Text Mining - Fake/True News Articles Analysis

Juan Picciotti

April 2022

Introduction

The term ‘Fake news’ was popularised by Donald Trump in 2016, and is defined as journalism that contains incorrect or misleading information.

The rise in technology and social media has hugely increased the amount and diversity of news sources that people have access to, with around 62% of adults in the US consuming news from social media in 2016, relative to 49% in 2014 (Shu et al, 2017). In addition, the younger generation tends to rely on social media to educate themselves on political news and world events (Rubin, 2017).

While access to greater amounts of information is clearly beneficial, this increases the potential for the propagation of misinformation if the sources are not legitimate, which can have highly negative real world consequences. For example, misinformation regarding Covid-19 was a direct cause of lower vaccination rates for certain segments of the population.
To counteract this, identifying fake news is now a great cause of focus for governments and institutions (Yale Law School, 2017).

Detection of fake news is a highly complex task due to the fact that it is created with the intention of deceiving the reader. The result of this is that people are almost no better than random at detecting fake news articles (Zhou 2020, Wang 2017). In addition, the advancement of AI techniques means that ‘bots’ can create huge quantities of highly realistic content for a low cost to the actor spreading the fake news (Stahl, 2018), meaning that it is almost impossible for a human fact checker to solve this detection problem.

Hence, identifying fake news through computational means is a highly relevant and important topic. In this paper we first discuss and seek to understand and identify differences between fake and true news articles, and then we build on the techniques used in other literature to produce a fake news detection classification model.

Datasets

The dataset we use for our analysis and to train the classification model is a collection of US political news articles, sourced from Kaggle (2017). The 'true' news data was originally sourced from trusted news website Reuters, and the 'fake' news data was sourced from Politifact, a website that aims to fact check news and identify misinformation. The dataset contains 21,192 and 22,851 true and fake news articles and their titles respectively, with publishing dates ranging from January 2016 to December 2017.

Topic Modelling

Topic modelling is a method for unsupervised classification of documents, which allows for deriving insights from text corpora. In this section we describe how we used one of the models that can be employed for this task - Latent Dirichlet Allocation - in our corpus and what results we obtained.

Theory

Latent Dirichlet Allocation, first introduced by Blei et al. (2003) in their famous paper published in the Journal of machine Learning research, in 2003, to text analysis, is based on the idea that documents are represented as a random mixture over latent topics.

Each topic is characterised as a distribution over words. LDA assumes that words of each document arise from a variety of topics where each topic is a Multinomial over a fixed vocabulary.

The topics are conveyed by all documents in the collection, as they are randomly drawn from a Dirichlet distribution; hence topic proportions vary across documents stochastically (Blei, 2007)

In mixed membership models, each document (indexed by m) is assumed to generate as follows:

- First a distribution over topics, θ_m, is drawn from a global prior distribution. In our case, uniformly distributed.
- Then, for each word in the document (indexed by n), we draw a topic for that word from a Multinomial distribution based on its distribution over topics ($z_{n,m}$ ~ Mult(θ_m)). Conditional on the topic selected, the observed word $w_{m,n}$ is drawn from a distribution over the vocabulary $w_{m,n}$ ~ Mult($\beta_{zm,n}$), where $\beta_{k,v}$ is the probability of drawing the v-th word in the vocabulary for topic k. LDA, assumes a Dirichlet prior for the topic proportion such that β_m~Dirichlet(α).

Although LDA is a powerful tool to analyse text, it makes some restrictive assumptions. Firstly, it assumes that topics within a document are independent of one another; these come from

assuming a Dirichlet distribution on the topics for a document. Secondly, the distribution of words within a topic is stationary, meaning that words are exchangeable within each document. Thirdly, topics are modelled entirely based on the text of the document.

Parts of Speech

In order to build a Topic Model, as for other applications of text analysis, after the pre-processing of the Documents, these are split into tokens (words and other terms) and a portion of them is removed due to carrying little value in terms of defining the underlying topics. Typically, predefined dictionaries of stop-words (tokens with small impact on topics) are used, in addition to manually generated data-specific collections of tokens.

Our strategy builds on this general approach by pre-specifying each token's role in the sentences they are part of. This classification into 'Parts of Speech (POS)' - syntactical tags: nouns, verbs, adjectives, etc. - allows for the removal of entire classes of tokens which offer little information for defining topics.

In our case, given the number of newspaper articles in our corpora (21,000+ true news, and 22,000+ fake news) and their lengths, the numbers of tokens are over 9 million (in true news), and 10.8 million (for fake news).

In order to make the analysis computationally cheaper, as well as removing stopwords, punctuation, numbers and symbols we also remove verbs, ad-positional phrases, determiners, pronouns, auxiliaries, adverbs, coordinating and subordinating conjunctions.

Specifying topics can be efficiently done by focusing on the people, places and events - public figures, countries, elections, pandemics, scandals, etc. - that are mentioned in the documents. This information is typically represented in the articles' nouns.

Despite adjectives usually being included for sentiment analysis and, arguably, could be disregarded for topic modelling, in our case they offer added value when considering n-grams (groups of n tokens that are found together). In our context, for instance, 'White' and 'House' alone have very different meanings from 'White House'.

Hence, for our analysis, we removed all POS except Nouns and Adjectives. As a result, almost 7 million tokens are removed from the fake news (leaving only 36.2% of the original quantity), and 6.5 million from the true news (leaving 39.6%).

For this purpose, we employed Python's NLTK tagger function, which has many different and precise categories for Parts of Speech, including many types of nouns, verb tenses, etc. In order to summarise the data, we grouped the POS into fewer sets (see Fig. 5).

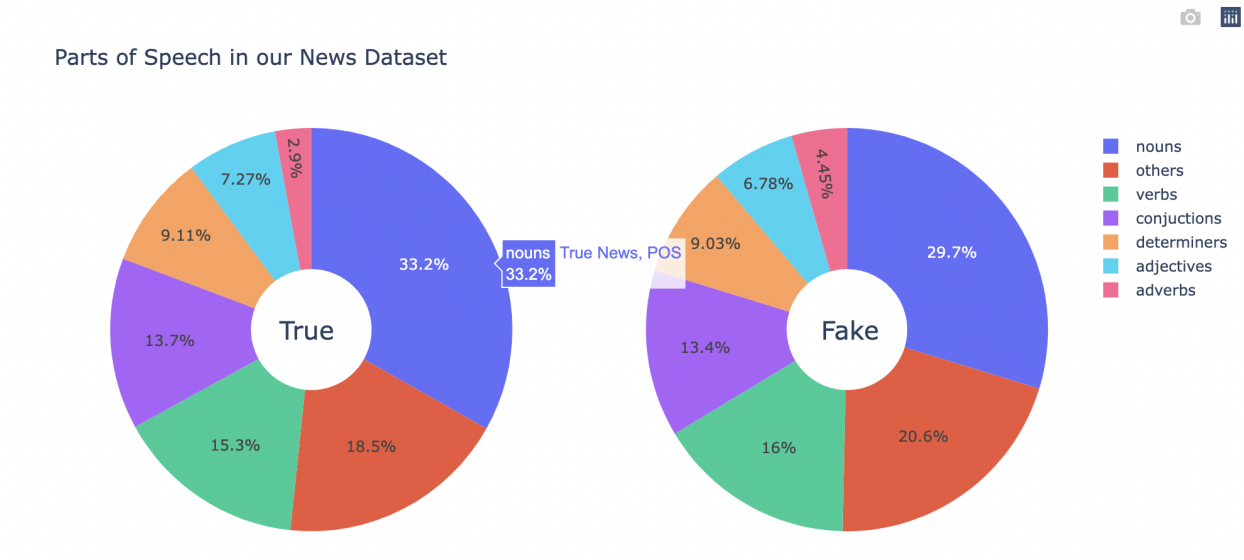
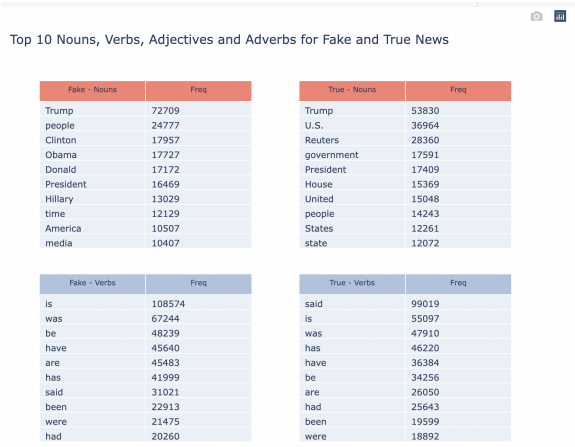


Fig. 5: Proportion of word types per dataset. Note: this is an interactive chart that can be fully viewed in the associated HTML file.

A dictionary based approach could be considered, where all nouns and adjectives are saved into a list, and then any words not present in the list filtered out.

However, sometimes the exact same word plays different roles in a sentence: for instance, "He gave him the *look*" and "You *look* nice". *Look* acts as a noun in the first case, and as a verb in the second. Hence, we take the approach of filtering out tokens considering their roles in the sentences instead of the words themselves.



Top 10 Nouns, Verbs, Adjectives and Adverbs for Fake and True News

Fake - Nouns	Freq
Trump	72709
people	24777
Clinton	17957
Obama	17727
Donald	17172
President	16469
Hillary	13029
time	12129
America	10507
media	10407

True - Nouns	Freq
Trump	53830
U.S.	36964
Reuters	28360
government	17591
President	17409
House	15369
United	15048
people	14243
States	12261
state	12072

Fake - Verbs	Freq
is	108574
was	67244
be	48239
have	45640
are	45483
has	41999
said	31021
been	22913
were	21475
had	20260

True - Verbs	Freq
said	99019
is	55097
was	47910
has	46220
have	36384
be	34256
are	26050
had	25643
been	19599
were	18892

Table 2 - Top 10 Nouns and Verbs in each corpus

Next, we perform some additional general preprocessing, which includes lowercasing, stemming and other basic tasks.

LDA Model

We use the gensim library to construct our models, one for each corpus. The three main arguments gensim's LDA model requires are the following:

- The preprocessed corpus
- The associated dictionary
- The number of topics

We create bigrams and trigrams with gensim's native function, modifying the arguments 'min_count' and 'threshold'. The higher their values, the harder it is for words to be combined into bi/trigrams. We additionally create the associated dictionary employing the library's tools.

As for defining the number of topics we want to work with, this task can be done by computing different diagnostic values (e.g. Held-Out Likelihoods, Semantic Coherence, Residuals methods) for a varying number of topics, and then choosing an "optimal" *k*. In our case, we simply intuitively analyse results for values of *k* between 8 and 12, and find that 8 produces the most clear and interpretable topics.

Additional important arguments for gensim's LDA model are:

- alpha: A-priori belief on document-topic distribution.
- eta: A-priori belief on topic-word distribution.

Since we have no prior insights on how topics or words within topics could be distributed, we did not input these.

Topic Modelling Results

We manually label the resulting topics as can be seen in Table x. Note that we also create an in-detail interactive dashboard that allows for more in-detail of the topics. See the attached HTML file to access the dashboard, and Appendix A.2 for instructions on how it can be used and interpreted.

Fake	True
Police shootings	UK elections
US elections between Clinton and Trump	US elections
Political affairs in the US	Puerto Rico State
Syrian war	Refugees
Trump	Industry - Regulation in the US
FBI investigation on emails from Clinton	Russian interference in US Elections

Fake	True
Media depiction of politics in the US	Iran - Kurdish
Racism in the US	China, North Korea and Foreign Affairs

Table 3 - Manually labelled topics produced from LDA analysis

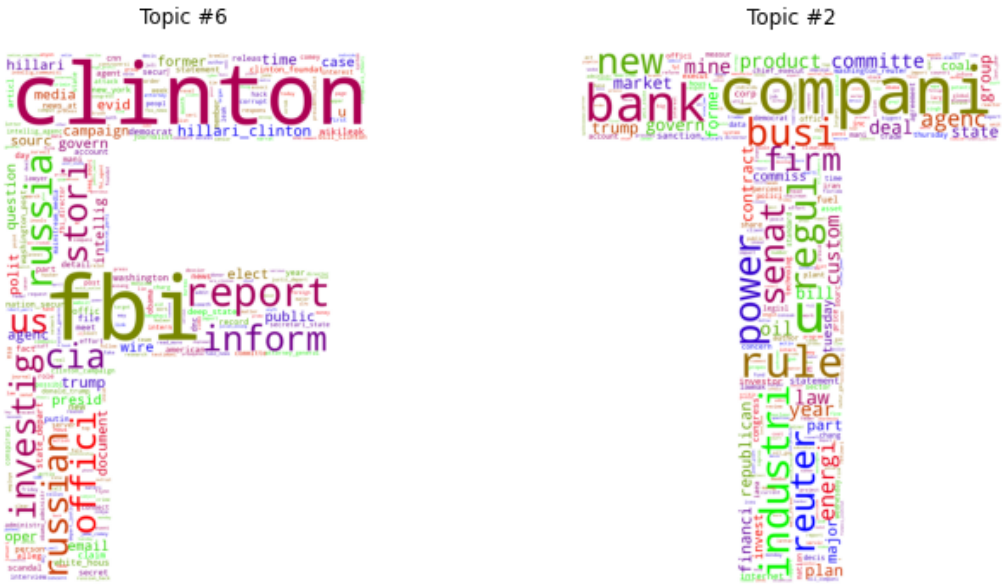


Fig. 6 - Word-clouds depicting terms usage in different topics in Fake and True news. Note: Please see notebook for all 16 topics (8 from each corpora).

As an example, we plot word clouds (Fig. 6) to observe the most frequent words for the ‘FBI investigation on emails from Clinton’ topic for the fake news and ‘Industry - Regulation in the US’ topic for the true news.

We see some overlap of the topics for the fake news and the true news articles, for example the ‘US elections’ topic for true and the ‘US elections between Trump and Clinton’ are clearly very similar. In addition, we can consider that the topics ‘Refugees’ for the true news and ‘Syrian war’ for fake news are overlapping, as well as ‘Industry - Regulation in the US’ for true news and ‘Trump’ for the fake news.

However, we see that some themes are more unique to each of the datasets. For instance the ‘Iran-Kurdish’, and ‘China, Korea and Foreign affairs’ themes for the true articles, and ‘Racism in the US’ for the fake articles. These results consolidate some of the initial findings in the earlier more basic polarisation analysis.

Furthermore, the dashboard also allows us to see that there is a sharp clustering in the case of true news, where most of the topics within the true news dataset are evenly distributed and separated. As for fake news, they are less separated and few topics dominate the distribution with “Police Shootings” and “Trump” taking 40% of the share.

It should be noted that these conclusions are stochastic in nature, therefore could vary slightly with new runs.

Conclusions

Some topics overlap both the fake and the true datasets (specifically around the US elections), whereas others remain specific to the true news (e.g. foreign affairs) and specific to the fake news (e.g. racism).

We also observe that topics within the true news are evenly distributed and separated, whereas for the fake news, they are less separated and few topics dominate the distribution. Bibliography

Adfonts Media, 2022. *Interactive Media Bias Chart.* [Online]. [Accessed 29 March 2022]. Available from: https://adfontesmedia.com/interactive-media-bias-chart/

Ahmed, S., et al., 2020. Development of fake news model using machine learning through natural language processing. Available from: https://arxiv.org/pdf/2201.07489.pdf

D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003. Available from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com

D. M. Blei and J. D. Lafferty, “A correlated topic model of science,” The annals of applied statistics, vol. 1, no. 1, pp. 17–35, 2007. Available from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-1/A-correlated-topic-model-of-Science/10.1214/07-AOAS114.pdf

Jin, Z., et al, 2017. "Novel Visual and Statistical Image Features for Microblogs News Verification," in IEEE Transactions on Multimedia, vol. 19, no. 3, pp. 598-608. Available from: https://ieeexplore.ieee.org/document/7589045

Kaggle. 2017. *Fake and Real News Dataset*. [Online]. [Accessed 23 March 2022]. Available from: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset?select=True.csv

Mikolov, T., et al., 2013. Efficient estimation of word representations in vector space. Available from: https://arxiv.org/pdf/1301.3781.pdf

Reddy H. et al. (2020). Text-mining-based Fake News Detection Using Ensemble Methods. International Journal of Automation and Computing, vol. 17, no. 2, pp. 210-221. Available from: https://www.mi-research.net/en/article/doi/10.1007/s11633-019-1216-5

Rubin, V.L., 2017. Deception detection and rumor debunking for social media. In *The SAGE handbook of social media research methods* (p. 342). Sage. Available from: https://core.ac.uk/download/pdf/61692768.pdf

Shu, K. et al, 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), pp.22-36. Available from: https://arxiv.org/pdf/1708.01967.pdf

Shrestha, M., 2018. Detecting Fake News with Sentiment Analysis and Network Metadata. *Earlham college, Richmond*. Available from: https://portfolios.cs.earlham.edu/wp-content/uploads/2018/12/Fake_News_Capstone.pdf

Stahl, K., 2018. Fake news detection in social media. *California State University Stanislaus*, *6*, pp.4-15. Available from: https://www.csustan.edu/sites/default/files/groups/University%20Honors%20Program/Journals/02_stahl.pdf

Wang W. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. 422-426. Available from: https://www.researchgate.net/publication/318740546_Liar_Liar_Pants_on_Fire_A_New_Benchmark_Dataset_for_Fake_News_Detection

Yang, Y. et al., 2018. TI-CNN: Convolutional neural networks for fake news detection. Available from: https://arxiv.org/pdf/1806.00749.pdf

Yale Law School. 2017. *Fighting Fake News Workshop Report.* [Online]. [Accessed 30 March 2022]. Available from: https://law.yale.edu/sites/default/files/area/center/isp/documents/fighting_fake_news_-_workshop_report.pdf

Zhou, X. et al., 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-40. Available from: https://arxiv.org/pdf/1812.00315.pdf

Appendix

Interactive dashboard instructions:

This dashboard is based on the pyLDAvis library, and can be interpreted as:

Left side panel:

- It offers a representation of how topics are distributed in the 2-dimensional space (based on PCA). The larger the bubble, the more frequent the topic in the documents is.

  Typically, topic models with low numbers of topics have large bubbles than tend to not overlap. Topic models with high numbers of topics, on the other hand, have many overlapping small size bubbles.

- Intertopic Distance is an approximation of semantic relationship between the topics. Those that share many terms will be closer and tend to overlap.

Horizontal Bar Graph:

- The blue bars in the graph show the frequency distribution of the words in all of the documents.
- The red portion of the bars describe the frequency of each word, given a topic.

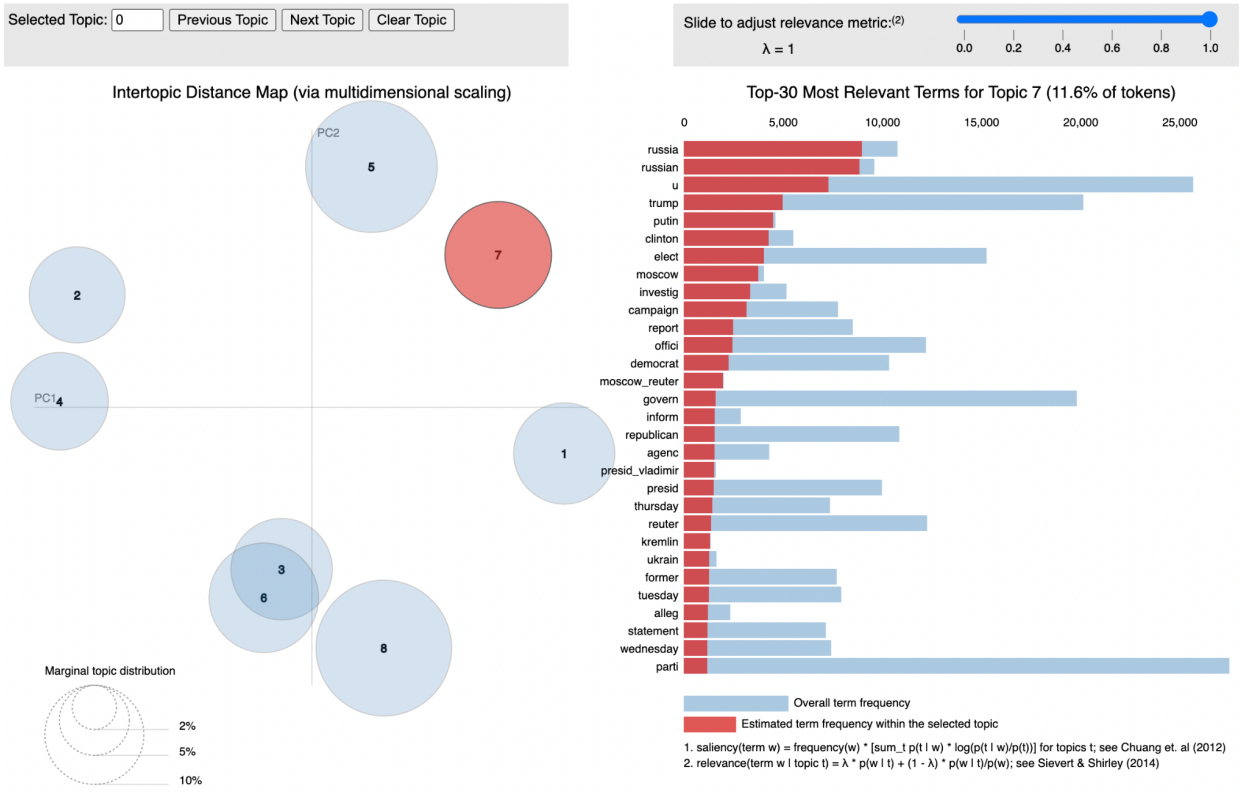
Interactivity:

- You can select a topic by clicking on a topic bubble. This will show the top 30 words associated with it, and their proportion in terms of total frequency.

- Hovering over the specific words, the topics containing those words are highlighted. The higher the proportion of that word in a topic, the larger the size of the bubble.

- You can play with the lambda parameter to re-rank words in topics based on their frequency. Decreasing the lambda parameter, increases the weight of the ratio of the frequency of word given the topic / Overall frequency of the word in the documents.

We attach two html files (one for each topic model) with the fully functional dashboards for in detail analysis of the topics.



Interactive visualisations of how the topics (tokens) are distributed in the corpora (topics). Please see html files to interactively explore the results of our LDA models