topic of sports is a mostly positive one as shown in both the histogram and Document Sentiments. There are several terms usually associated with the negative sentiment but the overall context is positive.

Cluster 2 (Conflict in Ukraine):





As expected, the news articles covering themes of war and crisis will dominate the negative sentiment portion of the histogram. The cluster in the bottom left quadrant makes up a huge majority of all the negative sentiment across all documents. There are, however, cases of positive sentiment as well, but we can say that, overall, the topics surrounding the conflict in Ukraine share a more negative than a positive sentiment.

Overall **Sentiment Analysis** has provided a useful context for the latent themes uncovered in the **Latent Semantic Analysis**, and allowed us to have a richer interpretation of document clusters and their emotional tone.