UseNet `Sci.` Text Mining Project

Description

This analysis of UseNet message board data from 1993 focuses on the `Sci.` groups. In other words, using a text mining approach, what were hot topics of discussion in science in the year 1993? Basic first steps include reading all of the UseNet data into a data frame, cleaning the data, tokenizing, and grouping by newsgroup. Moving forward, we filter for newsgroups containing "sci." First, we analyze top key words within sci. newsgroups using the TF-IDF (Term Frequency-Inverse Document Frequency) metric, which conveys the importance of a word in a particular document within the context of all of the documents in the dataset. Second, Topic Modeling Analysis calculates and visualizes correlations between newsgroups in the full dataset.

Analysis Of Words Within Newsgroups Using TF-IDF Metric

Figure 1 presents the results of TF-IDF analysis on UseNet 'Sci.' newsgroups from 1993, the terms with the highest TF-IDF scores within each group. In 'sci.crypt', terms like 'encryption', 'clipper', and 'cipher' indicate a focus on encryption technologies and related debates. 'sci.electronics' features 'wiring', 'gfci', and 'grounding', suggesting discussions on electrical safety and installations. 'sci.med' shows concerns about 'candida', 'patients', and 'symptoms', reflecting health-related conversations. Lastly, 'sci.space' centers around 'orbit', 'lunar', and 'spacecraft', pointing to a fascination with space exploration and celestial bodies.

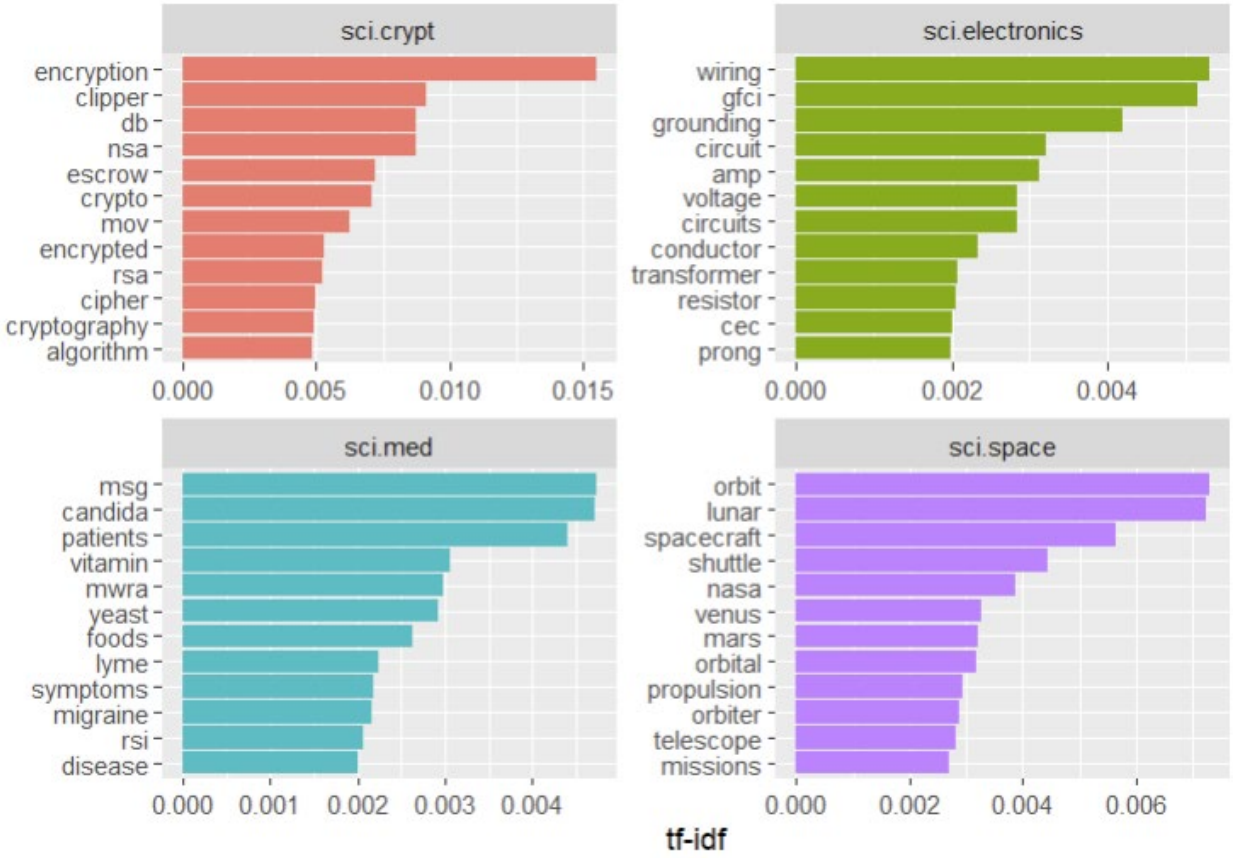

Figure 1: TF-IDF for Sci. UseNet Groups

Topic Modeling Analysis

Religion & Politics

The thickest correlation lines in Figure 2 connect `alt.atheism`, `soc.religion.christian`, and several `talk.politics.` groups, suggesting overlapping discussions of both religious and political viewpoints in this 1993 UseNet board. Religiou and politics are the most interconnected and contentious topics, with strong connections illustrating opposing viewpoints.

Computer Hardware

The comp. newsgroups all cluster together closely, suggesting that discussions of various hardware and software options were strongly linked.

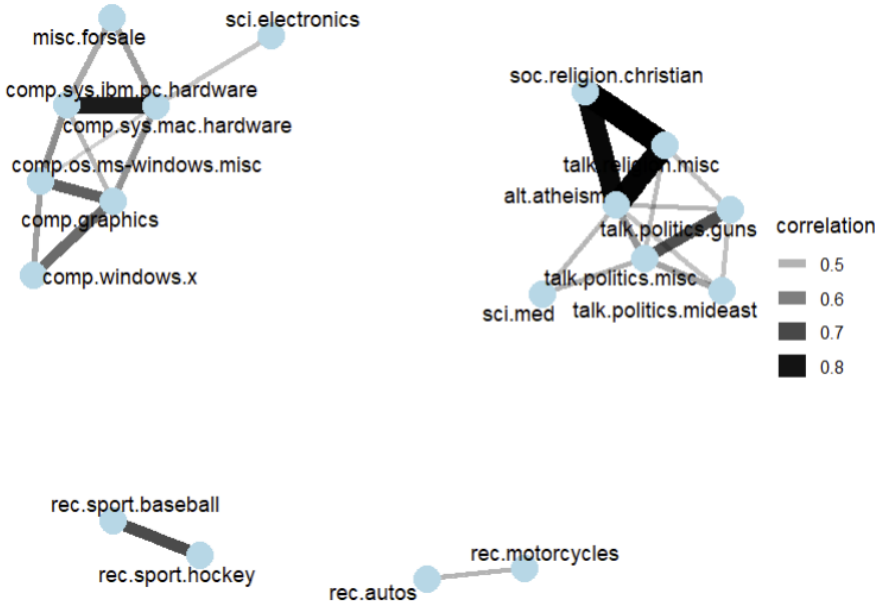

Figure 2: Newsgroup Correlations >0.4.

Thicker lines represent stronger correlations.

Sentiment Analysis by Word

Figure 3 showcases the contributions of various words in the full 1993 UseNet dataset in terms of the words that most strongly influence the perceived sentiment of the broader document. The most positive were “true,” “god,” and “win”; the most negative were “bad,” “hell,” and “wrong.” Figure 4 dives into sentiment for the four `sci.` groups and displays with the strongest influence on overall sentiment per group.

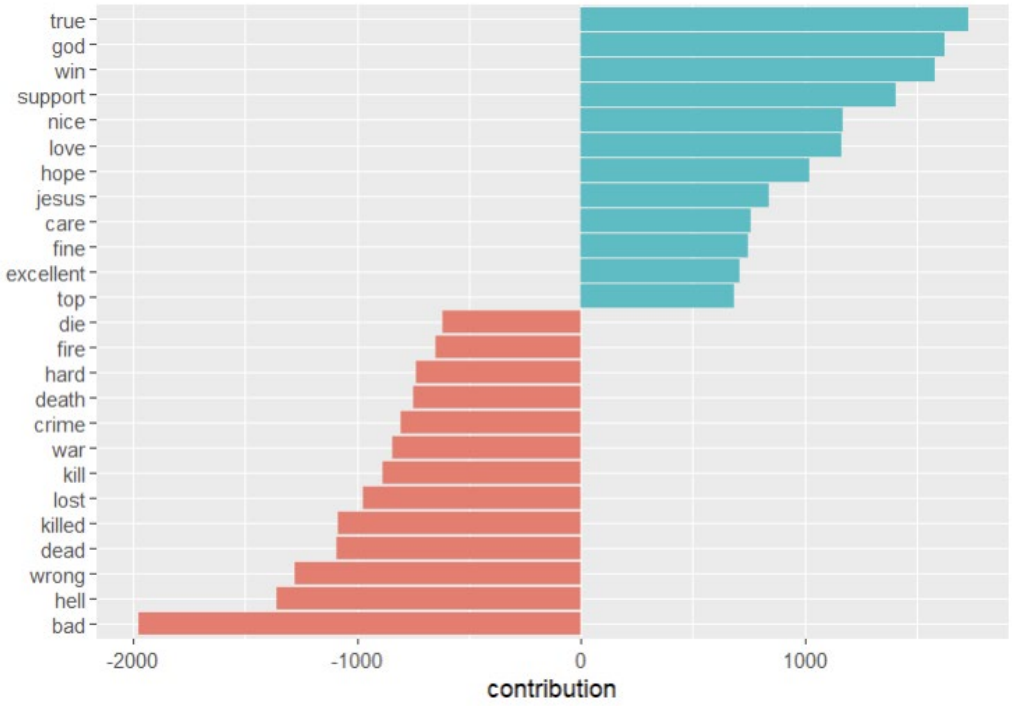

Figure 3: Average Contribution by Word

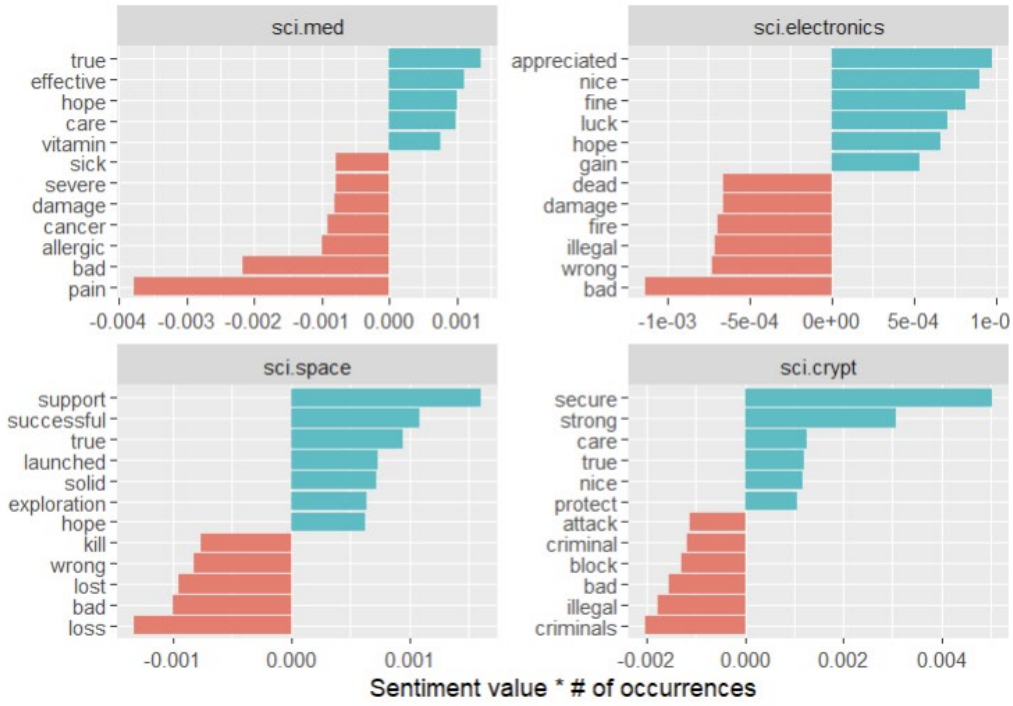

Figure 3: Sentiment Analysis by Word – sci.*

Sentiment Analysis by Message

The two messages in Figure and Figure 5 represent the "most positive" messages by sentiment in the categories sci.space and sci.electronics, respectively. Here they are for your enjoyment.

```
> print_message("sci.space", 61094)
In my first posting on this subject I threw out an idea of how to fund
such a contest without delving to deep into the budget.  I mentioned
granting mineral rights to the winner (my actual wording was, "mining
rights.)  Somebody pointed out, quite correctly, that such rights are
not anybody's to grant (although I imagine it would be a fait accompli
situation for the winner.)  So how about this?  Give the winning group
(I can't see one company or corp doing it) a 10, 20, or 50 year
moratorium on taxes.
Tom Freebairn
```

Figure 4: Most Positive Sentiment – sci.space

```
> print_message("sci.electronics", 53836)
Mixers have a wide variety of implementations; the Mini-Circuits
part you mention is a doubly-balanced diode mixer, but active ones
(BJT, FET) seem more popular in consumer receivers.  You might
call MCL; they have a nice catalog.
The universal answer for wide-coverage, theory+practice, RF design
is the _ARRL Handbook_, published by the American Radio Relay
League, the radio amateur organization.  Any technical bookstore
can order you one.  The book is superb, with lots of accessible
theory, construction projects, and generally interesting stuff.
You might also check out _Solid State Design for the Radio Amateur_
(I think), by Hayward and <someone>.  This has sharper design
and test information about subsystems like mixers.
Peter Monta    monta@image.mit.edu
MIT Advanced Television Research Program
```

Figure 5: Most Positive Sentiment – sci.electronics

N-gram Analysis

Figure 6 explores the sentiment value and frequency of bigrams (two-word phrases) in a text corpus, particularly highlighting the impact of negations. It visualizes how words following specific negations ("can't," "don't," "no," "not," "without," and "won't") shift in sentiment. Generally, positive words paired with negations become negative (e.g., "can't win"), and vice versa (e.g., "no problem"). The length of each bar represents the frequency of the bigram, showcasing which combinations are most common. This analysis provides valuable insight into how negations influence the overall sentiment of the text.

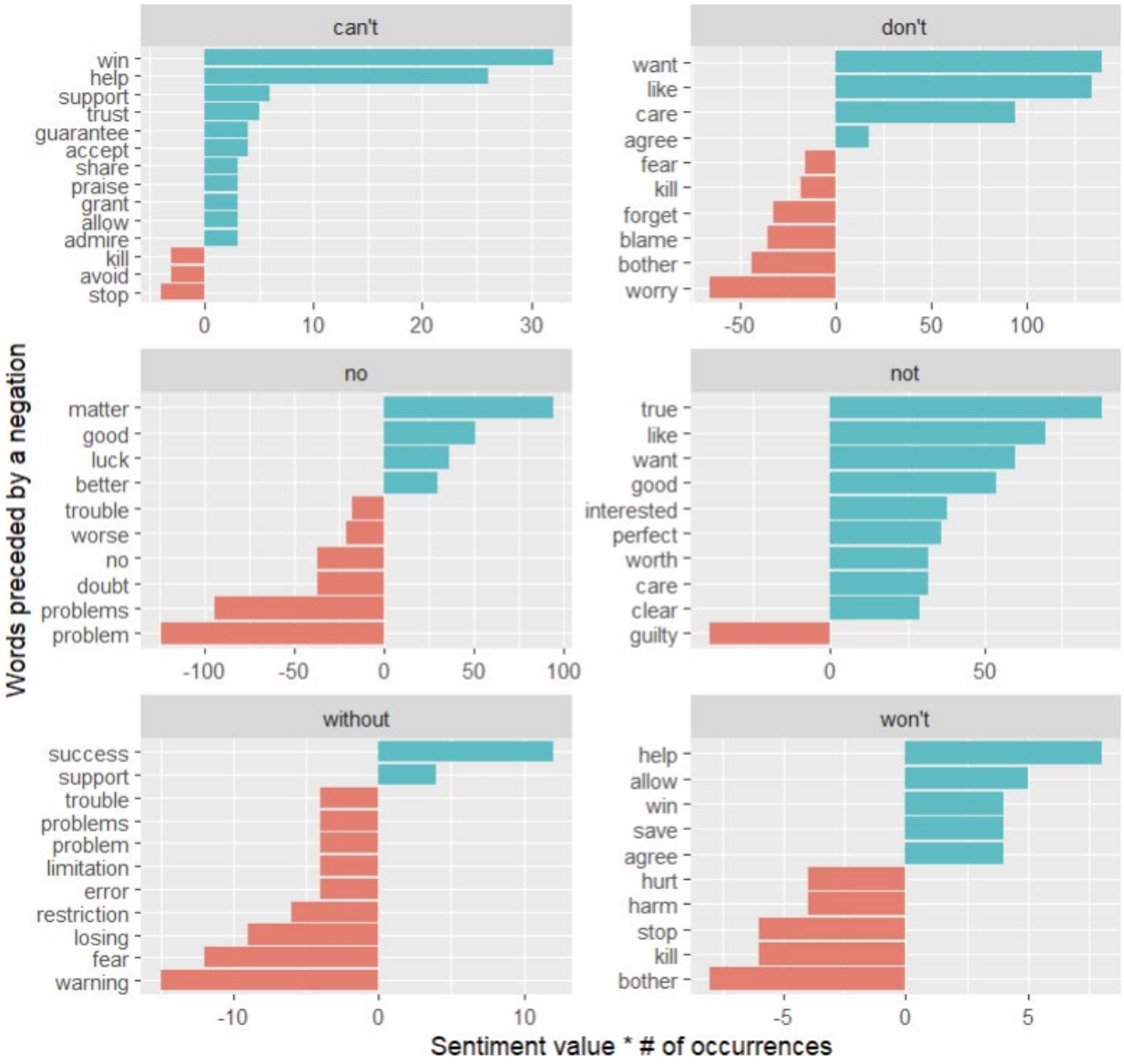

Figure 6: n-gram Analysis