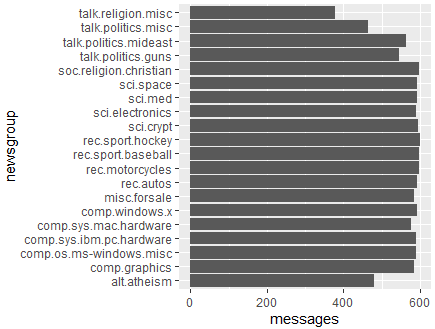
**Part1: Description and preprocessing**

We will read in all the messages from the 20news-bydate folder. The 20news-bydate folder contains sub-folders. We will use the sub-folders to create a training set of data and a test set of data. Read in files with read\_lines(), map(), and unnest().

Create a training folder that contains the 20news-bydate-train data, make function that reads folders, and create a data frame to hold the title of the newsgroup, the message id, and the text that goes with it.

The newsgroup column identifies which of the 20 newsgroups the article came from, the id column identifies a unique message within that newsgroup, and the text column contains the text associated with the newsgroup and id number (i.e. the message).

View names of all 20 newsgroups and the total number of messages within each newsgroup.

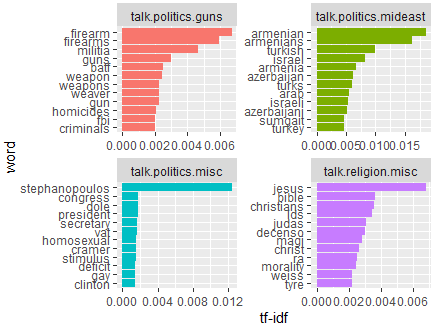


Next we will pre-process the text from the messages within the newsgroups (i.e. clean the data). All the messages include a header like this “from:” or “in\_reply\_to:” and may include email signatures which occur after “–”. Use cumsum() from dplyr and str\_detect() from stringr to remove these lines of text.Some text also included quotes by other writers (use regex), and messages 9704 and 9985 contain limited to no textual data so lets remove them too.Next we tokenize the data using unnest\_tokens() to seperate the messages by each word. Then, we remove stop words from the data.

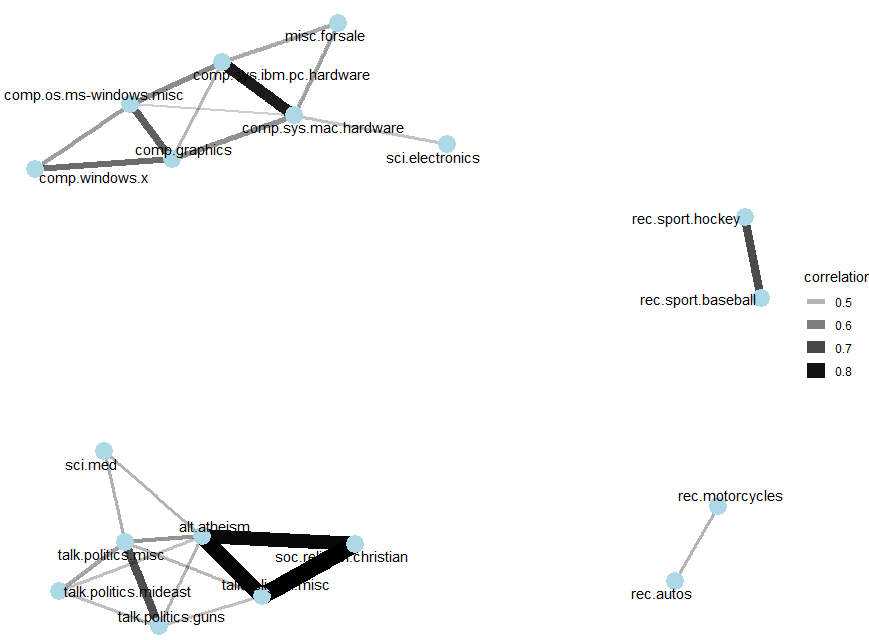
**Part2: Conclusion**

**Words by newsgroups:**

As we know tf-idf identifies important words in each newsgroup. We can also view the highest ranking tf-idf words for a specific topic.

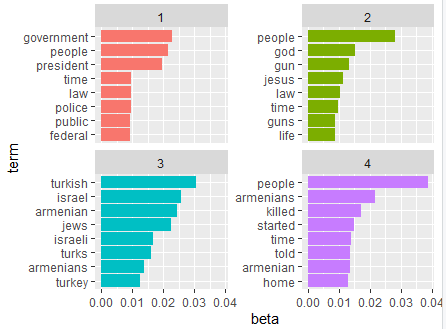


The newsgroups who most often used the same words are talk.religion.misc and soc.religion.christian. These two newsgroups use the same words about 83% of the time.

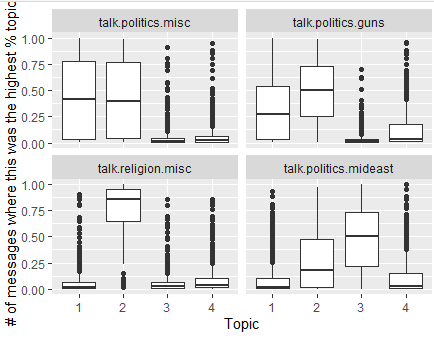


The network suggest there are four main topics: computers, sports, politics, and religion.

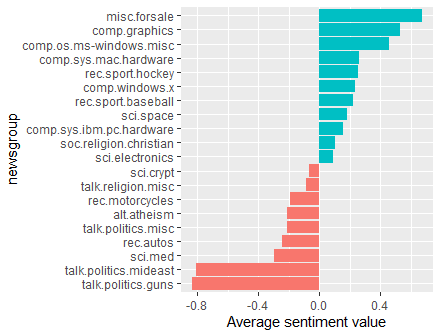
**Topic Modeling**



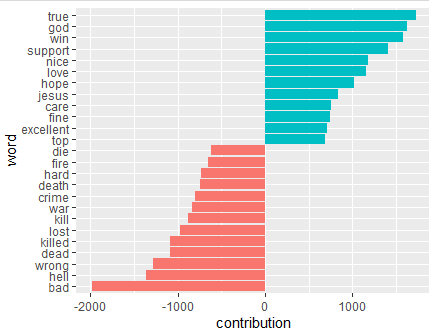
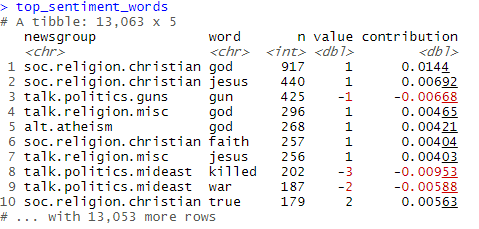
As the network predicted the main topics were politics, religion, sports and computer but here we can see politics, religion and social.

 to be specific the topics were about politics, religion and politics of Mideast.

**Sentiment Analysis**

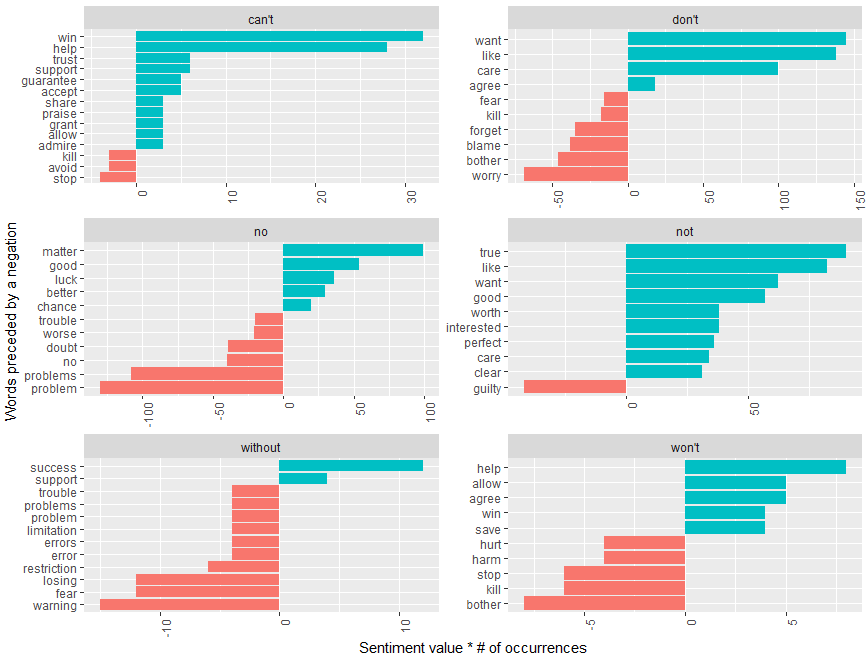


Misc.forsale has the highest score in sentiment and talk.politics.guns has the least score.

Talk.politics.guns promoted words such as guns which have a negative score resulting in having a low sentiment score for the topic.

**N-gram:**



Sentiments identified as negative came when followed by words such as without the most and then followed by wont. Can’t win is miss represented as positive instead of a negative sentiment. Similarly not true, don’t want, won’t help and without success are showing in positive light. Won’t bother and no problem are also largely misrepresented.