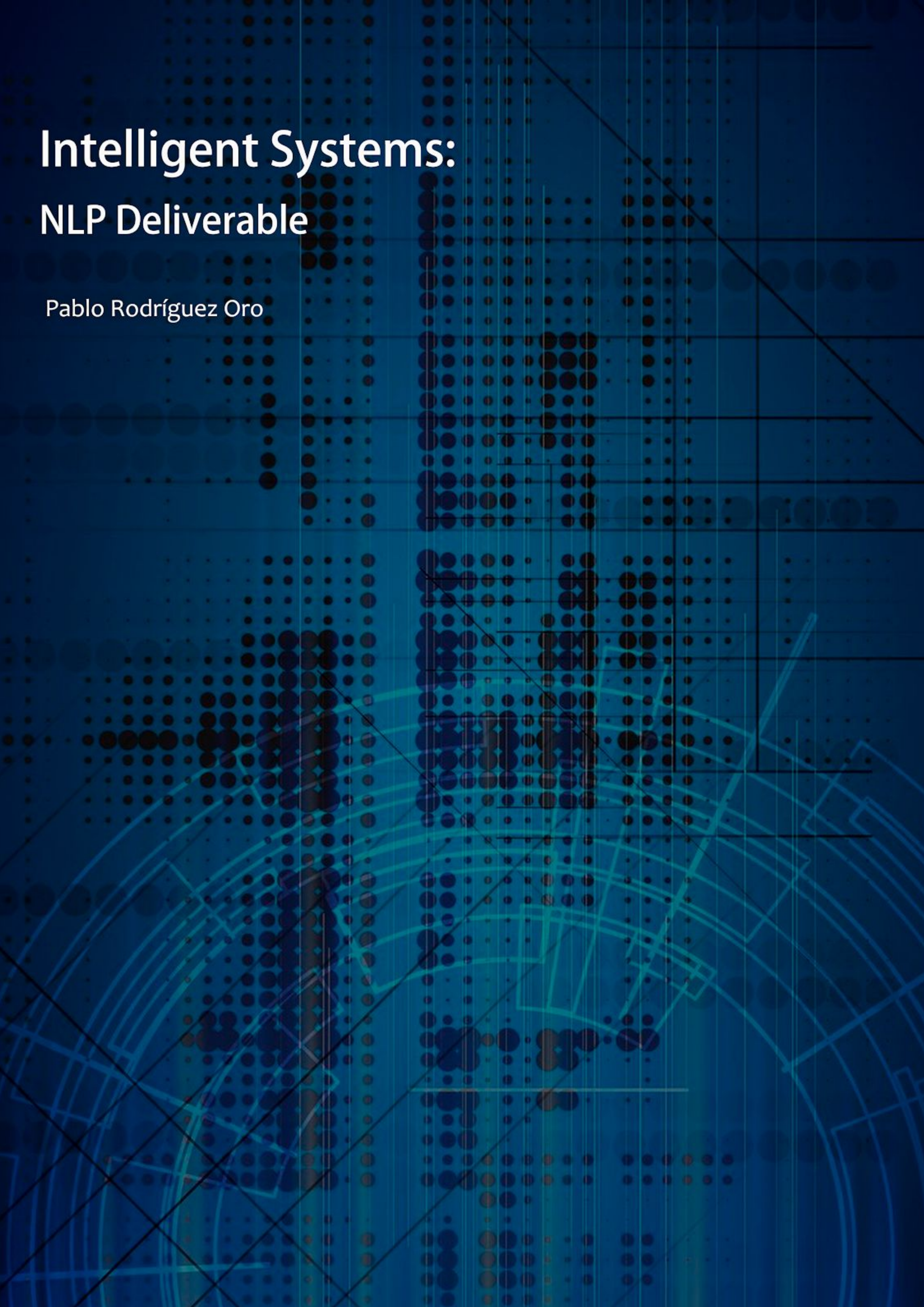


Intelligent Systems:

NLP Deliverable

Pablo Rodríguez Oro

Intelligent Systems: NPL Deliverable

1. Problem to Solve

Nowadays, the public's trust and perception of reliability in the press has been brought into question. One of the main concerns is journalists' bias when covering particular subjects like politics or public policies. This concern makes sentiment analysis of news reports a particularly interesting subject.

In this paper, the PerSenT dataset containing over three thousand news reports covering a wide array of subjects will be used to get an overview on some of the most covered topics and the general sentiment of the press towards them. In particular, this paper will look at the general sentiments towards some of the most talk about popular figures.

2. Experiments Done & Results

-Data Exploration and Preparation:

The first step in any data science project is to take a deeper look into the data to properly assess how to approach the project.

The PerSent dataset has the following variables:

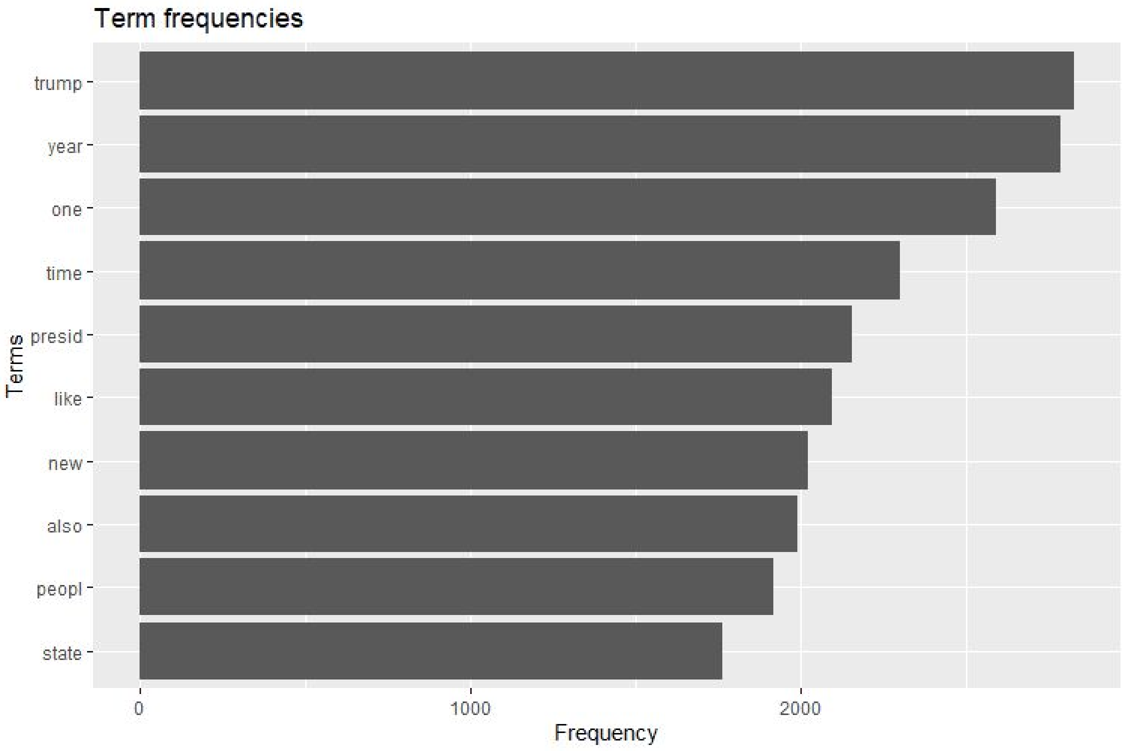
- **Index**
- **Title:** Title of the article
- **Target entity:** Main target of the article. Usually a person or institution
- **Document:** The content of the article itself
- **Masked document:** The article with the Target entity removed
- **True sentiment:** The overall sentiment of the article
- **Paragraph(n):** A series of variables containing the sentiment of each paragraph in each article

For the purpose of this work, the paragraph variables and the masked document will be dropped.

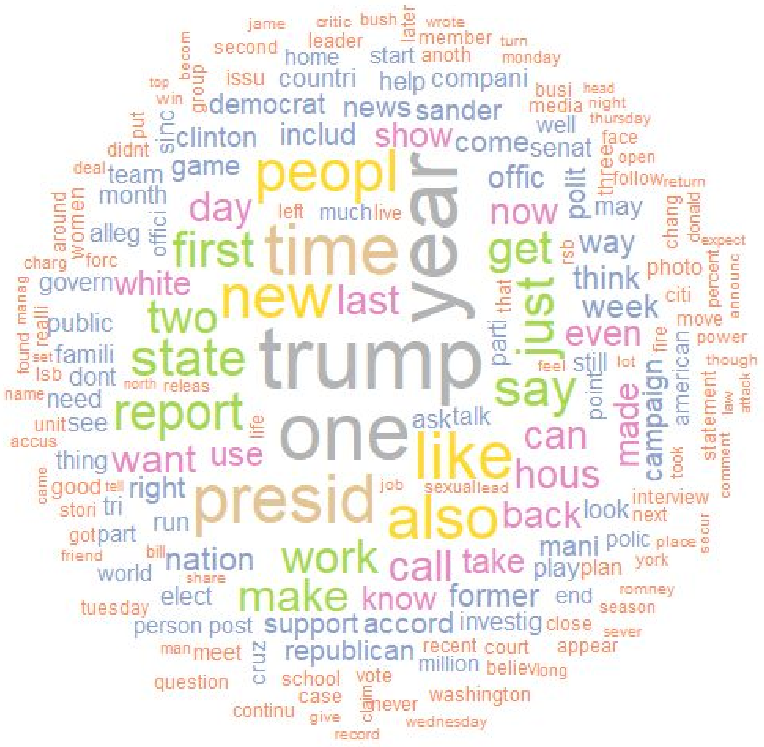
In order to use the text data, it is necessary to remove non-useful characters, like "@", "*" or "/" from the Document. In addition, punctuation, numbers and stop words are removed as well. Finally, the words in the Document are stemmed.

-Metrics:

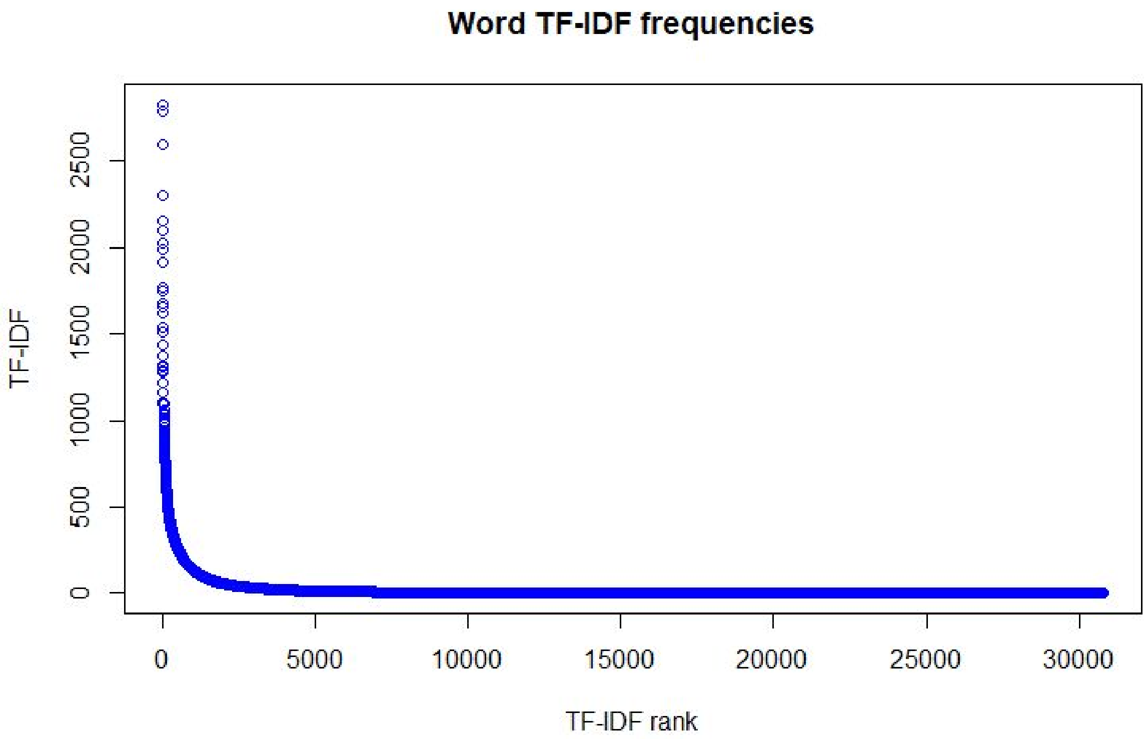
After having prepared the data, we now can take a look at some basic metrics in the text. For instance, let’s look at the ten most common words:



As we can see, the most common words are Trump, year, and one. Let’s also generate a word cloud with the 200 more frequent words:



Now, we can measure the relevance of the words in the document using TF-IDF weighting, obtaining the TF-IDF curve and ten highest rated words:



Trump	year	one	time	president	like	new	also	people	state
2828	2788	2593	2302	2154	2097	2024	1990	1918	1764

-Sentiment Analysis

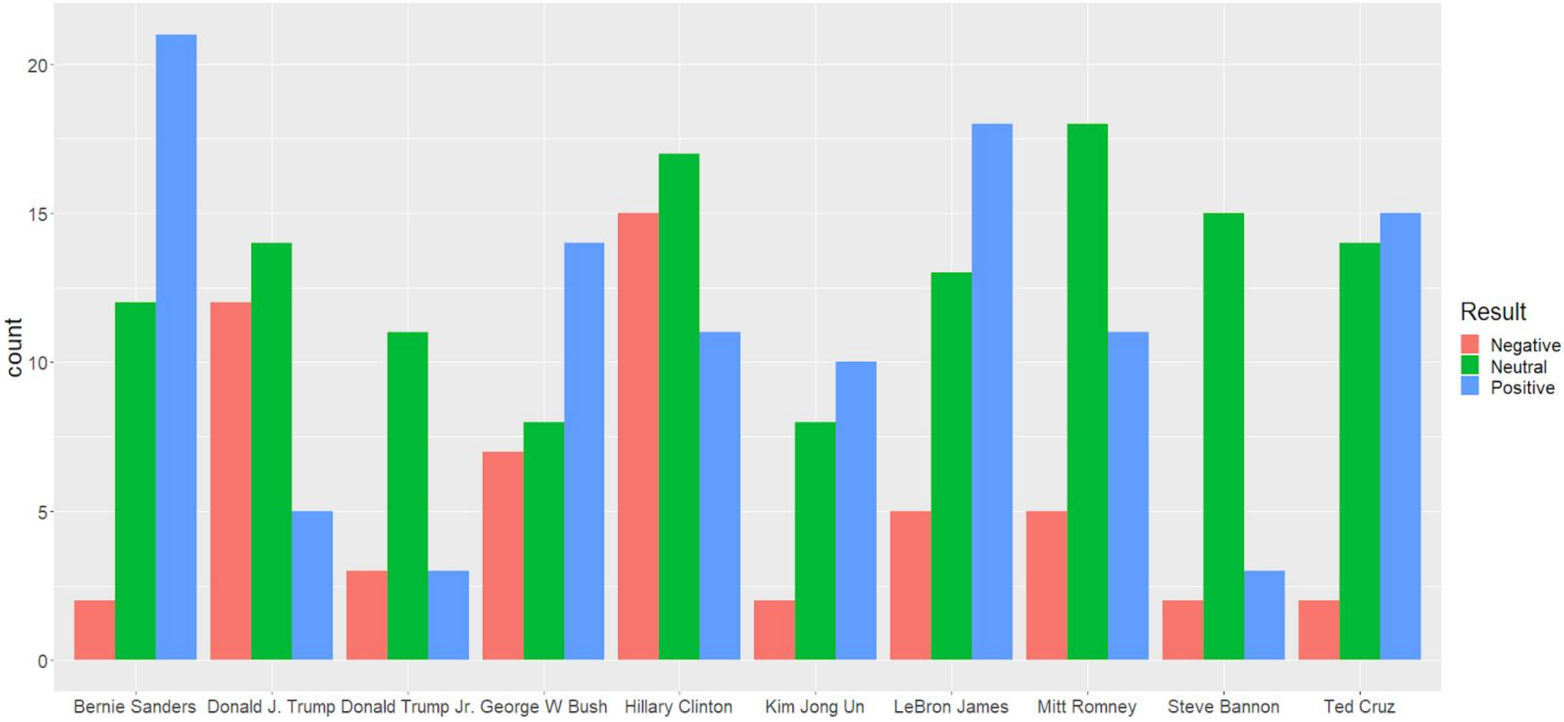
For the sentiment analysis of the Documents, two dictionaries will be used: Loughran-McDonald dictionary (LM) and Harvard General Inquirer dictionary (GI). With these, we obtain a series of score values, with numbers over 0 being associated with positive sentiment and values below 0 with negative sentiment. However, with the implementation of a neutral sentiment class, values below 0.025 and over -0.025 will instead be associated with this sentiment.

After applying the sentiment analysis, the following results were obtained:

Sentiment	**Percentage(%)This Work**	**Percentage(%) PerSenT**
Positive	32.94	52.38
Negative	26.47	10.46
Neutral	40.57	37.14

As it can be seen, most of the news reports have a neutral overall sentiment towards their main target (40.57%), whereas only 26.47% have a negative sentiment. Comparing these results to the ones obtained by the PerSenT team, we obtained a lower percentage of positive and neutral sentiment and a higher for negative.

Now we can see how some of the most popular figures in the media were portrayed:



3. Conclusions

In this paper a series of NLP methods and tools were applied to the PerSenT dataset with the goal of analysing the overall sentiment of the press and towards popular figures. This sentiment was mostly neutral and positive, with just over a quarter being negative, particularly focus of political figures like Donald Trump or Hillary Clinton.