X (Twitter) as a Free and Accessible Source of Public Sentiment

Using Computational Techniques to Analyse Brexit Sentiment on X

Brodie Knight
University of Cambridge
nk624@cam.ac.uk

Problem: Developing a Method to Fetch a Large Volume of Tweets from X

In order to do any analysis on data from X, we first need to download the Tweets we want to analyse. Since the takeover of what was formerly Twitter by Elon Musk, the cost of accessing large volumes of Tweets through the official X API has increased immensely. This makes it prohibitively expensive for small scale/independent projects that wish to use X data, and means that important research can no longer be done.

The only affordable alternative is to scrape the data from the web pages directly. A scraper bot accesses X data in the same way that users can see Tweets by going to the X website. Essentially, a scraper crawls through the web page, picking up the data it needs. However, this is much slower and less efficient than using an API, and there are often measures put in place by the websites to prevent scraping.

Solution: Scraping via 'Nitter' using Distributed Tor Circuits

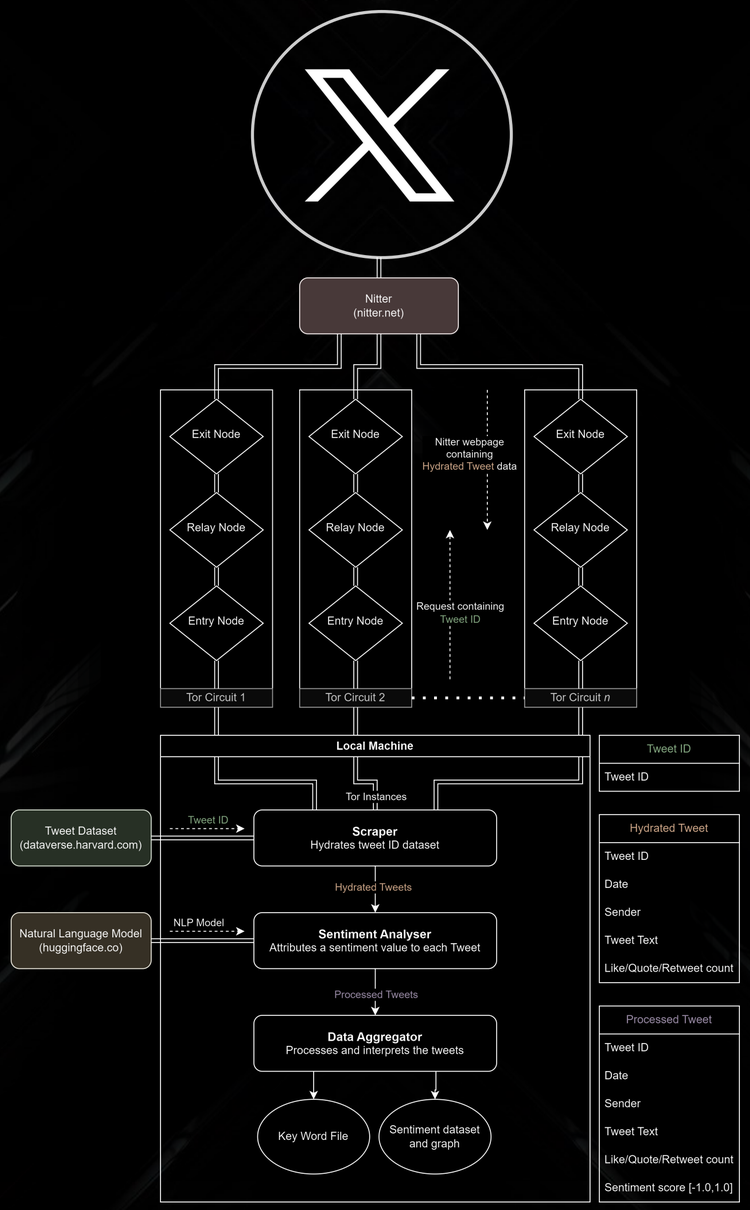
The X website itself is not ideal for scraping since every time we download a webpage we get a lot of unnecessary data in the form of adverts, images and scripts. It turns out that the 'Nitter' website - which was designed to display exactly the same Tweets as X but without the ads or distractions - is a much more suitable website to scrape, since there is much less bloat on the Nitter webpages.

A Tor circuit is essentially a Proxy which establishes a new internet identity. The Tor network is sometimes known as the Dark Web because of its use by criminals. It is possible to overcome Nitter's download rate limitations by distributing a large number of requests through up to 250 Tor Circuits simultaneously. Using this method resulted in an increase in max download rate from 180 to 12000 Tweets per minute.

A complete suite of software was coded for the purposes of this project, described by the flowchart in the centre.

nitscrape https://github.com/Merlotec/nitscrape
tsa https://github.com/Merlotec/tsa

Tweet Download Method Comparison	Price per million Tweets downloaded (USD)	Max Download Rate (Tweets per minute)
X API	5000	6000
Simple Nitter Scraper	0	180
Nitter Scraper with Distributed Tor Circuits	0	12000



Correlation Between Sentiment of Brexit Tweets and Brexit Polls

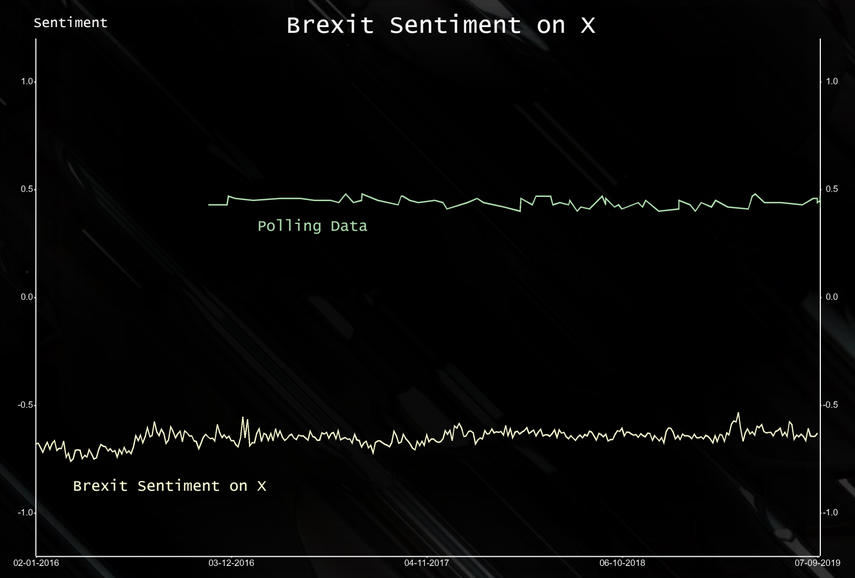
The secondary aim of this project is to carry out large scale language analysis and automated sentiment analysis of Tweets about Brexit using Natural Language Processing AI. The topic of Brexit is interesting in its own right and correlations that exist between Brexit Tweets and public opinion may also exist for other topics.

The sentiment analysis computer program automatically assigns a sentiment score to each Tweet based on how positive/negative the language is. An average sentiment value for each period is taken to get a combined sentiment value for that period.

Unfortunately there was no statistically significant correlation found between sentiment scores in Tweets and the national polling figures on Brexit opinion - see graph below. **The correlation coefficient was (-0.1980) at a significance level of 0.6704.**

Keyword Analysis of Brexit Tweets

Keyword analysis was also done throughout different time periods, which determined the most commonly used topic related words (ignoring common/neutral words). This delivered a good insight into what Brexit topics were discussed on X at various times. Further research could be done to investigate how changes is these trending key words relate to topics in the legacy media.



UNIVERSITY OF CAMBRIDGE

LAIDLAW FOUNDATION