Project Notes:

Project Title: Detecting Fake News Using a Machine Learning Model Based on Lexical Characteristics of Text
Name: Anne Wu

Note Well: There are NO SHORT-cuts to reading journal articles and taking notes from them. Comprehension is paramount. You will most likely need to read it several times so set aside enough time in your schedule.

Contents:

Knowledge Gaps:

This list provides a brief overview of the major knowledge gaps for this project, how they were resolved and where to find the information.

Knowledge Gap	Resolved By	Information is located	Date resolved
How do I develop an AI?	9/6/22: knowledge gap created	n/a (too broad, will make more specific knowledge gap)	10/14/22
How can I create an algorithm that simulates a network? (not sure if i will utilize)	9/28/22: knowledge gap created	n/a (not going in this direction)	12/1/22
How do I implement a machine learning algorithm in Python?	10/14/22: knowledge gap created	Article 19 (though I will still learn more as time progresses)	12/12/22
What machine learning algorithm will be most effective for my project?	Directly goes with previous knowledge gap 10/14/22: knowledge gap created	Article 15 (though this question is much more broad and still actively researched; could be a direction for my project)	12/12/22
What is a NLP (Natural Language Processor)? What type of NLP will I need to develop in order to achieve my goal?	10/15/22: knowledge gap created	Article 14	11/1/22
How can I utilize statistics for my project?	12/12/22: knowledge gap created		

Literature Search Parameters:

These searches were performed between 09/02/2022 and XX/XX/2022.
List of keywords and databases used during this project.

Database/search engine	Keywords	Summary of search
Gordon Library Database	Misinformation artificial intelligence	- https://wpi.primo.exlibrisgroup.com/permalink/01WPI_INST/1pchs3f/cdi_scopus_primary_2010526909 (Article 5 in TOC)
Gordon Library Database	misinformation news reasons	- Article 8 in TOC
Google Patents	Misinformation fake news	- Grants 1 and 2 in TOC
Gale OneFile: Psychology	misinformation fake news	Got nothing, seems like results were more on the lines of general deception instead of misinformation online
The New York Times	misinformation fake news online	- Found an article on the Gale OneFile database for the New York Times - Ended up searching the actual New York Times website for the same article since the article had videos that the database couldn't display - Article 13 in TOC

Article #1 Notes: Template

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	
Source citation (APA Format)	VOLKOVA, S. (2021). *Prediction of social media postings as trusted news or as types of suspicious news* (United States Patent No. US11074500B2). https://patents.google.com/patent/US11074500B2/en?q=misinformation+fake+news&oq=misinformation+fake+news
Original URL	https://dl.acm.org/doi/pdf/10.1145/3308560.3316739
Source type	
Keywords	
Summary of key points + notes (include methodology)	
Research Question/Problem/ Need	
Important Figures	
VOCAB: (w/definition)	
Cited references to follow up on	
Follow up Questions	

Article #1 Notes: These 5 Social Media Habits Are Linked with Depression

Article notes should be on separate sheets

Source Title	These 5 Social Media Habits Are Linked with Depression
Source citation (APA Format)	Rettner, R. (2018, June 1). *These 5 Social Media Habits Are Linked with Depression*. Livescience.Com. https://www.livescience.com/62718-social-media-habits-depression.html
Original URL	https://www.livescience.com/62718-social-media-habits-depression.html
Source type	Website
Keywords	Social Media, Depression, Mental Health
Summary of key points + notes (include methodology)	A study analyzed information about 500 undergraduate students that regularly used various social media sites. How people used social media connected to depression, since people with depression exhibited different behaviors on social media. However, this is only an association, so it doesn't mean social media causes depression.
Research Question/Problem/ Need	Does social media use connect to symptoms of depression?
Important Figures	n/a
VOCAB: (w/definition)	Undergraduate: student at college who has not yet earned a degree.
Cited references to follow up on	https://www.livescience.com/34718-depression-treatment-psychotherapy-anti-depressants.html (Webpage that details information about depression) https://www.livescience.com/61996-personality-social-media-addiction.html (details on Social Media Addiction) https://www.livescience.com/18324-facebook-depression-social-comparison.html (Facebook use's connect to depressive symptoms) https://www.livescience.com/58121-social-media-use-perceived-isolation.html (social media use can lead to perceived isolation)

	https://www.livescience.com/52148-social-media-teen-sleep-anxiety.html (Social media use effect on teens) https://about.fb.com/news/2017/12/hard-questions-is-spending-time-on-social-media-bad-for-us/ (Facebook's blog post about social media's impact on people)
Follow up Questions	Do people with depression gravitate towards using social media as opposed to interacting in real life? Would there be any significant changes if people with different backgrounds and in different age groups were included in a similar study? What external factors outside of social media contribute to depression? Would those factors correlate to social media use？

Article #2 Notes: Artificial Intelligence Learns to Learn Entirely on Its Own

Article notes should be on separate sheets

Source Title	Artificial Intelligence Learns to Learn Entirely on Its Own
Source citation (APA Format)	Hartnett, K. (2017, October 18). *Artificial Intelligence Learns to Learn Entirely on Its Own*. Quanta Magazine. Retrieved August 18, 2022, from https://www.quantamagazine.org/artificial-intelligence-learns-to-learn-entirely-on-its-own-20171018/
Original URL	https://www.quantamagazine.org/artificial-intelligence-learns-to-learn-entirely-on-its-own-20171018/
Source type	Website
Keywords	Artificial Intelligence, Go, tree search
Summary of key points + notes (include methodology)	A computer program called AlphaGo Zero, which initially only knew about the rules of Go, was able to develop various strategies to play Go just by playing against itself multiple times for 3 days. The primary algorithm that powers this program is the "tree search", letting the program look ahead and preview the various moves that can be made and their results. AlphaGo Zero in particular is able to remember the outcomes of the search and utilize them in future games.
Research Question/Problem/ Need	How is artificial intelligence able to learn how to improve itself in games like Go?
Important Figures	n/a
VOCAB: (w/definition)	n/a
Cited references to follow up on	http://web.eecs.umich.edu/~baveja/ (Satinder Singh's website. Was not involved in the research but could provide more info on AI) https://www.usgo.org/who-aga (Website for the American Go Association)

	https://www.nature.com/articles/nature24270 (paper that goes into detail about AlphaGo Zero, but under a paywall) https://www.quantamagazine.org/is-alphago-really-such-a-big-deal-20160329/ (Article that details the original AlphaGo)
Follow up Questions	What are some other examples of programs that utilize the “tree search” algorithm? What benefits does a machine playing against itself have with programs outside of strategy games? (science, medicine, etc.) Is this program based on a neural network? What are the advantages of using a neural network?

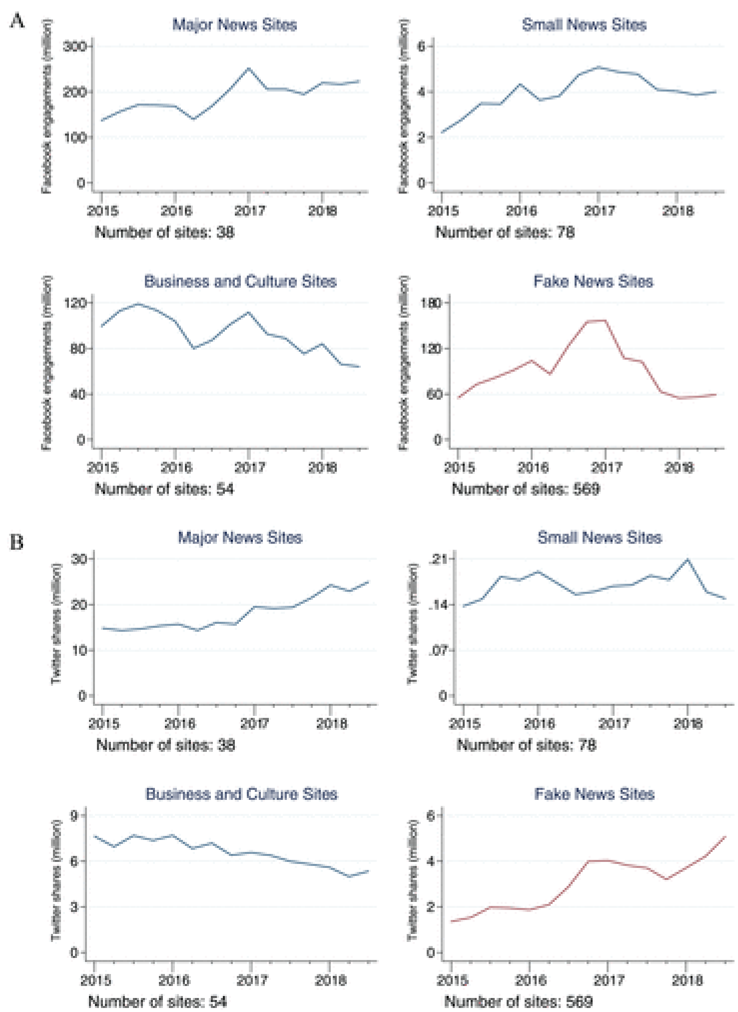
Article #3 Notes: Trends in the diffusion of misinformation on social media

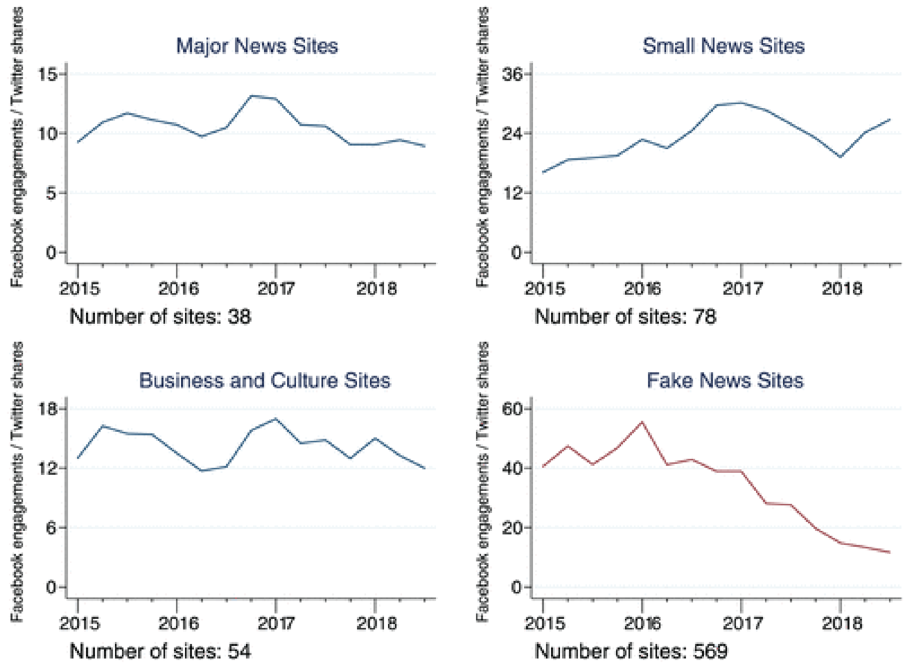
Article notes should be on separate sheets

Source Title	Trends in the diffusion of misinformation on social media
Source citation (APA Format)	Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics, 6.* https://doi.org/10.1177/2053168019848554
Original URL	https://doi.org/10.1177/2053168019848554
Source type	Research Article
Keywords	Social Media, Misinformation, Fake News, Facebook, Twitter, False Content

Summary of key points + notes (include methodology)	The spread of misinformation is a big problem on the internet, exacerbated by the fact that social media makes it incredibly easy to share that information. The study tracked the interactions on Facebook and the shares on Twitter of articles on a number of fake news websites. The trends show that while the rate of engagements were relatively stable for other sites, spread of fake news was far more inconsistent and grew around the time of the 2016 election. For Facebook, the spread declined after that while for Twitter, the spread continued to increase.
	—
	- Fake news may have played large role in 2016 election and its resulting political divisions
	- Evidence of how serious the misinformation problem is is limited
	- False stories still seem to be a problem on Facebook even after its news algorithm had been altered
	- Efforts to fight misinfo are “not working” and its “becoming unstoppable”
	- Data collection method:
	    - Find misinformation spread on social media from Jan 2015 to Jul 2018
	    - Used 569 sites identities as fake news sites
	        - Fake news site: sites identified as sources of false stories in 5 studies or online lists
	    - The Facebook and Twitter shares of these sites from BuzzSumo
	        - BuzzSumo tracks the amount of user interactions with web content for various social media sites
	    - To compare, these sites were also measured
	        - Major news sites

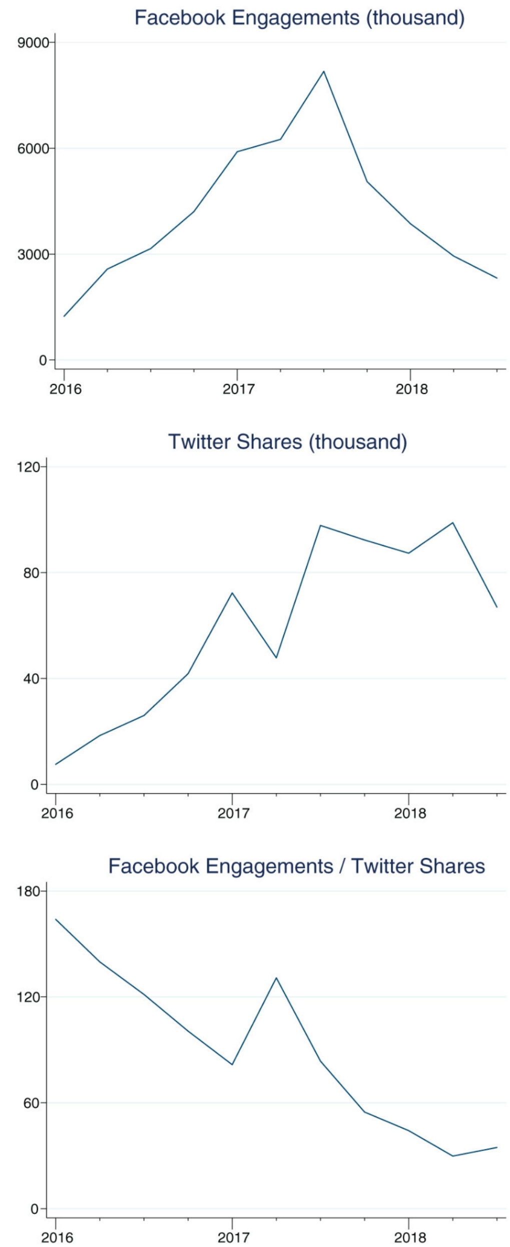
        - Small news sites that don't provide misinformation
        - Business and culture sites
- Results
- Data:
    - Collected sites might be weighted towards misinformation that Facebook is aware of instead of the opposite
    - List most likely excludes many small sites or sites that were only active for a short period of time
    - Mainly comprised of sites with a major US audience
    - Facebook engagements and twitter shares are not directly comparable
        - They are summed and then the **average by quarter** is found
    - BuzzSumo data is found for 569 of the 672 fake news sites that met the criteria and all of the other sites used for comparison
        - The sites that weren't found were small and the vast majority were inactive by 7/21/2018
- Results
    - Interacts for major news sites, small news sites, and business and culture sites remained relatively stable and were similar for both facebook and twitter
    - Interaction with fake news sites changed a lot and had very different trends on the two platforms
        - Sharp decrease after 2016 election for facebook
        - Continued to increase after 2016 election for twitter
    - Ratio of Facebook engagements to Twitter shares:
        - Stable for major news, small news, and business and culture sites
        - Sharp decline for fake news from 45:1 during election to 15:1 two years later
            - Though suggests that spread of misinfo has decline on Facebook, it's important to note how large the quantity for both twitter and facebook is, even if facebook takes up the vast majority
- Interpretation of data:
    - Overall conclusion is that the spread of misinformation has declined, but it has not stopped
    - Facebook still plays a big role in the diffusion of misinformation, even after algorithm changes
    - Database for false stories far from complete, even though attempted to be made as comprehensive as possible
    - Declines could be due to undersampling
    - The trends of the spread of fake news on Facebook and Twitter are relatively similar up until after the 2016 election. Twitter kept increasing, Facebook declined

	- Maybe indicates some change in Facebook way of processing misinfo that Twitter didn't have - BuzzSumo data has the chance of being inaccurate since information is not individually verifiable - The few sites that BuzzSumo didn't have data on also could have been a factor
Research Question/Problem/ Need	What are the trends of the spread of misinformation on social media?
Important Figures	 **A: Depicts Facebook engagements** - **Relatively Steady for News and Business and Culture sites** - **Fake news engagements peaked at around 2016 election, has decreased since** **B: Depicts Twitter Shares** - **Relatively Steady for News and Business and Culture sites** - **Had a peak at around 2016 election, has continued to rise**



Depicts ratio between facebook engagements/twitter shares

- **I honestly don't know how useful this metric is**
- **Claims to be the main part of the data, but there are so many ways to interpret it that I don't feel that it's reliable.**
- **Facebook engagements and Twitter shares are also not directly comparable since Facebook Engagements is a much more broad category.**

<table>
<tr><td></td><td>

Depicts the facebook engagements and twitter shares of 9540 urls that spread misinformation</td></tr>
<tr><td>VOCAB: (w/definition)</td><td>Scrape: copy data from a website using a computer program
Diffusion: the spreading of something more widely
Caveats: a modifying or cautionary detail to be considered when evaluating, interpreting, or doing something
Average by Quarter: the mean of values taken during a calendar quarter (3 months)</td></tr>
<tr><td>Cited references to follow up on</td><td>https://journals.sagepub.com/doi/suppl/10.1177/2053168019848554/suppl_file/appendix.pdf (article’s appendix)</td></tr>
</table>

	https://time.com/5112847/facebook-fake-news-unstoppable/ (Time article saying that the spread of misinformation is “unstoppable”) Allcott, H, Gentzkow, M (2017) Social media and fake news in the 2016 election. Journal of Economic Perspectives 31(2): 211–236. Lazer, DM, Baum, MA, Benkler, Y, et al. (2018) The science of fake news. Science 359(6380): 1094–1096.
Follow up Questions	Why are current ways to prevent the spread of misinformation, like Facebook’s news algorithm, ineffective? This article was written before the 2020 election, so was there any prevalent uptick in fake news sharing around that time? What are the motivations for creating misinformation, and are all motivations malicious? Most of that data collection was done manually, so are there any effective methods of collecting this information that are more automated without losing accuracy?

Article #4 Notes: MISINFORMATION and CONSPIRACY THEORIES about the COVID-19 VACCINES

Article notes should be on separate sheets

Source Title	MISINFORMATION and CONSPIRACY THEORIES about the COVID-19 VACCINES have spread across SOCIAL MEDIA, infiltrating the sunny world of WELLNESS INFLUENCERS at a time when the STAKES COULDN'T BE HIGHER.
Source citation (APA Format)	Phelan, H. (2021, April). MISINFORMATION and CONSPIRACY THEORIES about the COVID-19 VACCINES have spread across SOCIAL MEDIA, infiltrating the sunny world of WELLNESS INFLUENCERS at a time when the STAKES COULDN'T BE HIGHER. *Harper's Bazaar*, (3691), 130+. https://link.gale.com/apps/doc/A659005242/PPOP?u=mlin_c_worpoly&sid=bookmark-PPOP&xid=6c5428cb
Original URL	https://link.gale.com/apps/doc/A659005242/PPOP?u=mlin_c_worpoly&sid=bookmark-PPOP&xid=6c5428cb
Source type	Magazine Article
Keywords	COVID-19, vaccine, misinformation, social media, influencers

Summary of key points + notes (include methodology)	This article details the reasons why false information about the COVID-19 vaccine has been so prevalent. It mostly attributes the spread to social media influencers that are able to utilize the ignorance of their audience to spread their erroneous beliefs. Anti-vaxxers already existed before COVID-19 existed, but the general doubts people had about the vaccine served to make their voices louder.
	- Starts with description of a social media account that promises quick fixes to problems without proof and how people are drawn to it
	- For herd immunity, 85% of people need to get a vaccine, but only 60% of Americans are planning to get it
	- Why people skeptical of COVID-19 vaccine (at first):
	    - Was fastest developed vaccine
	    - Heavily politicized
	    - In POC communities, systemic racism made them doubt medical establishments
	    - However, there was still resistance after studies supporting the safety and effectiveness of these

	vaccines…
	- The discussion about COVID-19 vaccine has made the voices of anti-vaxxers louder
	    - Most-followed social media accounts of anti-vaxxers increased following by 7.8 million since 2019
	    - Two anti-vax books in top 5 results of “vaccine” on Amazon
	    - Russian bots were used to spread misinfo and anti-vax messaging from 2014 to 2017
	    - Almost half of all twitter accounts spreading misinfo were bots possibly deployed by China and Russia
	- Dr. Christiane Northrup
	    - “den mother to the New Age and anti-vaxx communities.”
	    - Certified OB/GYN
	    - Said that the COVID-19 vaccines would lower people’s enlightenment
	    - Videos contain her ASMR voice and occasional harp playing
	    - Advocates for QAnon, an american far-right political movement revolving around false claims
	    - Has published a book which contains both sound medical advice and nonsense pertaining to “Shamanic Imprint Removal” and false anti-vax info
	        - Section titled “Vaccines: Helpful or Harmful” says for reader to decide for themself while also giving false info about the dangers of vaccines
	            - One reader like that she wasn’t trying to “jam a message down my throat” like doctors would
	- QAnon and anti-vax campaigns seem innocuous at first, encourage to “do your own research”
	    - However, one cannot rely on their own intuition on data, they need to be experienced in the field and peer review
	- People go to people that they trust to get info about something they don’t know. This is why influencers are so effective
	- Anti-vaxxers often emphasize personal responsibility as opposed to protecting others
	- Though there are many influencers that spread misinfo, there are also those that combat this misinfo.

Research Question/Problem/ Need	What are the causes and effects of the spread of misinformation pertaining to the COVID-19 vaccine? How can we push back against this misinformation?

Important Figures	n/a
VOCAB: (w/definition)	**Holistic:** “characterized by the treatment of the whole person, taking into account mental and social factors, rather than just the symptoms of a disease” **Insidious:** “proceeding in a gradual, subtle way, but with harmful effects” **Rhetoric:** “the art of effective or persuasive speaking or writing, especially the use of figures of speech and other compositional techniques” **Geopolitical:** ”relating to politics, especially international relations, as influenced by geographical factors.” **Osteopath:** “a licensed physician who aims to improve people’s overall health and wellness by treating the whole person, not just a condition or disease they may have.” **Pernicious:** “having a harmful effect, especially in a gradual or subtle way.” **OB/GYN:** Doctor who specializes in women’s health
Cited references to follow up on	“A 2018 study out of George Washington University found that Russian bots were instrumental in fueling the online debate around vaccines between 2014 and 2017, uncovering thousands of Twitter accounts that had been used to spread misinformation and anti-vaccine messaging in the U.S.” Find this study somehow “In late April, researchers at Carnegie Mellon University found that nearly half of all Twitter accounts tweeting misinformation about the coronavirus were likely bots deployed, they hypothesized but could not substantiate, by China or Russia.”
Follow up Questions	Why would countries like China and Russia want to spread misinformation in other countries besides itself? In what ways can misinformation affect the political landscape of a country like the US? What is the purpose of influencers spreading misinformation about COVID-19? Is it just preying on the ignorance of others to gain money? How has the anti-vax movement developed over time?

Article #5 Notes: Contrasting the Spread of Misinformation in Online Social Networks

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Contrasting the Spread of Misinformation in Online Social Networks
Source citation (APA Format)	Amoruso, M., Anello, D., Auletta, V., Cerulli, R., Ferraioli, D., & Raiconi, A. (2020). Contrasting the Spread of Misinformation in Online Social Networks. *The Journal of Artificial Intelligence Research, 69,* 847-879. https://doi.org/10.1613/jair.1.11509
Original URL	https://doi.org/10.1613/jair.1.11509
Source type	Journal Article
Keywords	Social Networks, False Information, News, Algorithms, Public Safety

Summary of key points + notes (include methodology)	This paper's main goal is to do two things: create algorithms that can identify sources of misinformation and place monitors that are able to block this misinformation. To do this, accounts in a social network are treated as a network of nodes. Infected nodes, as in accounts that can spread misinfo, have a 0 to 1 chance of infecting other nodes. The algorithm that they developed seems to be more effective than previous solutions and has a very quick performance speed.
	—
	1 Introduction
	- Many people use social media in their day to day lives
	    - Allows ease of communication and creates bonds of trust
	- How users interact with each other can cause contect to go viral (go into a vast audience)
	- However can social media also cause spread of inaccurate or completely false information
	    - Can be by mistake or with malicious intent
	    - Ex. attract a specific niche for ad revenue or to change public opinion
	    -
	- Results of spread of misinfo can be undesirable or very dangerous
	    - Misinfo about vaccines -> refuse vaccines for children which lowers herd immunity
	    - Misinfo about Ebola on Twitter 2014 -> overall panic and spread of harmful medical advice

	- Misinfo about COVID-19 - Political misinfo -> influence in voting decisions - Also can cause unstable financial markets

- 3 steps for fighting misinfo (paper focused on last two points):
    - Recognize misinfo
    - **Identify sources**
        - Understand goals for spread
        - Know who sources are
        - Allows for further action for 3rd step
        - But for many cases not possible to 100% identify source
            - Find list of "suspects"
    - **Limit ability for further spread**
- Placing monitors on users, both suspects and normal users
- Monitored users must consent to it or just respond to reports of detected misinfo by users without monitors

1.1 Our Contribution

- Used a directed weighted graph and independent cascade model
- Each node represents a user
- If they have been exposed to misinfo, the node is considered "infected" and have a chance to spread to neighbouring nodes
- Independent Cascade model and epidemics models have been used to model general spread of info on social media
- Deliberately created false info news tend to be more novel and get more emotional reactions, which results in more shares

1.2 Related Works

- Source identification
    - Treat the spread of misinfo like an infectious disease (simple epidemic models)
        - Global parameter that shows probability that user will be exposed to misinfo
            - Fails to account that there are more factors that influence if people are exposed (ie: spread from neighbors/ familiar people)
    - New model called rumor centrality from Shah and Zaman (2011)
- Limit diffusion of misinfo
    - Two main approaches
    - 1st approach: Monitor spread of true info with fake info, those exposed to true info cannot be infected by false info (true info overpowers false info)
        - Must perfectly know starting points / sources of misinfo to do this approach

	- Node Protector problem: find smallest set of nodes that start with true info that can counteract spread of false info - Some studies also try to maximize spread of true info

- 2nd approach: having nodes that serve to be a **blocker** of misinfo
    - IS it possible to combine these two techniques?
    -
- Approach used in this paper: Place monitors throughout network that can detect misinfo and block it
    - Best to have as few monitors as possible
    - Limit as many nodes exposed to misinfo as possible
    - These two somewhat contracting goals makes this very difficult
- Purpose is to Strengthen model proposed by Zhang et al. (2015a)
-

2 Source Identification

- Section is about approaches to identity sources, both for known and unknown # of sources

2.1 The Approach

- Considering a subset of the whole network
- First only look at single source of misinfo
    - Use a Arborescence of the subset to find a root (starting node)
        - Root can be considered the starting source of misinfo
        - computing spanning arborescences is well studied
- Look at multiple sources of misinfo
    - Can consider multiple arborescence models in the same system
        - Branching: “forest of disjoint arborescences”

2.2 MILP Formulation

- Oh good jesus what am i looking at
- What even is a MILP approach

3 Monitor Placement

- Trying to find the minimum number of misinfo monitors needed for the least amount of spread
- A lot of math and algorithm talk

4 Experiments

- Validate proposed approaches by using examples from real

	world data
	- Tests conducted on some high-end computer ("CentOS Linux 7, equipped with an Intel Xeon E5-2650 v3 processor running at 2.3GHz and 128 GB of RAM")
	- Algorithms and Independent Cascade Model implemented in Python
	4.1 Source Identification
	- Tests for source identification:
	    - 12 instances
	        - 10 directed graphs
	        - 2 undirected graphs
	    - 8 instances from Social category of Konect database (http://konect.uni-koblenz.de)
	- Results:
	    - MLIP model is very fast,
	        - at most 30 seconds for a single test
	        - 9 out of 12 instances test takes at most 3 seconds
	    - Nodes that are considered sources by the model have distance 0
	    - 80% of true sources of misinfo are identified correctly for 6 out of 10 instances
	    - 63.33% of sources (19 out of 30 sources) identified correctly for Advogato and Youtube links
	    - Only result where identification correctness was below 50% was political blogs
	    - Undirected instances (Facebook) generally had worse results, potentially showing the used method isn't effective for this type of case
	    - A larger solution space make the used method more effective
	    - Success rate of the paper's algorithm is 5 times more than that of NNT from a past study
	    - Paper's algorithm outperforms MMSC (Zhang et al. (2015a) because less monitors are placed and less nodes are infected
	5. Conclusions and Future Work
	- Potential future work is to consider when location of seeds change over time. First steps for this taken by Auletta et al. (2020)

Research Question/Problem/ Need	Create algorithms that are able to reliably identify sources of misinformation in a network and place monitors that can block further spread.

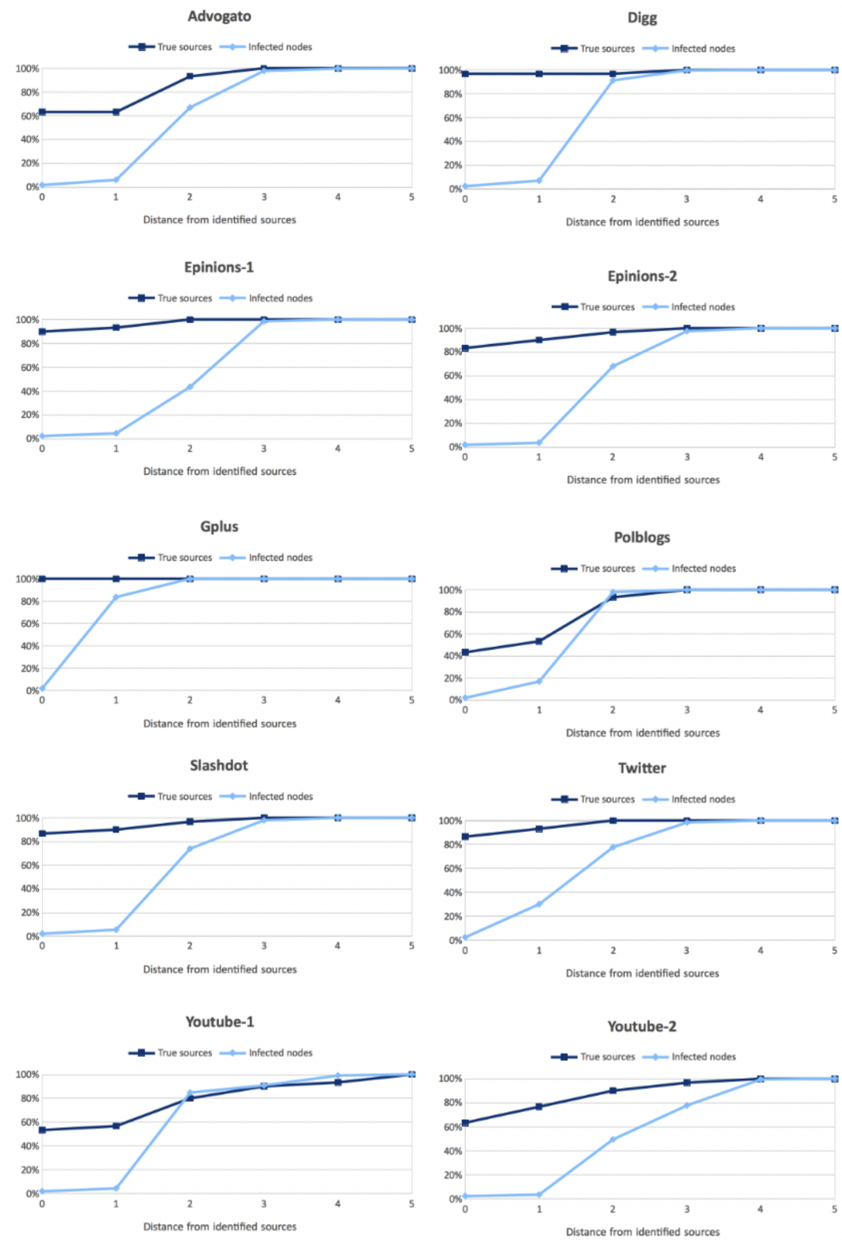
<table>
<tr>
<td>Important Figures</td>
<td>


Figure 1: Source identification model performances on directed social graphs

Graph shows what percentage of source nodes of misinfo are caught based on the algorithm’s assumptions of where the roots are.
</td>
</tr>
<tr>
<td>VOCAB: (w/definition)</td>
<td>
Node: a point in which lines or pathways intersect or branch; a central or connecting point
Heuristic: proceeding to a solution by trial and error or by rules that are only loosely defined
Arborescence: Apparently “tree diagram” in french, treelike; in this context it’s a system that starts from only one node, called the root
</td>
</tr>
</table>

	Cascade: a process whereby something, typically information or knowledge, is successively passed on.
Cited references to follow up on	Auletta, V., De Nittis, G., Ferraioli, D., Gatti, N., & Longo, D. (2020). Strategic monitor placement against malicious flows. In *ECAI '20.* Zhang, H., Alim, M. A., Thai, M. T., & Nguyen, H. T. (2015a). Monitor placement to timely detect misinformation in online social networks. *InICC '15*, pp. 1152–1157. (mentioned a lot and basis for algorithm to limit spread of misinfo)
Follow up Questions	How does the type of misinformation being spread affect how the information is spread? Can knowing how information spreads help in the recognition of misinformation? Can you use recognized misinformation to more easily identify possible sources / creators of misinfo? Can this method of analysis of networks be used for forming a network of news articles?

Article #6 Notes: Real-Time Prediction of Online False Information Purveyors and their Characteristics

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

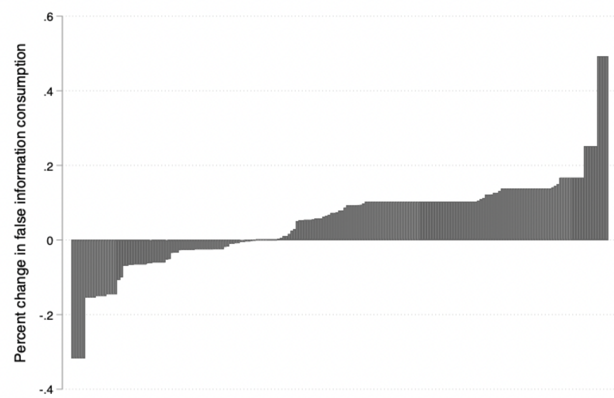
Source Title	Real-Time Prediction of Online False Information Purveyors and their Characteristics
Source citation (APA Format)	Doshi, A. R., Raghavan, S., & Schmidt, W. (2020). Real-Time Prediction of Online False Information Purveyors and their Characteristics [SSRN Scholarly Paper]. https://doi.org/10.2139/ssrn.3725919
Original URL	https://dx.doi.org/10.2139/ssrn.3725919
Source type	Research Article
Keywords	False information, false information campaigns

Summary of key points + notes (include methodology)	This paper details a way to detect sources of misinformation just with domain registration data. It mainly uses fake news domains that were in operation during the 2016 presidential election. This can be used as a first line of defense against misinfo even before articles start to appear on the domain.
	—
	1 Introduction
	- Disinfo, misinfo, and other forms of “fake news” becoming very common online
	- False info campaigns have targeted:
	    - Nike
	        - To damage reputation and do economic harm
	    - Competitors of a company
	        - Using Facebook accounts for commercial disinfo campaign
	    - Local and national communities and governments
	        - Fabrication of explosion at chemical plant in Atlanta Georgia
	        - Russia, China, And Iran possibly spreading misinfo about COVID-19 in US
	- Data used from 2016 US presidential election
	- Machine learning models can detect website registration data to find domains that will likely :

	- Make false info - Make false info that will be highly exposed to others - Will shut down after an event of interest has ended

- Lot of evidence of spread of misinfo in political settings
    - 156 news sites in 2016 presidential election shared 37.6 million times on social media
- False information campaign techniques used in political contexts most likely will also be used in non-political settings
- False info detection is an active research area
    - Find motivations for spread of deceptive info
    - Analyze how false info is shared on social networks
    - Recent analysis on how social media platforms respond to false information
- Past automated efforts to prevent spread of disinfo
    - First efforts focused on distinct word usage and network characteristic of social media spread
    - Other efforts used more complex textual models and user characteristics to identify false info
    - Overall almost always uses article text/social media content for features in models
- Model detailed in paper only uses info that is known when domain is registered, which is needed for every domain on the internet
    - Means that can be used earlier than other techniques, even before context actually appears on the domain
- Those who spread false info more rely on websites that seem to be trustworthy news outlets than exclusive spread on social media now
- False info articles and sites that contain them seem to be harder to detect and combat
- Effort needed to combat fake news much more than effort to create it
- Samples used are domains known to spread false info and other domains that existed at the same time
- Classifiers were trained on the sample's International Corporation for Assigned Names and Numbers (ICANN) data, release date of articles, and browsing history from U.S. internet users
- Model methods:
    - 1st: only using domain registration data to predict if domain will create false info
        - Inspired by Guzman, J. and S. Stern (2015). Where is Silicon Valley? Science 347 (6222), 6069.
    - 2nd: Predict outcome based on how much false info is consumed from dataset of browsing leading up to 2016 election

	- 3rd: identify false info providers with certain type of operating profile, may indicate domain's purpose
	- Early-identity finding system could:
	  - help false info be eliminated more quickly
	  - Complement other models that detect false info via content
	  - Combination of verification tools could reduce chance of identification errors
	2 Data
	- Database provided by Mozilla Corporation
	  - Recruited 2680 US Firefox users to monitor their web browsing habits in months leading up to 2016 election
	  - 26,310 of 2,670,124 webpage visits were to false info sites
	- False info domains and content database
	  - From Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives 31 (2), 211236.
	  - Found fake articles from fact checking services from Buzzfeed, Snopes, and Politifact
	  - Sample of 883 fake news articles on 363 domains
	- Domains registered before election from DomainTools
	  - Also gives domain registration data
	    - Name of domain
	    - Extension (.com, .gov, .org)
	    - Names and contact info of registrant
	    - Site administrators
	    - billing administrators
	    - technical administrators
	    - Registration date
	- 2.1 Outcome Measures
	  - 3 outcomes:
	    - Is the site a false information domain?
	    - What is the efficacy of the false info domain?
	      - Based on average domain visits whenever false info article from any of the noted databases is published
	    - Did the domain shut down by June 2017?
	      - 27% of domains
	- 2.2 Feature Extraction
	  - Using various data from the domain registration data to create a set of 957 features
	3 Methods
	- Used LASSO, form of penalized regression
	4 Results
	- Used range of 0 to 1 to classify outcome
	  - If greater than or equal to 0.7, labeled with outcome

	- If less than 0.7, not labeled with outcome - 1st model acts as first line of defense and can be used as an indication for further monitoring - Domains that ended up shutting down operations after the election were actually more effective during the period of the election than those that continued operations -
Research Question/Problem/ Need	We can use just the domain registration data of a website to know if the domain will produce misinformation in the future the moment it is created.
Important Figures	Figure 2: Average change in false-information consumption on the release date of a false information article  *Note:* $n = 335$. Each bar represents the average false-information consumption as a percent change over the surrounding five days (from two days prior to two days after) of a false information article on the date it was published. Shows the percentage change of false-information consumption any time one false info article from any of the detected sources is released. Shows the range from two days before to two days after the article is released.
VOCAB: (w/definition)	Domain Registration Data: the data that results from reserving a domain on the internet for a certain period of time Salient: most noticeable or important. Classifiers: an algorithm in machine learning that organizes data in groups Contemporaneously: existing, occurring, or originating during the same time
Cited references to follow up on	Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1. (machine learning algorithm used in methods)

Follow up Questions	Since the data here is based on articles from the 2016 US election, will there be significant changes when looking at data from different contexts? The user interactions detected here only come from users of Firefox that gave permission to have their browsing data be used in a study. Will changes occur if we look at a browser like Google Chrome? How can you get information from sites that are inactive now? How does the spread of misinformation change when there are no significant events occuring?

Article #7 Notes: Defending Against Neural Fake News

Article notes should be on separate sheets

Source Title	Defending Against Neural Fake News
Source citation (APA Format)	Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. https://doi.org/10.48550/ARXIV.1905.12616
Original URL	https://doi.org/10.48550/arXiv.1905.12616
Source type	Journal Article
Keywords	Fake News, Neural News, natural language generation, disinformation, Grover, false news

Summary of key points + notes (include methodology)	This paper introduces a generative model called "Grover", which is able to both generate news and detect machine created news. Specifically, based on parameters like the body text, title, and author of the article, the AI can generate other components. This is important because as time goes on, more and more fake news will be produced by machines. It is noted that Grover's machine created fake news is noted to often be of better quality than human written fake news. A surprising part of this study is that Grover was accurately able to detect its own machine-generated fake news and those of other AIs. It is also able to differentiate machine generated news and human written news accurately.
	—
	1 Introduction
	- Mainly focused on Grover, a model that is able to detect and generate computer-made fake news articles
	- Fake news is made to:
	    - Gain ad revenue
	    - Influence the opinions of others
	    - Change the results of elections
	- Majority of disinformation seems to be human made
	- As time goes on, more and more fake news articles will be made by computers
	- Humans rate disinformation from Grover as more trustworthy than disinformation created for humans
	- Pretrained language models are about to detect Grover's fake news with 73% accuracy
	- Grover is able to detect its own fake news with 92% accuracy

-

2 Fake News in a Neural and Adversarial Setting

- Many types of fake news ranging for satire to propaganda
- Paper mainly focuses on news articles: stories and metadata with false info
- Fake news written by humans for two big reasons:
    - Monetization (ad revenue)
        - News usually going to be viral content
    - Propaganda
        - News advances a certain agenda, has to be persuasive
- Big market for fighting misinfo on internet
    - Platforms like Facebook promote trustworthy sources and disable account that spread misinfo
    - Users of platforms use:
        - Tools: NewsGuard and Hoaxy
        - Websites: Snopes and Politifact
        - All these tools rely on manual fact checking, little to no automation
    - Main approach to automate fake news:
        - Point out stylistic biases in text
        - Good for social media platforms
    - Fact checking not completely reliable due to cognitive biases
        - Backfire effect
        - Confirmation bias
- Framework: adversarial game, with two players:
    - **Adversary:** generate fake stories that have certain attributes / purposes. Seems realistic to humans and verifier. Will be referred to as "fake news generator"
    - **Verifier:** classify if stories are real or fake
        - Access to unlimited real stories
        - Limited # of fake news stories
    - As verifiers get better, so will adversaries

3 Grover: Modeling Conditional Generation of Neural Fake News

- **Mainly details methods**
- Can expect fake news generator to make targeted content
- Many generative models have realistic text but don't know when to stop (not "controllable generation")
- Grover's generated text is realistic and controllable.
- Document treated as text field with start and end marker
- 5 components for generated news article:
    - Domain
        - What site the article is published
    - Date
        - When the article is published
    - Authors

- Generated names
- Headline
    - Title of article
- Body
    - Article content

- Discussion of methods used by Grover that I don't understand at the moment
- If some components are known, Grover can generate other missing components
- Grover's architecture based on GPT2
- Dataset used is RealNews, a bunch of news articles from Common Crawl
    - https://commoncrawl.org/ (seems to be an organization dedicated to collecting and sharing various types of data)
    - Limited to 5000 news domains indexed by Google News
    - Used Newspaper Python Library to extract article content (body)
- Trained Grover models on randomly sampled sequences in RealNews with length 1024

3.1 Language Modeling results: measuring the importance of data, context, and size

- **Exactly what it says: the results**
- Grover improves when it's given the full metadata as opposed to when no context is provided
- Section 3.2 is something about why "Nucleus Sampling" was used

4 Humans are Easily Fooled by Grover-written Propaganda

- 4 classes of articles considered
    - **Human News:** Reliable human written news
    - **Machine News:** Grover written articles based on metadata of Reliable human written news
    - **Human propaganda:** Human-written propaganda articles
    - **Machine propaganda:** Grover written articles based on metadata of Human-written propaganda articles
- Qualified workers at Amazon Mechanical Turk rate these articles based on 3 criteria:
    - Stylistic consistency
    - Content sensibility
    - Overall trustworthiness
- Quality of Machine news not as good as human news
- Quality of Machine propaganda better than human propaganda
    - Noticable difference in overall trustworthiness
- Data shows that machine generated fake news could become bigger concern if it gets more widespread

5 Neural Fake News Detection

- Models for role of verifier by detecting at Human or Machine written:

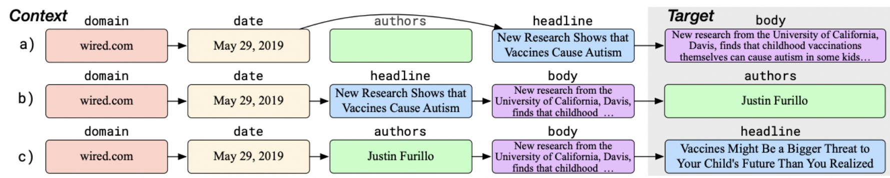
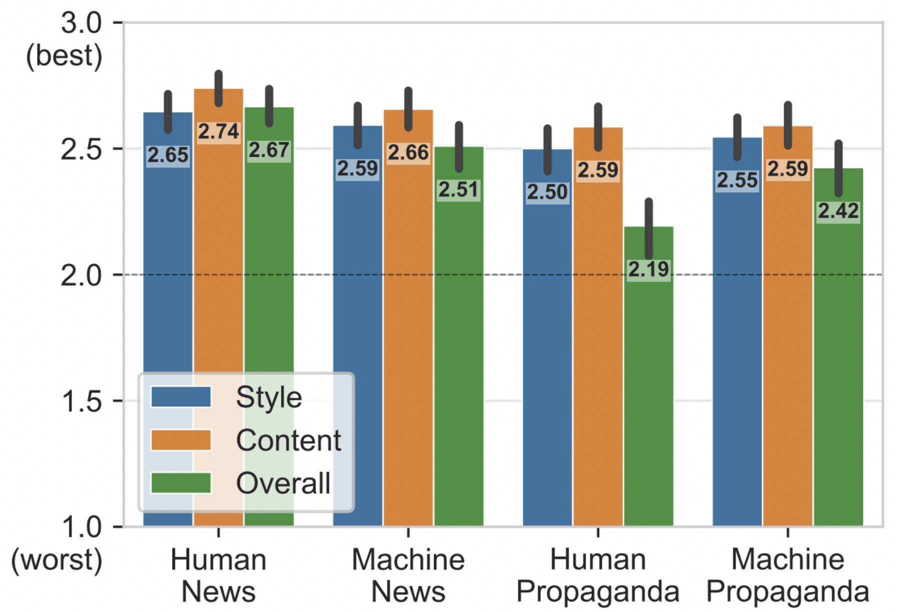
- Grover
- GPT2
- BERT
- FastText
- 2 ways to evaluate:
    - **Unpaired**: each verifier given a single news article and has to determine if its from a human or a machine
    - **Paired**: each verifier is given two articles with the same metadata, one being human written and one because machine generated. It has to determine which one is human written and which one is machine generated
- **Results:**
    - Much harder to determine with unpaired setting compared to paired setting
    - Grover could accurately track roughly 90% of the time
        - If larger generator accuracy below 81%
        - If discriminator larger accuracy above 98%
    - Other verifiers performed worse than Grover

6 How does a model distinguish between human and machine text?

- Exposure Bias
    - Other modals are not trained on computer generated articles, only on human generated ones
- Variance-reduction
    - If a model had reduced variance it leaves an "artifact"
- Basically:
    - Too little variance = bad
    - No limits to variance = also bad
- Grover most likely able to detect own fake news because can detect tail the best

7 Conclusion: a Release Strategy for Grover

- How humans interpret Grover's generated articles proves that neural news generation can be very dangerous
- There are Defenses to these models, like Grover itself
- Next steps:
    - Training Grover was relatively inexpensive, so getting generators of fake news will only get easier and easier
    - Since Grover is both an effective creator and detector of fake news, releasing these types of models is an important means of defense
    - Future in generation? (lot of lingo i don't understand here)
    - Further studies in other types of threat models
    - Analyzing computer generated real news
    - Discriminators that are effective mainly rely on structure (?) of known news articles. Find a way to instead use a model of the world, like humans do when they check misinfo
    - Using deep neural networks, like those used for videos

	on YouTube to check for inappropriate content, to check news articles as well
Research Question/Problem/ Need	How can we effectively detect neural fake news, or news that is generated by an AI?
Important Figures	 Figure that shows how Grover is able to fill in information for news articles based on already known information about it.  Stats of how, on average, workers at Amazon Mechanical Turk felt about the style, content, and overall quality of Human and Machine written material. Grover is noticeably good at writing propaganda.
VOCAB: (w/definition)	**Adversarial game:** - Adversarial: opposed; hostile - So Adversarial game most likely game in which two or more computers/players are pitted against each other **Backfire effect:** When someone is shown evidence that their current beliefs are wrong, they tend the reject that belief and strengthen their own argument (https://effectiviology.com/backfire-effect-facts-dont-change-minds/) **Confirmation bias:** The tendency to look for information that supports one's own stance as opposed to the opposite stance (Casad, B. J. (2019, October 9). confirmation bias. Encyclopedia

	Britannica. https://www.britannica.com/science/confirmation-bias) **generative models:** models that are trained on a large amount of existing data to generate new data like said given data. (https://openai.com/blog/generative-models/)
Cited references to follow up on	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. (source mentions GPT2's architecture, grover's architecture is noted to be similar) Code for Grover: https://github.com/rowanz/grover
Follow up Questions	How can news generators also function as a fake news detector? (Are their systems reverse engineered in some way?) Can these sorts of generators also be trained on a certain type of fake news source to produce stronger results for that niche? (ie: medical, political) How can this sort of model be implemented into social media algorithms? What was the progress of previous models before Grover?

Article #8 Notes: Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share

Article notes should be on separate sheets

Source Title	Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share
Source citation (APA Format)	Effron, D. A., & Raj, M. (2020). Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share. *Psychological Science, 31(1*), 75–87. https://doi-org.ezpv7-web-p-u01.wpi.edu/10.1177/0956797619887896
Original URL	https://doi-org.ezpv7-web-p-u01.wpi.edu/10.1177/0956797619887896
Source type	Research article
Keywords	Morality, Misinformation,

Summary of key points + notes (include methodology)	The researchers did four experiments to find if repeated encounters with certain misinformation will cause people to think that spreading it is less unethical. All the experimentals had a similar format: people were shown 6 fake news headlines multiple times (except for experiment 2) and then polled to see how unethical they felt it would be to spread this headline with 6 new headlines mixed in. This correlation seemed to be positive for all experiments conducted.
	- Sometimes people feel that spreading misinfo can be morally right if it supports their viewpoint
	- If they feel that the spread of misinfo is permissible,
	    - they won't take action to stop it
	    - They won't hold spreaders of misinfo accountable
	    - More likely to spread it themselves
	- 14% of US adults and 17% of UK adults have admitted to spreading news that they knew was fake at the time
	- Fake news: "articles that are intentionally and verifiably false, and could mislead readers" (Allcott & Gentzkow, 2017, p. 213)
	- Fake articles are more likely to go viral on social media, making people come across it multiple times
	- People are more likely to believe a news headline if they encounter it multiple times

- Hypothesis: prior exposure to a piece of fake news will reduce a person's judgment of how unethical it would be to spread it, regardless of if they believe the news
    - Previously encountered info is easier to process ("feels more fluent") than new information
    - If they can process the information, they will associate that information with the truth
    - Due to this association, repeated info can have a "ring of truthfulness" regardless of beliefs
    - People can instinctively believe information even if they explicitly say it is false
- illusory-truth effect: judging repeated statements as more accurate
- If someone knows a claim is false when they first encounter it, when they encounter it again they may forget and misunderstand they fluency of the statement as truth
- This research assumes that the claims are known to be false and finds how they are morally judged

Experiment 1: will 4 previous encounters with a fake-new headline make the headline seem less unethical to publish?

Method

Participants:

- 150 US participants on Prolific Academic
    - Participants are diverse and are less familiar with experimental procedures compared to Amazon Mechanical Turk workers
        - Amazon Mechanical Turk workers were used in "Defending against Neural Fake News"
    - Experiment 2 and its pilot experiment was conducted first, but experiment 1 presented first for clarity
    - Participants that failed a reading-comprehension question, were on a mobile device, or lived outside the US could not participate in the experiment.
    - Dataset contained 1648 observations and 138 people
        - 75 men
        - 63 women
        - M = 34 years
            - Is this mean or median?
        - SD = 13
            - Probably standard deviation
        - Range = 18-74
        - 95 Democrat leaning
            - Does this majority democrat population affect the results?

- 22 Republican leaning
- 21 lead towards no party
- Some stats thing

Materials

- Stimuli: 12 actual fake-news headlines in regards to american politics with photographs for a fact checking website
- Half of headlines appealed to Republicans, other half appealed to Democrats

Procedure

- Adapted from procedure of Pennycook et al. (2018)
- Has two phases
    - *Familiarization phase*
        - Participants saw 6 of 12 headlines 4 times
        - Each time headline is shown, rated each headline on:
            - "How interesting is this headline?"
            - "How engaging is this headline?"
            - "How funny is this headline?"
            - "How well-written is this headline?"
        - After each rating they completely a distractor task
    - *Judgment phase*
        - Participants are shown all 12 headlines
            - Half seen in familiarization phase
            - Half are new
        - Message introducing judgment phase:
            - "For this part of the study you will be asked to read a series of fake news headlines that were recently published online. The information in these headline is not real. Non-partisan fact-checking websites have confirmed that these headlines describe events that did NOT happen."
            - This is so that the subjects are not influenced by illusory-truth, making it very clear that all of the headlines are false
    - Randomly determined:
        - Headlines shown in familiarization phase
        - Order in which filler items were completed
        - Order in which headlines appeared in both phases

Measures

- *Moral condemnation*
    - Participants moved slider to indicate how:
        - Unethical it would be to publish each headline

- Acceptable it would be to publish each headline
    - The two items were averaged
- *Intended social-media behaviors*
    - Participants asked how likely they would ___ if an acquaintance shared the headline on social media on a scale of 1 to 7:
        - "Like" it
        - Share it
        - Post a negative comment
        - block/unfollow the acquaintance
    - Each item was analyzed individually
- *Accuracy beliefs*
    - Mainly to see if illusory-truth had any effect on findings
    - Participants asked to rate each headline's factual accuracy
        - 4 point scale was used, much like previous research on illusory truth
- *Comprehension check*
    - After judgment phase
    - Check if participants understood that they judged fake news articles
    - Asked if:
        - All headlines were true
        - All headlines were false
        - Some were true and some were false
            - If this was chosen, the responder was given the 12 headlines again to choose which they thought were true and which they thought were false

Results

- *Moral condemnation*
    - As expected, headlines that were previously seen were rated as less unethical to publish than those that were new
- *Intended social-media behaviors*
    - Participants indicated that they were most likely to like and share headlines that they have previously seen
    - Less likely to block or unfollow the person who posted the previously seen headlines
    - Effects were mediated by moral judgments
        - Consistent with expectation that exposure to headlines will affect social media behaviors by softening judgments
        - What are these moral judgments?
    - Posting a negative comment seemed to not be dependent on if they were looking at a new headline

	or previously seen headline
	- *Accuracy beliefs*
	    - Previously seen headlines were still seen as unethical to publish compared to new headlines (no illusory truth effect)
	Discussion
	- Repeat encounters with a fake news headline can
	    - Reduce people's moral issues for publishing it
	    - Increase want to promote on social media
	    - Decrease chance of blocking or unfollowing someone who shared it
	- Illusory-truth effect was not replicated, which is expected
	    - Placed emphasis on the fact that all headlines were false before judgment phase
	    - Participants didn't believe previously seen headlines any more than the new headlines
	    - Unlikely that participants forgot that the headlines were false
	
	Experiment 2: Will a single encounter with a fake-news headline make it seem less unethical to publish?
	- Is a large-sample, preregistered replication
	Methods
	- Participants
	    - 800 US workers on Amazon Mechanical Turk in June 2018
	        - Informed by previous experiment with 596 participants
	    - Could not participate if they
	        - Failed a reading-comprehension test
	        - Responded from a mobile device
	        - Responded from a non-US ip address
	    - 9536 observations from 796 people
	        - 467 women
	        - 326 men
	        - 3 nonbinary
	        - 458 leaned Democrat
	        - 223 leaned Republican
	        - 115 no party leaning
	        - M = 34 years
	        - SD = 12
	        - Range = 18-76
	    -
	- Procedure and measures
	    - Identical to that of Experiment 1 except:

- In familiarization phase, six headlines were only shown once (as opposed to 4)
- Filler was only how interesting they felt each headline was

Results

- Like in experiment 1, headlines were rated to be less unethical to publish if they have seen them before

Discussion

- Just encountering a fake news article once is enough to have people say that it is less morally reprehensible to publish it

Experiment 3: If people are encouraged to think deliberately (haha Thoreau) about the false claim as opposed to intuitively, will the spread of that misinformation seem more unethical to them?

- Argued that previous encounters with fake news headlines make them feel intuitively, even if it is known that they are false
- On that basis, shifting moral judgment from intuition to deliberation should lessen this effect

Method

- 2 x 2 factorial design with 12 repeated measures
    - What does this mean
- Participants
    - Requested 600 complete responses from Prolific Academic in November 2018
    - Could not participate if they
        - Took part in Experiment 1
        - Failed a reading comprehension test
        - Responded from a mobile device
        - Responded from a non-US IP address
    - Headlines were shown 4 times, like in experiment 1
    - 8,731 observationsfrom 761 people
        - 407 men
        - 345 women
        - 9 nonbinary
        - M = 33 years
        - SD = 12
        - Range = 18-76
        - 509 democrat leaning
        - 147 republican leaning
        - Others not party leaning
    - More than 600 responses were obtained because
        - Some people submitted incomplete by anayliable responses
        - Prolific Academic did not count responses

	submitted more than 20 min after the beginning of the experiment
	- Hypothesis: repeated exposure would have a smaller effect in deliberative-thinking condition than in the intuitive-thinking condition
	Procedure:
	- First part is very similar to Experiment 1 (participants views 6 headlines four times, provide filler ratings, introduce judgment phase saying that every headline is false)
	- At this point it deviates into two groups
	  - Those assigned to deliberate thinking condition
	    - were told to “take time to deliberate,” “think very hard,” “ignore any gut feelings,” and “generate clear reasons” about the ethics for publishing every headline
	    - Before rating a headline, they had to type two reasons for their choice
	  - Those assigned to intuitive thinking condition
	    - Assigned to quickly rate headline’s ethicality via “their first instinct,”, to “pay attention to [their] feelings”, and to not “think too hard.”
	    - Could not type their reasoning for ratings
	- Everything else for Experiment 1’s judgment phase is basically the same
	- For the final part, participants completed the three-item cognitive reflection test (CRT)
	  - Purpose is to “assesses individual differences in deliberative thinking”
	  - Since participants were conditioned into thinking deliberately if they were in the “deliberate thinking condition” group, researchers thought individual differences would not be made apparent
	  - These results did not significantly moderate results, so they aren’t expanded upon
	Results
	- Moral condemnation
	  - Hypothesis: previously seen headlines get less moral condemnation than new, effect is reduced if encouraged to think deliberately
	  - Hypothesis seems to be correct: effect was halved with deliberate thinking group compared to intuitive thinking group
	- Intended social media behaviors
	  - No evidence that deliberate thinking affected intended social media behaviors
	- Moderated mediation analysis
	  - Data to support that is weak

Experiment 4: Will repeated exposure to false headlines have effects on moral judgments beyond accuracy, likeability, and popularity? (participants were not warned that the headlines were false for this experiment to test for generalizability)

- Measured potential factors of repeated exposure that were not accounted for in the previous experiments
    - How much people like it
    - How popular the participate thinks it is
- Also tested if repeated exposure could increase the chance of someone sharing the headline in a experimental setting

Method

- Participants
    - "posted slots for 300 U.S. Prolific Academic users in March 2019"
        - Sample size chosen because double the number of Experiment 1's provides good point for comparison
    - Specifically requested an equal number of Democrats and Republicans because all previous experiments had majority Democrats
    - Same requirements for participation as previous studies
        - Including a captcha test to detect bots
            - Why wasn't this in any of the earlier ones
    - 3,552 observations from 296 participants
        - 147 men
        - 147 women
        - 2 nonbinary
        - M = 34 years
        - SD = 13
        - Range: 18-74
        - 151 Democrat
        - 142 Republican
        - 3 did not complete politics measure
- Procedure
    - Very similar to previous experiments
    - Participants were NOT informed that the headlines were fake until the very end of the experiment
    - Measures are somewhat different
- Measures
    - Moral condemnation
        - Asked how unethical it would be to share the

	headline (as opposed to publishing the headline)
	- Control variables
	  - How accurate did they feel the headline was
	  - How much did they like the headline
	  - How popular did they think the headline was
	  - Also had exploratory measure of how well written the participants thought the headline was (not expanded upon)
	  - All these use a 100 point scale via a slider unlike the other studies
	- Sharing intentions and behavior
	  - Two measures to see potential consequences of moral condemnation
	    - Behavioral-intentions measure: how likely would you share the headline if someone you know shared it? (answered after other questions mentioned before)
	    - Behavioral measure: after all the headlines were rated, told:
	      - “We would like to run a study where research participants see some of the headlines you’ve been looking at today. These participants can choose to click on the headlines to read the full article.”
	      - Would be shown the 12 headlines again in a random order, told to choose 4 headlines to share with next “study”
	      - Expectation is for people to select more previously seen headlines to share
	Results
	- Dependent measure: moral condemnation
	  - Effects of prior expose to moral condemnation is independent on judgements of accuracy, liking, and popularity
	- Sharing intentions
	  - More inclined to share previously shared headlines compared to new headlines
	- Sharing behavior
	  - Shared more headlines that they previously encountered

Discussion

- Repeated encounters = reduction of moral condemnation

General Discussion

- More encounters with a piece of misinfo = said misinfo seeming less unethical to spread
- Suspect that people associate fluency with truth
- Other explanation is that fluency feels good and encourages positive feelings, regardless if fluency associates with belief in truth
    - Seems less likely, tho
- Repeat encounters make people think the headline is popular and, therefore, reliable
- Findings are separate from illusory truth effect

Research Question/Problem/ Need

How does previous exposure to fake news headlines affect how people perceive their ethicacy?

Important Figures

Table 1. Results of Experiment 1

		Previously seen headlines		New headlines							
Measure	Response range	*M*	95% CI	*M*	95% CI	Mean difference	d_z	*b*	*SE*	*z*	*p*
Moral condemnation	0–100	66.37	[62.49, 70.55]	70.44	[66.32, 74.38]	–4.07	–0.26	–3.83	0.99	3.86	< .001
Intentions to "like"	1–7	1.73	[1.54, 1.89]	1.55	[1.38, 1.73]	0.18	0.24	0.16	0.05	3.18	.001
Intentions to share	1–7	1.62	[1.44, 1.80]	1.49	[1.31, 1.68]	0.13	0.27	0.13	0.04	3.19	.001
Intentions to post negative comment	1–7	2.06	[1.83, 2.29]	2.15	[1.92, 2.38]	–0.09	–0.14	–0.09	0.05	1.68	.092
Intentions to block	1–7	2.29	[2.04, 2.55]	2.47	[2.22, 2.72]	–0.17	–0.23	–0.17	0.06	3.06	.002
Accuracy beliefs	1–4	1.45	[1.35, 1.56]	1.43	[1.33, 1.53]	0.02	0.04	0.02	0.02	0.80	.425

Note: We computed 95% confidence intervals (CIs) from the multilevel regression model.

Table 2. Results of Mixed Regression Models for Each Dependent Measure in Experiment 3

Measure and predictor	Initial analysis: Interaction model	Initial analysis: Main-effect model	Robustness check: Interaction model	Robustness check: Main-effect model
Moral condemnation				
Headline type	−2.84***	−2.19***	−2.87***	−1.99***
Thinking condition	3.33*	4.03*	3.19†	4.14*
Headline Type × Thinking Condition	1.39†		1.92*	
Intentions to "like"				
Headline type	0.08**	0.06**	0.08**	0.05*
Thinking condition	−0.16*	−0.18*	−0.14*	−0.18**
Headline Type × Thinking Condition	−0.04		−0.07*	
Intentions to share				
Headline type	0.03	0.03	0.04	0.02
Thinking condition	−0.16*	−0.17*	−0.14*	−0.16*
Headline Type × Thinking Condition	−0.01		−0.04	
Intentions to block				
Headline type	−0.19***	−0.17***	−0.18***	−0.16***
Thinking condition	0.29*	0.31*	0.30*	0.32**
Headline Type × Thinking Condition	0.03		0.05	

Note: Predictors were dummy coded (headline type: 1 = previously seen, 0 = new; thinking condition: 1 = deliberative, 0 = intuitive). Thus, coefficients are main effects only in the models without an interaction term. The robustness check excluded participants' responses to any headlines they misidentified as true. The regression models also included a fixed intercept, random intercepts for participants, and fixed effects for headline type. Significance tests were two-tailed for simple and main effects and one-tailed for interactions because we preregistered directional predictions for the interactions.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

VOCAB: (w/definition)	**Preregistered:** practice of documenting a research plan before the study is conducted **Attenuate:** reduce the force, effect, or value of **distractor task:** a stimulus or an aspect of a stimulus that is irrelevant to the task or activity being performed. In memory studies, an item or task may be used as a distractor before the participant attempts to recall the study material to be remembered https://dictionary.apa.org/distractor **Replication:** In scientific research, the repetition of an experiment to confirm findings or to ensure accuracy. one-tailed tests:
Cited references to follow up on	None (there are interesting things here but they deviate from the objective of my project)
Follow up Questions	How can we account for the behaviors of humans when developing an AI? Can we depend on people to follow the instructions or advice of others? Even if people are directly told that certain news is false, will they subconsciously take that to heart? How can we make people think more critically about the news and

	media that they consume?

Article #9 Notes: Better Language Models and Their Implications

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Better Language Models and Their Implications
Source citation (APA Format)	Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., Sutskever, I., Askell, A., Lansky, D., Hernandez, D., & Luan, D. (2019, February 14). *Better language models and their implications*. OpenAI. https://openai.com/blog/better-language-models/#task6
Original URL	https://openai.com/blog/better-language-models/
Source type	General Article
Keywords	Language Models, GPT-2, transformer-based. synthetic text
Summary of key points + notes (include methodology)	A general OpenAI blog post detailing general information about GPT-2, a language model that is able to predict the next words of a given input. The style of writing that GPT-2 models is adaptable based on the writing style of the input. However, it often takes multiple attempts for the model to output an adequate result. GPT-2 is also able to produce ok results for other language tasks, those they are not as good as results from models specifically designed for those tasks. — Intro - Purpose of GPT-2 model is to predict next word over 40 GB of internet text - Is a transformer-based language model with 1.5 billion parameters - Trained on dataset of 8 million web pages - Objective: generate the next word given all the previous words - Improvement on previous modal GPT - Training data come from outbound links from reddit with more than 3 upvotes (filtered by humans) - This is doubtful though since there are a lot of bots on reddit - Outperforms other models trained on specific domains, even

	though GPT-2 isn't trained on a domain specific data set
	Samples
	- Modal is able to continue a text input and adapts to the style of the given text
	- This generated text is mostly coherent and has human level quality, though repetition and logic errors do crop up
	- If the topic of the input is well-represented in the trained data, the generated data will be reasonable 50% of the time
	    - If the topic isn't well represented the modal can perform poorly
	- Often takes multiple attempts to get reasonable result
	- Shows that as more time goes by, language models will become easier and easier to customize
	Zero-Shot
	- Though the model is not trained on any domain-specific data, it performs well on specific modeling tasks and better than models that are domain-specific
	- Also performs ok on other language tasks (question answering, reading comprehension, summarization, translation), though they are far from models that are specifically designed to do these tasks
	    - With more computations and data these may get better
	Policy Implications
	- Good uses:
	    - AI writing assistance
	    - Better dialogue agents
	    - Improved translation
	    - Better speech recognition
	- Malicious purposes:
	    - **Misleading news articles**
	    - Impersonation online
	    - Automated false or abusive content posted on social media
	    - Automated spam/phishing content
	- Technology is reducing the cost for the production of false content and disinformation campaigns
	- malicious actors, some being political, are being use to target individuals to silence them
	    - Further generation of images, text, audio, and videos could strengthen these actors
	        - Halting these generations results in progress in AI halting as a whole, so effective countermeasures must be taken soon
	Release Strategy
	- Due to fears of GPT-2 being used for malicious purposes, versions of GPT-2 were slowly released over time
	    - This is very different from grover's strategy, though

grover was also able to detect computer generated content

- Governments should incentivize monitoring the effects of AI technology and their capabilities

Research Question/Problem/ Need

Develop a language model that is able to produce comprehensible text based on a starting input.

Important Figures

DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	**70.70%**	63.7%	92%+
LAMBADA	accuracy (+)	**63.24%**	59.23%	95%+
LAMBADA	perplexity (–)	**8.6**	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	**93.30%**	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	**89.05%**	82.3%	92%
Penn Tree Bank	perplexity (–)	**35.76**	46.54	unknown
WikiText-2	perplexity (–)	**18.34**	39.14	unknown
enwik8	bits per character (–)	**0.93**	0.99	unknown
text8	bits per character (–)	**0.98**	1.08	unknown
WikiText-103	perplexity (–)	**17.48**	18.3	unknown

GPT-2 achieves state-of-the-art on Winograd Schema, LAMBADA, and other language modeling tasks.

Shows how GPT-2 performs with various datasets, performing better than all previous records

VOCAB: (w/definition)

Misinformation: false information that is spread, regardless if it was intended to mislead others
Disinformation: Misleading information that is deliberately spread for malicious purposes
(the difference between misinformation and disinformation is **intent**)
(Source: "Misinformation" vs. "Disinformation": Get Informed On The Difference. (2022, August 15). Dictionary.com. https://www.dictionary.com/e/misinformation-vs-disinformation-get-informed-on-the-difference/)
Malicious Actors: Another term for threat actors. People or groups of people that can threaten cybersecurity
Parameter: Variables estimated and used by a machine learning model based on given inputs

Cited references to follow up on

This article is connected to a technical paper: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
Code for the model: https://github.com/openai/gpt-2

Follow up Questions

Can machine generated text result in the increase in cheating in academic settings?

	Will there be a point where machines seem to understand the content of their own writing? At what point will language models be able to do various language tasks with ease? How do language models perform on languages outside of english?

Article #10 Notes: Language Models are Unsupervised Multitask Learners ABANDONED

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Language Models are Unsupervised Multitask Learners
Source citation (APA Format)	Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
Original URL	https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
Source type	Research Article
Keywords	
Summary of key points + notes (include methodology)	
Research Question/Problem/ Need	
Important Figures	
VOCAB: (w/definition)	
Cited references to follow up on	
Follow up Questions	

Article #11 Notes: Counteracting neural disinformation with Grover

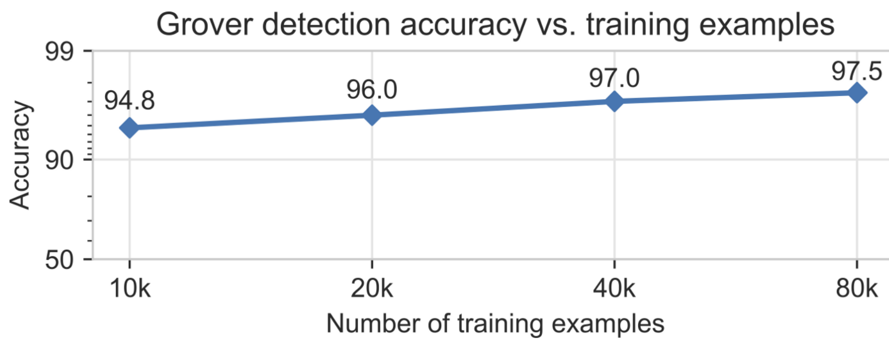
Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

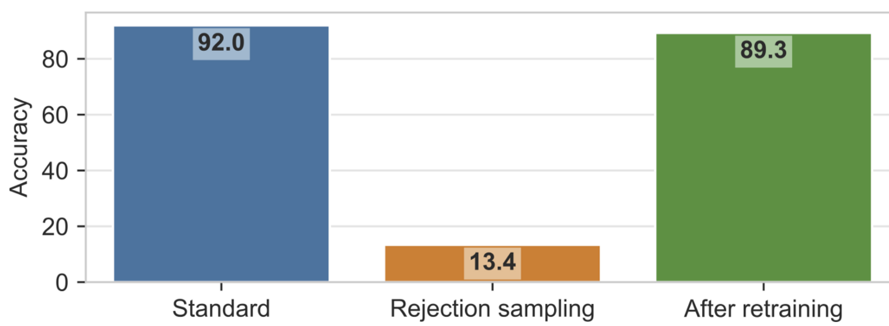
Source Title	Counteracting neural disinformation with Grover
Source citation (APA Format)	Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019, June 18). *Counteracting neural disinformation with Grover*. Medium. https://blog.allenai.org/counteracting-neural-disinformation-with-grover-6cf6690d463b
Original URL	https://blog.allenai.org/counteracting-neural-disinformation-with-grover-6cf6690d463b
Source type	General Article
Keywords	Fake News, Neural News, natural language generation, disinformation, Grover, false news
Summary of key points + notes (include methodology)	This is a blog post that extends on the research done in “Defending against Neural Fake News”. Its purpose is to answer questions in regards to the paper and provide results that expand beyond the scope in the paper. Some of the extensions observed were that Grover is able to detect human written news well, can become stronger as it is trained on more training data, can defend against a rejection-sampling attack if retrained, and the types of news it has a harder time detecting. — Directly connected to “Defending Against Neural Fake News” - Grover is good at spotting fake news because it is also able to generate it well - Purpose of articles is to answer questions and show experimental results that go beyond the scope of the paper **Part 1: New experimental results that further show that Grover is a good detector of misinformation** - Generators of fake news is also familiar with its own peculiarities, as well as those of similar language models

	- In machine learning, Ai-performance increases when there is more training data
	    - Therefore, if there is more data available, the AI will be more accurate in tracking that data
	    - "Grover's detection accuracy when given as few as 10k [machine-written fake news] articles is 94%, but it increases to up to 97.5% when trained on 80,000 [machine-written fake news] articles."
	- Fake news that Grover is good at detecting as fake:
	    - Grover model generated news (written in OG paper)
	    - OpenAI GPT-2 written news
	        - Used publicly released GPT-2 models
	        - "Without ever having seen any GPT-2 generations during training (a zero-shot setting), Grover correctly classifies 96.1% of them as machine-written."
	    - Human-written fake news
	        - Grover trained with 30k examples each of real news, machine written news, and human written fake news
	        - Grover able to detect both types of fake news with over 95% accuracy
	        - Extends on https://aclanthology.org/D17-1317/
	Part 2: See how others may try to undermine Grover's fake news detention
	- rejection-sampling attack: assume an environment where the attacker has access to the discriminator Grover is using and can generate articles until one slips through
	    - Makes big and unlikely assumption that adversary knows everything about verifier while verifier knows nothing
	    - If Grover can't be further trained in this setting (unlikely), accuracy becomes around 13%
	    - However, once retrained using generations from attacker, accuracy becomes over 89%
	    - Conclusion: rejection-sampling attack only gives adversary temp advantage
	    - Adversaries only are successful if their generations are short, which is unrealistic to assume for a news article
	- Some news is harder to detect than others
	    - High rate of successful detection for news written in the style of major news outlets (slate, bbc, The Guardian, etc)
	    - However, financial news is much harder for Grover to detect
	        - "americanbankingnews.com" only has a 60.5%

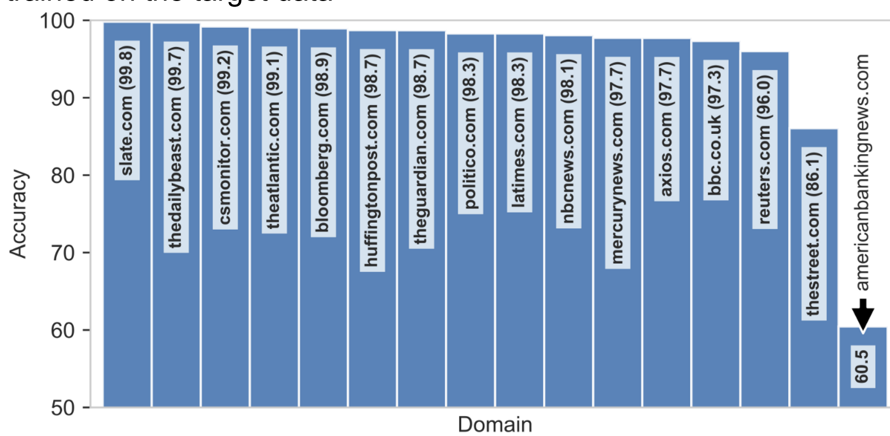
<table>
<tr><td></td><td>
detection rate
- Possibly due to many articles being automatically generated based on the price of stocks
</td></tr>
<tr><td>Research Question/Problem/ Need</td><td>Further expansion on “Defending Against Neural Fake News”. Mainly shows that Grover can detect news from a variety of different sources, including human written fake news.</td></tr>
<tr><td>Important Figures</td><td>



Shows how Grover’s detection accuracy becomes higher when it is trained on more content



Shows how Grover is able to counteract Rejection sampling if it is trained on the target data



Shows Grover’s accuracy in detecting misinfo based on the style of certain domains

</td></tr>
</table>

VOCAB: (w/definition)	Zero-shot setting: Problem setup in machine learning when during testing, model looks that data that it wasn't exposed to during training rejection-sampling attack: continually generating distributions until one falls through the cracks of the discriminator Discriminator: The classifier that can distinguish real data from fake data Crawl: (of a program) systematically visit (a number of web pages) in order to create an index of data.
Cited references to follow up on	https://aclanthology.org/D17-1317/ (shows that fake news often has common traits in their language) https://docs.google.com/forms/d/e/1FAIpQLSfPMUXH1fdUxI3TchRS9RaEJ_W-W-adZhQ-_MymHnGkGOIBkA/viewform (form to apply for access to the dataset they used and the Grover-Mega Model)
Follow up Questions	How would an adversary be able to obtain access to Grover's discriminator? Should we create preventive measures to counteract that? Would it be possible for a machine learning model to continuously retrain itself during testing? (seems unlikely though since during testing the model cannot know if info is true or false) What are effective ways to prevent certain data from being accounted for in training data? Are there any models that can effectively fact check short statements?

Article #12 Notes: Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking
Source citation (APA Format)	Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. https://doi.org/10.18653/v1/D17-1317
Original URL	https://aclanthology.org/D17-1317/
Source type	Research article
Keywords	Fact-checking, political, language
Summary of key points + notes (include methodology)	This research article mainly observes how certain vocabulary can indicate how reliable a source is and what type of unreliable source if it is unreliable. The paper also details the creation of an algorithm to see if certain statements crawled from Politifact are true or not, first using a 6-point scale and a 2-point scale. — **1 Introduction** - Words in news and politics can have big impact on beliefs and opinions - Work dedicated to fact checking tripled since 2014 (was written in 2017) - Some organizations are dedicated to fact checking the words of prominent figures, like PolitiFact - Politifact has a ranked system with 6 levels to assess a statement. Most statements are not ranked completely true or

	completely false
	- Previous work focused on only binary scale from tracking misinfo, but political fact checking needs to be more nuanced
	- Two scales for false information:
	    - Intent to decisive
	    - Trustworthiness
	- Purpose is to see how the language of political quotes can indicate truthfulness and deception
	- Also looked at a 6 point scale for detecting truthfulness using database from politifact
	2 Fake News Analysis
	News Corpus with Varying Reliability
	- 3 unreliable news types:
	    - Satire: intended to be humorous and not be serious
	    - Hoax: convince readers of a story intended to instill fear
	    - Propaganda: mislead reader into believing in a political/social agenda
	- Satire intent is not malicious, humor should be obvious
	- Satire and hoaxes often invent stories
	- Propaganda combines truth and lies to make things seem ambiguous, conflicting readers
	- Used lexical resources to trusted and fake news articles
	- See what article types use what types of words (?)
	- Tracking use of
	    - Subjective words
	    - Hedging
	    - Words that indicate dramatization (researchers crawled this by themselves using Wiktionary)
	Discussion
	- First person and second person pronouns used more in less reliable / deceptive news
	    - Possibly because editors of real news edit out personal language
	    - Previous work says it indicates imaginary writing
	- Words used to exaggerate (subjectives, superlatives, and modal adverbs) mostly used in fake news
	- Words used to demonstrate concrete ideas / figures (comparatives, money, and numbers) occur more in real news
	- Trusted sources more often use assertive words, less likely to use hedging, showing less vagueness
	- Trusted sources use more “hear category words” (???), citing sources more
	- Satire prominently uses adverbs
	- Hoaxes use less superlatives and comparatives
	- Propaganda use more assertive verbs and superlatives

	- Mimicking the language of real news
	News Reliability Prediction
	- Categorize news into 4 categories:
	  - Trusted
	  - Satire
	  - Hoax
	  - Propaganda
	- Articles that are collected split into 20k articles for training and 3k articles for testing
	- Articles from training and test sets are from different sources so model can’t rely on patterns of specific authors
	- Model is 65% accurate, which is much higher than random but still leaves room for improvement
	- N-grams (parameters?) weighted most for trusted news were
	  - Specific locations (“washington”)
	  - Specific times (“on monday”)
	- N-grams weighted most for satire
	  - Indications of flippant remarks (“reportably”, “confirmed”)
	- N-grams weighted most for hoaxes
	  - Controversial topics (“liberals”, “trump”)
	  - Dramatic cues (“breaking [news]”)
	  - “Youtube” and “video”: indicates relying on video sources
	- N-grams weighted most for propaganda
	  - Abstractions (“truth”, “freedom”)
	  - Specific issues (“vaccines”, “syria”)
	  - “Youtube” and “video”: indicates relying on video sources
	3 Predicting Truthfulness
	Politifact Data
	- Fact checks individual statements from public figures
	- Ran by journalists who actively fact check sources
	- Each quote graded on truthfulness on a 6-point scale (“True” to “Pants-on-Fire False”)
	  - Scale creates more nuance beyond basic True or False with no in between
	- Most statements not said to be fully true or fully false
	- Created model to grade politifact statement in two ways:
	  - 6 point scale
	  - 2 point scale (3 truthful ratings in true, other three as false)
	Model
	- Trained LSTM model, Maximum Entropy model, And Naive Bayes model
	Classifier Results
	- LSTM performs better with text-only inputs

	- Other models perform better when additional parameters are added
	4 Related Work
	Deception Detection
	- Psycholinguistic work postulates that certain speech patterns indicate lying or hiding the truth
	    - Hedge words can add indirectness to hide meaning
	- Many studies relating to "Linguistic aspects deception detection" in NLP applications
	    - People lie with intent for some sort of payoff
	Fact-Checking and Fake News
	- Political fact checking: research in how much it helps people's awareness
	- Study unique linguistic styles in clickbait articles (Biyani et al. (2016))
	- The characterization of hoax documents on Wikipedia (Kumar et al. (2016))
	- Differences between fake news types also in previous work
	- Work in paper expands on this work by providing quantitative data
	Conclusion
	- See how truthfulness of news articles and public statements are indicated by vocabulary
	- Predict truthfulness of statements and analysis of types of fake news i.e. satire, propaganda, hoaxes
	- Fact checking = hard, but analysis of linguistic characteristics can increase understanding of differences of fake vs real news

Research Question/Problem/ Need	How does the vocabulary used in news or statements indicate the truthfulness of said news or statement?

Important Figures

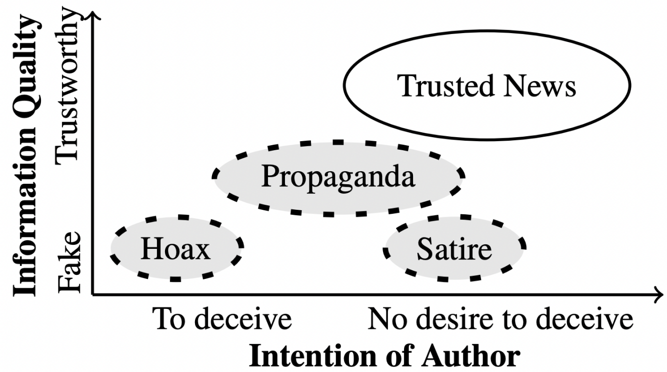


Figure 2: Types of news articles categorized based on their intent and information quality.

Lexicon Markers	Ratio	Example Source	Example Text	Max
Swear (LIWC)	7.00	Borowitz Report	... Ms. Rand, who has been **damned** to eternal torment ...	S
2nd pers (You)	6.73	DC Gazette	**You** would instinctively justify and rationalize **your** ...	P
Modal Adverb	2.63	American News	... investigation of Hillary Clinton was **inevitably** linked ...	S
Action Adverb	2.18	Activist News	... if one **foolishly** assumes the US State Department ...	S
1st pers singular (I)	2.06	Activist Post	**I** think its against the law of the land to finance riots ...	S
Manner Adverb	1.87	Natural News	... consequences of **deliberately** engineering extinction.	S
Sexual (LIWC)	1.80	The Onion	... added that his daughter better not be **pregnant**.	S
See (LIWC)	1.52	Clickhole	New Yorkers ... can bask in the **beautiful image** ...	H
Negation(LIWC)	1.51	American News	There is **nothing** that outrages liberals more than ...	H
Strong subjective	1.51	Clickhole	He has one of the most **brilliant** minds in basketball.	H
Hedge (Hyland, 2015)	1.19	DC Gazette	As the Communist Party USA website **claims**...	H
Superlatives	1.17	Activist News	Fresh water is the single **most** important natural resource	P
Weak subjective	1.13	American News	... he made that **very clear** in his response to her.	P
Number (LIWC)	0.43	Xinhua News	... **7 million** foreign tourists coming to the country in **2010**	S
Hear (LIWC)	0.50	AFP	The prime minister also **spoke** about the commission ...	S
Money (LIWC)	0.57	NYTimes	He has proposed to lift the state **sales tax** on groceries	P
Assertive	0.84	NYTimes	Hofstra has **guaranteed** scholarships to the current players.	P
Comparitives	0.86	Assoc. Press	... from fossil fuels to **greener** sources of energy	P

Table 2: Linguistic features and their relationship with fake news. The ratio refers to how frequently it appears in fake news articles compared to the trusted ones. We list linguistic phenomena more pronounced in the fake news first, and then those that appear less in the fake news. Examples illustrate sample texts from news articles containing the lexicon words. All reported ratios are statistically significant. The last column (MAX) lists compares which type of fake news most prominently used words from that lexicon (P = propaganda, S = satire, H = hoax)

How certain types of language are used in fake news. Ratio is how much type of language is used in fake news compared to real news. Last column is the type of fake news that vocab is used the most.

<table>
<tr>
<td></td>
<td>
<table>
<tr><th></th><th colspan="3">More True</th><th colspan="3">More False</th></tr>
<tr><th></th><th>True</th><th>Mostly True</th><th>Half True</th><th>Mostly False</th><th>False</th><th>Pants-on-fire</th></tr>
<tr><td>6-class</td><td>20%</td><td>21%</td><td>21%</td><td>14%</td><td>17%</td><td>7%</td></tr>
<tr><td>2-class</td><td colspan="3">62%</td><td colspan="3">38%</td></tr>
</table>
Table 4: PolitiFact label distribution. PolitiFact uses a 6-point scale ranging from: True, Mostly True, Half-true, Mostly False, False, and Pants-on-fire False.

How labels on Politifact are distributed based on both a 6-point and 2-point system
</td>
</tr>
<tr>
<td>VOCAB: (w/definition)</td>
<td>Hedging: avoid making a definite decision, statement, or commitment
Welsch t-test: (statistics) test to compare the means of two groups independent for each other that have a normal (bell-curve) distribution
Facetious: treating a serious situation with inappropriate humor
Hearsay: information that is unable to be reliably confirmed; a rumor
Entailment: deduction / implication</td>
</tr>
<tr>
<td>Cited references to follow up on</td>
<td>https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/ (Article on how Politifact work, which was the inspiration for this paper)</td>
</tr>
<tr>
<td>Follow up Questions</td>
<td>How useful is the additional nuance of a multi-point scale of truthfulness as opposed to a binary one?
What types of fake news are the most dangerous?
This article does not distinguish fake news created for monetary value and fake news created to spread an agenda. What are the language differences between those two categories?
Is assessing the truthfulness of satire particularly useful?</td>
</tr>
</table>

Grant #1 Notes: Prediction of social media postings as trusted news or as types of suspicious news

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Prediction of social media postings as trusted news or as types of suspicious news
Source citation (APA Format)	VOLKOVA, S. (2021). *Prediction of social media postings as trusted news or as types of suspicious news* (United States Patent No. US11074500B2). https://patents.google.com/patent/US11074500B2/en?q=misinformation+fake+news&oq=misinformation+fake+news
Original URL	https://patents.google.com/patent/US11074500B2/en?q=misinformation+fake+news&oq=misinformation+fake+news
Source type	Patent
Keywords	Neural network, social media, prediction

Summary of key points + notes (include methodology)	- Spread of fake info on social media can have serious impacts in the real world
	- False info varies based on intent
	- False info tends to fabricate stories instead of giving facts
	- Suspicious news content:
	    - Disinformation: false facts to deceive reader or to convince of a biased agenda
	    - Misinformation posts promoted or generated from propaganda
	    - Clickbait: eye catching headlines
	- Intent:
	    - Propaganda and clickbait: opinion manipulation, attention redirection, monetization, traffic attention
	    - Hoaxes: deceive reader
	    - Satire: NOT meant to deceive, rather to entertain and criticize, but can still be harmful
	    - “Massive digital disinformation” listed as one of the main risks to modern society in World Economic Forum Report

	Summary
	- Patent details Systems, computer implemented methods, and computer readable non-transitory storage media to predict if social media posts are trusted or suspicious
	- All embodiments have high accuracy and low error
	- Some embodiments
	- Some records have data on more than one language
	- Some cases use a neural network that uses parameters like text representation, linguistic markers, and user representation
	- Some use preset labels that are based on different types of fake news
	
	Detailed Description
	- Deception Detection often relies on:
	    - Hand-engineered features
	    - Shallow linguistic features
	    - Network features
	    - User behavior
	- Adding grammatical and syntactical features does not improve accuracy, but patterns of social interactions between users does
	- Combining multiple embodiments improves accuracy detection as a whole
	
	Examples and Comparisons
	- Satire and hoaxes can be harmful if they are shared out of context
	- Suspicious news often is created to build a narrative rather than report facts
	- Disinformation: false information spread to deceive
	- Conspiracy: belief that some sort of organization is responsible for an event
	- Propaganda: deliberate spread of misinformation
	- Hoax: mislead for political or financial gain
	- Clickbait: taking true stories but then making up details about those stories
	- Collected data from twitter one week before and after the Brussels bombing from suspicious and trusted news accounts
	
	- Neural network was used to sort twitter news accounts into 4 categories: propaganda, hoax, satire, clickbait
	- Input parameters:
	    - Tweet text
	    - Social graph: network of users associated with a social media post
	    - Linguistic markers of bias and subjectivity
	    - Moral foundational signals
	- Bias cues:

	- Hedges, assertive verbs, factive verbs, implicative verbs, report verbs
	- Features represented by using vector representations
	- Subjectivity Cues:
	  - strongly and weakly subjective words and positive and negative opinion words
	- Psycholinguistic Cues:
	  - Various categories in LIWC
	- Neural networks work better than logistic regression baselines
	- Syntax and grammar features can actually lower accuracy in some cases: mostly due to the unique language and length of tweets
	Linguistic analysis:
	- Verified news
	  - Less bias markers, hedges, and subjective terms
	  - Less harm/care, loyalty/betrayal, and authority moral cues
	  -
	- Satire is the most different from propaganda and hoaxes
	- Propaganda, hoaxes, and clickbait are the most similar
	Clickbait Score Prediction
	- Intent of clickbait: attention redirection, MONEY, traffic attraction
	- Using a regression model to get a clickbait score from 0 to 1
	- Not using handcrafted features; instead using a neural network to get machine trained features
	- A Clickbait challenge for this regression task provided a few datasets with labeled and unlabeled data
	- Content is a large factor that humans use to judge if content is clickbaity
	- Compared the performance of inputs containing:
	  - Only text of post
	  - Only text of article
	  - The post and the article
	- Linguistic cues were also added
	- Higher performing models used fewer epochs
	- Also developed models trained with noisy labels

Research Question/Problem/ Need	Suspicious news spreads very quickly and easily on social media, so we need a way to detect it on social media reliably.

Important Figures

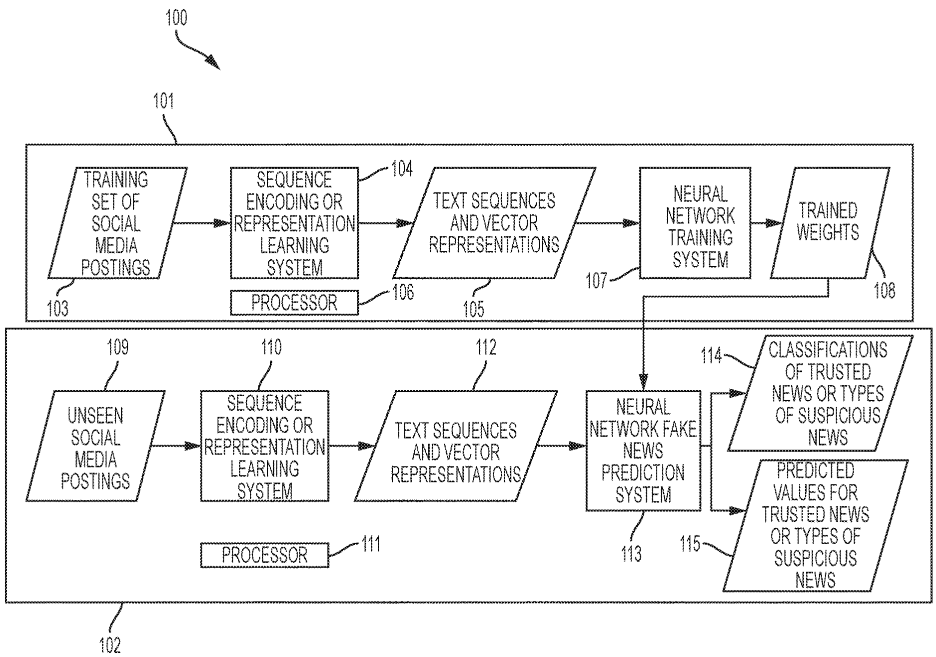


Figure 1 details a system that is able to predict if social media posts are trusted news or suspicious news

TABLE 1

TYPE	NEWS	POSTS	RTPA	EXAMPLES
Propaganda	99	56,721	572	ActivistPost
Satire	9	3,156	351	ClickHole
Hoax	8	4,549	569	TheDcGazette
Clickbait	18	1,366	76	chroniclesu
Verified	166	65,792	396	USATODAY

Twitter dataset statistics: news accounts, posts and retweets per account (RTPA).

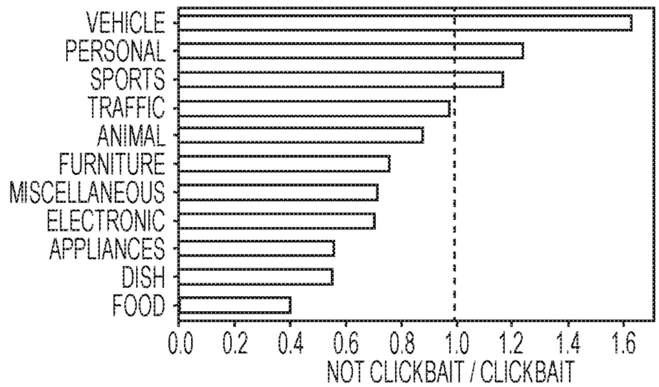


FIG. 5

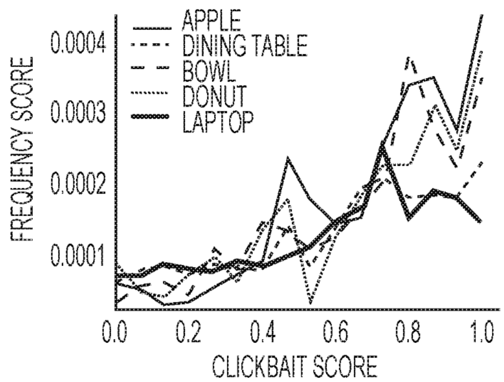


FIG. 6A

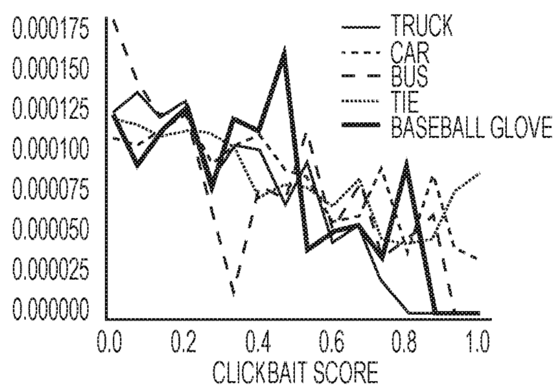


FIG. 6B

Charts and graphs show how certain objects in a thumbnail can indicate clickbait

- If there are food objects of electronic objects, the clickbait score increases

VOCAB: (w/definition)	Embodiment: a tangible or visible form of an idea, quality, or feeling.

	Schematic: a 2 dimensional representation of how components of something interact Moral foundations: def here https://moralfoundations.org/ Vector representation: "A vector is a tuple of one or more values called scalars." Epoch: the total number of iterations of all training data to train a machine learning model Noisy labels: when labels in a dataset are not 100% accurate
Cited references to follow up on	n/a
Follow up Questions	How do images indicate that something is fake news? What do we define as clickbait? Is clickbait always misleading? How do noisy labels improve/help train a machine learning model?

Grant #2 Notes: Machine learning to identify opinions in documents

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Machine learning to identify opinions in documents
Source citation (APA Format)	Dadachev, B., & Papineni, K. (2020). *Machine learning to identify opinions in documents* (United States Patent No. US10832001B2). https://patents.google.com/patent/US10832001B2/en?q=misinformation+fake+news&oq=misinformation+fake+news
Original URL	https://patents.google.com/patent/US10832001B2/en?q=misinformation+fake+news&oq=misinformation+fake+news
Source type	Patent
Keywords	Machine learning, opinion detection, nlp

Summary of key points + notes (include methodology)	- Machines currently only able to understand very little about content of news articles
	- Most approaches do not utilize a deep understanding of news content
	- Current work
	    - Subjectivity detection
	        - Often uses lexicons, which are limiting
	        - Subjectivity doesn't necessarily show anything about article content
	    - Sentiment analysis
	        - Trying to find the viewpoint of the author on a topic
	        - Does not provide info about what the article is actually about
	    - Stance detection
	        - Only detects if an article is for or against a predetermined topic
	        - Only viable for initially known topics
	Summary

	- Machine learning algorithm that can determine if statements in an article can be classified as opinion or fact
	Detailed Description
	- Two main components:
	    - Machine learning opinion classification model: see if portions of document are opinionated or not
	    - Summarization algorithm : ranks portions of a document by importance
	    - Many ways for these two parts to interact
	- This sorting can result in less time wasted reading articles and saving resources to load articles
	- Sometimes the opinion of an author is directly written in the article but in other points it can be less evident (ie sarcasm)
	- News articles often can be put into two different types:
	    - Neutral retelling of events
	    - Opinions of these events
	- Can be used to filter out parts of article with no substance
	- How opinionated an article is is heavily dependent on the topic and context of it
	- Neural networks often used
	- Classification models:
	    - Some use binary classification “opinion” “not opinion”
	    - Other have multi-class:
	        - “Reported opinion”: opinion of someone that is not the author themselves
	        - “Author opinion”
	        - “May be opinion”
	        - “Mixed fact and opinion”
	    - Output a classification score that can then be labeled
	- Viable training sets
	    - Opinion pieces from news corpus with opinion labels
	    - Documents with individual sections labeled
	        - This one is more effective but requires more resources
	- “Therefore, the user can avoid reading articles which feature redundant opinions, thereby again conserving…” I think this thought process is flawed because the vast majority of people only want to hear that their own opinion or assumption is right
	- Basically being able to accurate detect and present the main opinion of a article can waste less processing power
	
	- Computer program that can
	    - Be given a part of a text and classify it as opinion or not opinion
	    - Have a confidence score with that classification
	    - Have each portion of a article be assigned a confidence score to show if that portion can represent

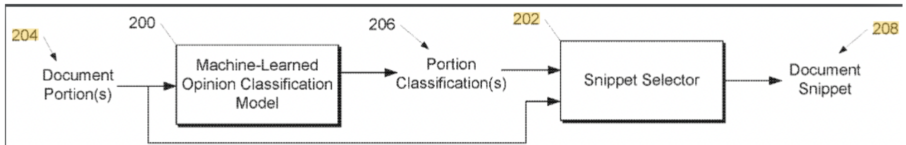
<table>
<tr>
<td></td>
<td>
the whole article
- Have one of the “opinion” sections of the text be shown in a informative display
- Have imputed feature data like
    - Lexicon data
    - Topic data
    - Content type
    - Surrounding content data
    - Story content data
    -
</td>
</tr>
<tr>
<td>Research Question/Problem/ Need</td>
<td>We need an effective way to summarize the viewpoints of an news article with a program.</td>
</tr>
<tr>
<td>Important Figures</td>
<td>


Figure 2

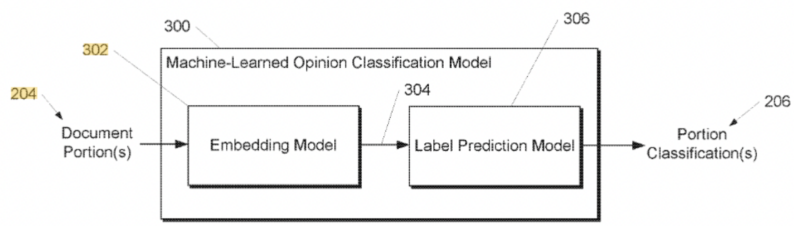
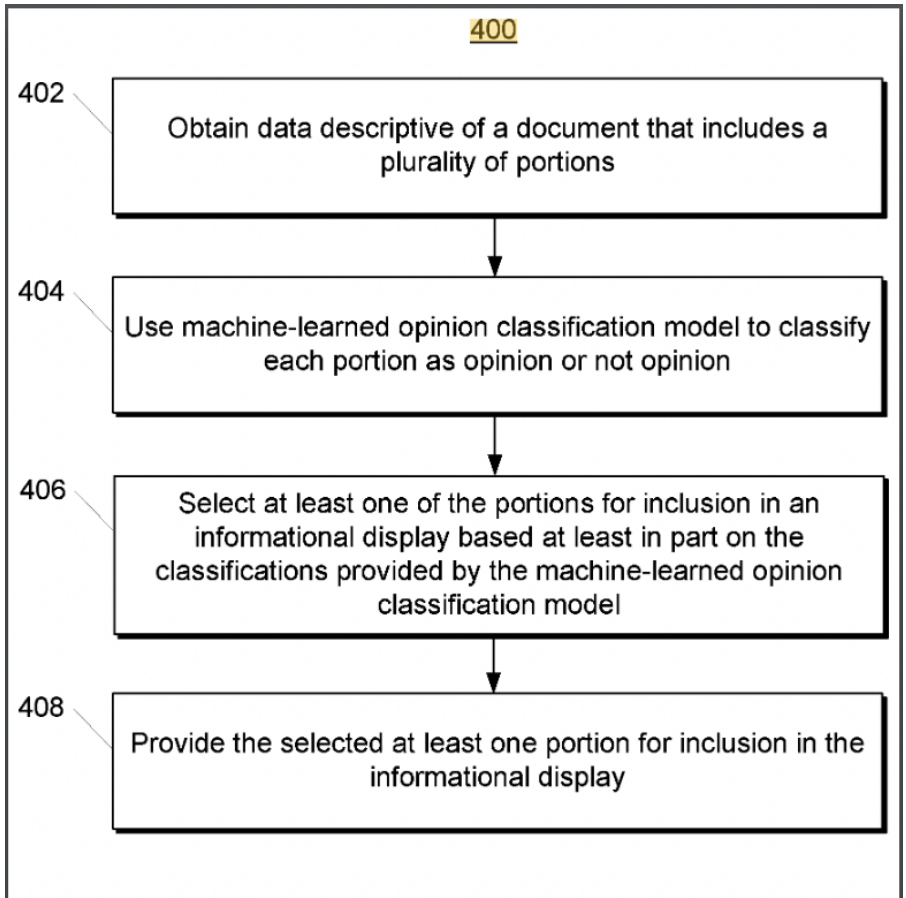


Figure 3
</td>
</tr>
</table>

	 Figure 4
VOCAB: (w/definition)	Plurality: to have more than one of Transitory: not permanent Snippet: a small part of a whole Recurrent: occurring often or repeatedly.
Cited references to follow up on	n/a
Follow up Questions	Is lexical data sufficient for opinion detection? Would opinion detection be an adequate strategy to use when detecting fake news? Is there any way for the computer model itself to understand what the opinion of the article is?

Article #13 Notes: Google Finds ‘Inoculating’ People Against Misinformation Helps Blunt Its Power

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Google Finds ‘Inoculating’ People Against Misinformation Helps Blunt Its Power
Source citation (APA Format)	Grant, N., & Hsu, T. (2022, August 24). Google Finds ‘Inoculating’ People Against Misinformation Helps Blunt Its Power. *The New York Times*. https://www.nytimes.com/2022/08/24/technology/google-search-misinformation.html
Original URL	https://www.nytimes.com/2022/08/24/technology/google-search-misinformation.html
Source type	General Article
Keywords	Online misinformation, Google, internet, pre-bunking, falsehoods
Summary of key points + notes (include methodology)	This article mainly details a study by Google, the University of Cambridge and the University of Bristol. In the study, the researchers try to test a method to prevent people from getting tricked by misinformation before they are even exposed to it. They used Google ad space to get people to watch videos related to misinformation techniques and how not to buy into them. It was shown that people who watched the video improved their ability to detect misinformation techniques by 5%. Overall, fact-checkers can only do so much, so it's important that the general public is informed about misinformation. - Falsehoods show up quickly and also spread quickly - It also takes time for fact checkers to debunk them - Researchers trying an approach to undermine misinformation before people see it, calling it “pre-bunking” - Found that showing videos about the tactics of misinfo to people make them more skeptical of misinfo afterwards

	- However, this may not work for people with extreme and hardened political beliefs
	- Tech companies have struggled to strike a balance for fighting against misinfo and lies without getting to the levels of censorship
	  - Though companies can help to address the problem, it is ultimately the user's job to differentiate because fact and fiction
	- Strategies used during midterm vote on social media:
	  - Partnering with fact-checking groups
	  - Warning labels
	  - Portals with vetted explainers
	  - Post removal
	  - User bans
	- Attempts have been made to prevent spread of misinfo, but they are not effective
	- Paper has 7 experiments with 30000 total participants
	- Used Youtube ad space to show users in the US 90-second animated videos meant to teach them about propaganda and misinfo techniques
	  - One million adults ended up viewing the ad for 30 seconds or longer
	  - Topics taught include:
	    - Scapegoating
	    - Deliberate incoherence
	    - Conflicting explanations to declare truthfulness
	  - Some participants within 24 hours of seeing the video were tested, found 5% increase in ability to know misinfo techniques
	  - One video starts with a girl holding a teddy bear with sad music and a narrator saying "What happens next will make you tear up", then proceeds to explain how emotional manipulation contributes to spread of false info
	- One of paper's authors says that pre-bunking played into a desire for people to not be tricked, then commenting that it was one of the few studies that worked on all the political spectrum
	- However, pre-bunking was not as effective for those with extreme political beliefs
	- Elections are also difficult to pre-bunk since their beliefs in regards to that are much more deep and difficult to change
	- Prebunking is also not a long term solution: effects lasts for only a few days to a month
	- Many other attempts at pre-bunking by other groups:
	  - Misinformation-identifying curriculum over two weeks
	  - Lists with tips to how to identify misinfo

	- Online games to help detect misinfo - A study in 2020 found that people that played online game Bad News could recognize common misinfo strats across cultures - Pre-bunking compared to vaccines: warning and weakened doses of misinfo can develop protection against real misinfo - Hard part of fighting misinfo is not known what rumor or conspiracy will spread next, but they follow a predictable pattern - Fact checkers can only do so much, so the general public needs to be taught trends in misinfo in order to not be taken advantage by it
Research Question/Problem/ Need	If people are informed about the tactics of misinformation and how it works, will they be more skeptical of falsehoods?
Important Figures	n/a
VOCAB: (w/definition)	Inoculating: vaccination silver bullet: a simple, seemingly magical, solution to a difficult problem, a pancrea Portals: A portal is a web-based platform that collects information from different sources into a single user interface and presents users with the most relevant information for their context. Fear-mongering: the action of deliberately arousing public fear or alarm about a particular issue.
Cited references to follow up on	https://www.science.org/doi/10.1126/sciadv.abo6254 (main research article discussed in study)
Follow up Questions	This technique of pre-bunking, at least, does not seem to be a permanent solution. What are solutions that have more long term effects? What are effective ways to draw people into commercials? (since that seemed to work here) How are we able to un-condition people that have extreme political views that are difficult to change? How can we teach machines the same principles of misinformation that humans are taught?

Article #14 Notes: A Gentle Introduction to Natural Language Processing

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	A Gentle Introduction to Natural Language Processing
Source citation (APA Format)	Vijay, R. (2022, July 28). *A Gentle Introduction to Natural Language Processing*. Medium. https://towardsdatascience.com/a-gentle-introduction-to-natural-language-processing-e716ed3c0863
Original URL	https://towardsdatascience.com/a-gentle-introduction-to-natural-language-processing-e716ed3c0863
Source type	General Article
Keywords	Natural Language Processing, Sentiment analysis

Summary of key points + notes (include methodology)	- Natural language processing is about making computers be able to learn and process human language
	- Types of NLP
	    - Machine Translation
	    - Natural Language generation
	    - Web search
	    - Spam filters
	    - Sentiment analysis
	    - Chatbots
	    - Etc
	- Data cleaning: removing unwanted symbols from text that the machine doesn't need to worry about
	- Preprocessing data: generally means transferring data into an understandable format
	    - Making all text lowercase
	    - Tokenization: splitting all text into individual words called tokens
	    - Stop words removal:

	- Stop words are words that do not contribute information to a document - NLTK has an inbuilt stop words list but it does not work for all situations - Can also build own stop words list - Stemming: reducing a word to it's stem/root word - Ex: love, loving, loved and all be reduced to the root love - Stems sometimes are not a word in the language ("movi" is the root for "movie") - Lemmatization: same as stemming but every stem is a valid word in the language - N-grams: combination of words used together - N = 1: unigrams (individual words) - N = 2, bigrams (two words) - N = 3, trigrams (three words) - And so on and so forth… - Used to preserve sequence information - Text data vectorization: converting text into numbers so that they can be used by algorithms - Bag of words (BOW) - Two sentences said to be similar if they have similar sets of words - BOW makes dictionary of unique words in the corpus provided - If word is present in a document, set the value to one, else set as 0 - Creates a matrix - Natural language toolkit (NLTK): open source library for NLP tasks - Syntax to import: `!pip install nltk` - Terms: - Text sentence: "document" - Collection of documents: "text corpus"
Research Question/Problem/ Need	What are the basics of NLP and how can you create a basic AI that can sort IMDB reviews?
Important Figures	n/a
VOCAB: (w/definition)	Token: a single element in a programming language Sentiment Analysis: NLP technique to determine of data is positive, negative, or neutral Stratify: arrange or classify Naive Bayes: probabilistic machine learning model that's used for classification task

Cited references to follow up on	n/a
Follow up Questions	What type of NLP should I use for my project? Will I need to combine multiple subcategories of NLP for my project? If I will need to do multiple tasks for my project, how would I do so? Will sentiment analysis be useful for my project or are there other tasks that are more important?

Article #15 Notes: A Survey on Natural Language Processing for Fake News Detection

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	A Survey on Natural Language Processing for Fake News Detection
Source citation (APA Format)	Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A Survey on Natural Language Processing for Fake News Detection. *Proceedings of the 12th Language Resources and Evaluation Conference*, 6086–6093. https://doi.org/10.48550/ARXIV.1811.00770
Original URL	https://doi.org/10.48550/arXiv.1811.00770
Source type	Research Article / review article
Keywords	NLP, fake news detection, automation

Summary of key points + notes (include methodology)	- Automated fake news detection = determining the truthness of claims in news
	- Relatively recent NLP problem but important since news has large social-political impacts in society
	- Detection of fake news is important and something that NLP can help with
	- Conventional solution: have experts manually check claims against evidence (ie PolitiFact)
	    - However, time consuming and expensive
	    - Cannot catch up with the spd in which fake news is being produced
	- Paper gives overview on automated fake news detection using NLP
	    - States challenges with fake news detection and Machine Learning solutions that solve this problem
	- Contributuions of the paper:
	    - First comprehensive review of NLP solutions for automated fake news detection
	    - See how fake news detection relates to existing NLP tasks
	    - Summarize the available datasets, NLP approaches, and results to guide new researchers
	Related Problems
	- Fact Checking

	- Assessing truthfulness of claims made by public figures
	- Many researchers do not distinguish between fake news detection and fact checking
	- Fake news detection focuses on news events, fact-checking is more broad
	- Rumor Detection
	  - No concrete definition
	  - (Zubiaga et al., 2018) defines it as separating statements as rumor or non-rumor
	  - Rumor: statement containing unverified information
	  - Rumor statement must have info that can be verified as opposed to subjective feelings
	  -
	- Stance Detection
	  - Finding what stance an author takes on an issue based on text
	  - Can be a subtask for fake news detection
	  - Not based on fact checking, more based on consistency
	- Sentiment Analysis
	  - Extracting the emotions of someone from text
	  - Not about accuracy of claim, instead about emotions expressed (possibly about the claim)
	Task Formulations
	-
	- Classification
	  - Most common strat
	  - Most classification methods are binary
	  - However, not all news in completely true or completely false
	  - Getting reliable labels can be difficult for training data
	- Regression
	  - Output is numeric score of truthfulness
	Datasets
	- Claims
	  - Might use but is not my main focus
	  - PolitiFact, Channel4.com, and Snopes: all contain manually labeled short claims in news
	- Entire-Article Datasets
	  - The dataset that I probably want for my project
	  - Fakenewsnet: ongoing data collection project for fake news research
	    - Consists of headline and body text of fake news articles based on PolitiFact and Buzzfeed
	    - Also contains data for social engagement on Twitter
	  - BS detector: data collected from a browser extension of the same name
	    - Searches all links on webpage to references to unreliable sources based on a manually made list of unreliable domains

- Posts On Social Networking Sites
    - Probably irrelevant to my project so will skip

Methods

- Preprocessing
    - Basically organizing text into a format that the computer is able to understand
    - Term Frequency-Inverse Document Frequency (TF-IDF) and Linguistic Inquiry and Word Count (LIWC) often used for converting tokenized text into features
    - Word sequences: word2vec and GloVe most often used
    - When an entire article is input:
        - Another step is finding the central claims
        - "Thorne et al. (2018) rank the sentences using TFIDF and DrQA system (Chen et al., 2017)"
- Machine Learning Models
    - Majority of research uses supervised models, others are less commonly used
    - Discussion mainly about classification models
    - **Non-neural network**
        - **Support Vector Machine (SVM)** and **Naive Bayes Classifier (NBC)** most commonly used
            - Usually used as baseline models
        - **Logistic regression** and **decision tree (like Random Forest Classifier)** also sometimes used
    - **Neural network**
        - **Recurrent Neural Networks (RNN)** like **Long Short-Term Memory (LSTM)** is popular in NLP
            - Works better when it comes to longer sentences
        - **Convolutional neural networks (CNN)**: good at many text classification tasks
            - Used to extracting figures with lots of different metadata
        - **Multi-source Multi-class Fake news Detection framework (MMFD):** combination of CNN and LSTM
            - CNN looks at patterns in text
            - LSTM looks at temporal dependencies in the text
            - Concatenation of outputs put through Fully Connected Network
        - Attention mechanisms
            - Put in neural networks for better performance
    - **Rhetorical Approach**
        - **Rhetorical Structure Theory (RST),** sometimes combined with **the Vector Space Model (VSM)**
        - **RST:** analytical framework for coherence of a story
        - Identified via text coherence and structure
    - **Collecting Evidence**

	- **RTE-based (Recognizing Textual Entailment)** used to gather evidence
	- Finding relationships between sentences
	- Textual evidence is needed for fact checking
	- **therefore , can only be used if dataset has evidence**
	-
	Results & Observations
	- Focuses on 3 datasets: LIAR, FEVER, and FAKENEWSNET
	- The other databases mentioned are more useful for rumor detection
	- Accuracy of models using LIAR
	  - LSTM more accurate than CNN
	  - One study added verdict reports, which raised accuracy by 4%
	  - Another study improve the performance of LIAR by 21% by replacing credibility history with speaker2credit, a larger credibility source (https://github.com/akthesis/speaker2credit)
	  - “The two papers also show the attention scores for verdict reports/speaker credit are higher than the statement of claim”
	- Accuracy of models using FEVER
	  - Models often use attention based methods with FEVER
	  - Best at verification and evidence-collection
	- Accuracy of models using FAKENEWSNET
	  - Many methods rely on social engagement data
	  - Performs the best when included with additional data
	Discussions & Recommendations
	- Requirements for a fake news corpus:
	  - 1. Availability of both truthful and deceptive instances;
	  - 2. Digital textual format accessibility;
	  - 3. Verifiability of “ground truth”;
	  - 4. Homogeneity in lengths;
	  - 5. Homogeneity in writing matters;
	  - 6. Predefined timeframe;
	  - 7. The manner of news delivery;
	  - 8. Pragmatic concerns;
	  - 9. Consideration for language and culture differences
	
	- New recommendations for datasets that expand on ones used before:
	  - Not practical to categorize with just “true” or “false”
	  - More choices tend to make ordinary people lead to similar conclusions as experts
	  - Binary classification at this point is pretty accurate
	  - Next step: classifying news in more categories than binary
	  - Many multi-class models do not consider the order of labels (ie. classifying a true article as false is more wrong than saying a true article is mostly true)
	    - Many models do not account for the example
	    - Develop possible method to keep track of this?
	    - Did actually consider this: make it more of a priority?

<table>
<tr><td></td><td>

Quote claims or articles from various speakers and publishers within the scope of dataset

Many types of fake news: some have harmful intent, others have more innocuous reasons
Satire can be distinguished well from both real and fake news via style analysis
Cannot assume that a domain or a publisher only provides either real or fake news
Data should be diverse and have different writing styles

Validate Entire Article

Claims are easier to analyze
“As a future task, we should consider how to evaluate the truthfulness of the entire-article and annotate them. For example, it may be preferable to add truthfulness scores to individual statements.”

Find the individual claims made in a full article and then see how accurate those claims are? Finding a sufficient dataset for this could be hard though

Common Models Critiques

Hand-crafted featured needed for non-neural network approaches, but could also use neural networks
“However, these hand-crafted features seem to learn something that is more useful and cannot be combined with hand-crafted features” what does that mean
“For example, Rashkin et al. (2017) shows that adding LIWC did not improve the performance of the LSTM model while non-neural network models are improved largely on their dataset.”
“However, relying too much on speakers’ or publishers’ information for judging may cause some problems”

</td></tr>
<tr><td>Research Question/Problem/ Need</td><td>A general summary of the process and gaps in fake news detection using a NLP approach.</td></tr>
<tr><td>Important Figures</td><td>
<table>
<tr><th>Name</th><th>Main Input</th><th>Data Size</th><th>Label</th><th>Annotation</th></tr>
<tr><td>LIAR</td><td>short claim</td><td>12,836</td><td>six-grade</td><td>editors, journalists</td></tr>
<tr><td>FEVER</td><td>short claim</td><td>185,445</td><td>three-grade</td><td>trained annotators</td></tr>
<tr><td>BUZZFEEDNEWS</td><td>FB post</td><td>2,282</td><td>four-grade</td><td>journalists</td></tr>
<tr><td>BUZZFACE</td><td>FB post</td><td>2,263</td><td>four-grade</td><td>journalists</td></tr>
<tr><td>SOME-LIKE-IT-HOAX</td><td>FB post</td><td>15,500</td><td>hoaxes or non-hoaxes</td><td>none</td></tr>
<tr><td>PHEME</td><td>Tweet</td><td>330</td><td>true or false</td><td>journalists</td></tr>
<tr><td>CREDBANK</td><td>Tweet</td><td>60 million</td><td>30-element vector</td><td>workers</td></tr>
<tr><td>FAKENEWSNET</td><td>article</td><td>23,921</td><td>fake or real</td><td>editors</td></tr>
<tr><td>BS DETECTOR</td><td>article</td><td>-</td><td>10 different types</td><td>none</td></tr>
</table>
Table 1: A Summary of Various Fake News Detection Related Datasets. FB: FaceBook.
</td></tr>
</table>

Author	Meta-data	Base Model	Acc.
Wang		SVMs	0.255
		CNNs	0.270
	+Speaker	CNNs	0.248
	+All	CNNs	**0.274**
Karimi		MMFD	0.291
	+All	MMFD	**0.348**
Long		LSTM+Att	0.255
	+All	LSTM(no Att)	0.399
	+All	LSTM+Att	**0.415**
Kirilin	+All	LSTM	0.415
	+All+Sp2C	LSTM	**0.457**
Bhatta-charjee	2-class label	NLP Shallow	0.921
		Deep (CNN)	0.962

Table 3: The Current Results for LIAR. +All means including all meta-data in LIAR. Bhattacharjee convert 6-class labels to 2-class labels.

Author	Model	Acc.
Thorne	Decomposable Att	**0.319**
		0.509
Yin	TWOWINGOS	**0.543**
		0.760
Hanselowski	LSTM (ESIM-Att)	**0.647**
		0.684
UNC-NLP Nie	Semantic Matching Network (LSTM)	**0.640**
		0.680

Table 4: The Current Results for FEVER. The results in boldface are the accuracy of evidence-collection task.

Author	Data	Model	Acc.
Shu	Buzz Feed	RST	0.610
		LIWC	0.655
		Castillo	0.747
		TriFN	**0.864**
Della		HC-CB-3	**0.856**
Deligiannis		GCN	**0.944**
Shu	Politi Fact	RST	0.571
		LIWC	0.637
		Castillo	0.779
		TriFN	**0.878**
Deligiannis		GCN	**0.895**
Della		HC-CB-3	**0.938**

Table 5: The Current Results for FAKENEWSNET. There are two sources of data separately: BuzzFeed and PolitiFact.

VOCAB: (w/definition)	Veracity: conformity to facts; accuracy classification models: model that reads input and generates an output to

	classify the input into a category baseline models: simple model that acts as a baseline/reference for a machine learning project temporal dependencies: the impact of previous behavior on current behavior. concatenation: a series of interconnected things or events. regression
Cited references to follow up on	Nakashole, N. and Mitchell, T. M. (2014). Language-aware truth assessment of fact candidates. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1009–1019. **(used a regression model)**
Follow up Questions	How could we be able to calculate the accuracy of multi-class classification methods without making the accuracy look worse than it actually is? What are some ways to track claims in full articles? How can we connect a database of claims to articles? Why are some datasets more effective than others?

Article #16 Notes: Automated fake news detection using linguistic analysis and machine learning

Article notes should be on separate sheets

Source Title	Automated fake news detection using linguistic analysis and machine learning
Source citation (APA Format)	Singh, V., Dasgupta, R., Sonagra, D., Raman, K., & Ghosh, I. (2017, July). Automated fake news detection using linguistic analysis and machine learning. In *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation (SBP-BRiMS)* (pp. 1-3).
Original URL	http://sbp-brims.org/2017/proceedings/papers/challenge_papers/AutomatedFakeNewsDetection.pdf
Source type	(Extremely short) proof of concept research paper
Keywords	Fake News, Text Processing, Machine Learning

Summary of key points + notes (include methodology)	- Dataset used are
	    - "Kaggle Fake News": contains only articles that are false in some way: 345 articles randomly picked
	    - A created dataset of 345 "valid" articles from New York Times, National Public Radio, and the Public Broadcasting corporation
	- LIWC was used to obtain the linguistic features of the articles
	- 80% of data used for training and 20% for testing
	- Multiple Machine Learning models tested: Support Vector Model (SVM) seemed to be most accurate at 0.87
	- Different features between fake news and real news was also observed
	    - Are these percentage values?
	    - Fake news seems to use more authentic language (seems more honest and personal)
	- Shows that a linguistic approach to detecting fake news has potential
	- Using multiple features is useful
	- Contributions:

- Making a new real news dataset
- A machine learning model that it able to reach 87% accuracy
- Finding features that is associated with fake news

Research Question/Problem/ Need

How effective is automated fake news detection using linguistics?

Important Figures

Overall Accuracy	0.87			
Classification Report	**(Support Vector Machine)**			
	Precision	Recall	F1-Score	Support
Valid News	0.89	0.89	0.86	65
Fake News	0.86	0.90	0.88	73
Avg\total	0.87	0.87	0.87	138

I do not know where support came from or what it's supposed to mean.

	Mean (valid)	**Mean (fake)**	**Absolute Diff mean (credi-ble-fake)**	**Difference (Std. dev.)**
Word Count	1009.78	686.77	**323.00**	0.47
Authentic	16.72	24.04	**7.32**	0.47
Clout	76.49	70.37	**6.12**	0.53
Tone	42.25	36.33	**5.93**	0.24
Analytic	87.90	85.09	**2.81**	0.21

It's very unclear from this graph what these numbers mean. I'm assuming this is the mean average percent of the type of word used since that is how it is measured using LIWC, but the text itself does not specify this.
I'm also not really sure what the difference column means or if it is even useful information.

VOCAB: (w/definition)

Normalization: Transforming features in order for them to be a similar scale to each other.
Disclosing: more revealing
Precision: the fraction of true positions over the number of true positives and false positives
Recall: the fraction of true positives over the number of true positives and false negatives

Cited references to follow up on

n/a

Follow up Questions

Is the standard division calculated for all of the articles in the testing data or for each subset? (ie fake vs real)
Would different data comparison methods lead to different results?
Why do certain machine learning algorithms perform better with this

	task than others? Is looking at the difference in the means of the results really an accurate way of finding the differences of features between two groups?

Article #17 Notes: We Will Know Them by Their Style: Fake News Detection Based on Masked N-Grams

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	We Will Know Them by Their Style: Fake News Detection Based on Masked N-Grams
Source citation (APA Format)	Pérez-Santiago, J., Villaseñor-Pineda, L., & Montes-y-Gómez, M. (2022). We Will Know Them by Their Style: Fake News Detection Based on Masked N-Grams. In O. O. Vergara-Villegas, V. G. Cruz-Sánchez, J. H. Sossa-Azuela, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, & J. A. Olvera-López (Eds.), *Pattern Recognition* (pp. 245–254). Springer International Publishing. https://doi.org/10.1007/978-3-031-07750-0_23
Original URL	https://link.springer.com/chapter/10.1007/978-3-031-07750-0_23
Source type	Conference Paper
Keywords	N-Grams, Fake News detection, masking, machine learning, written style
Summary of key points + notes (include methodology)	- Fake news definition considered in paper: “fake news are news published by a media outlet, which includes: claims, statements, speeches, publications, among other types of information and its authenticity is not verifiable (false)” - Paper focused on analyzing fake news purely based on writing style, making it be able to be extended to various languages outside of english - The paper itself uses english and spanish. Do drastically different languages like chinese also work with this method?

- "Despite their promising results, these linguistic features are technically very demanding to be extracted, analyzed, understood and interpreted" in what way?
- Work done in this paper is not computationally expensive. Similar methods have been used for authorship analysis but not for automated fake news detection

Related Work

- Many other strategies for detection of fake news
- Style based detection is helpful in early stages of fake news detection when it hasn't been spread significantly
- Many have proposed using LIWC features
    - Likely approach for my project
- Most implementations need language and domain related resources
- **Approach in paper: mask semantic information and leave only lexical style patterns to avoid the above**
- Approach has been used for various other tasks but hasn't been done for fake news detection specifically
    - Called "text distortion"
- Motivation of fake news is to appeal to reader's emotions and beliefs, which is reflected in language used
    - Punctuations marks to emphasize personal opinions
    - Difference in use of numbers to provide reliability of info presented
    - Used to created "fine-grained masking strategy"

Style-Based Method for Fake News Detection

General order of process:

1. Set up list/set of tokens/words that will not be masked
    a. Terms associated with style (frequently used terms)
2. Any term not in set is masked while preserving the sequence of words in the text
3. N-grams of words extracted (que?)
4. BoW representation of document feed to trained classifier
5. Prediction

Selection of Lexicons

Selected from *k* highest frequent words from:

1. Most frequent words of the language
    a. Spanish words from "Current Spanish Reference Corpus" (CREA)
    b. English words from "British National Corpus" (BNC)
        i. Would there be a difference if an american dictionary was used instead
2. Most frequent words in the corpus
    a. Aka the words most frequently used in the new articles themselves in provided news datasets

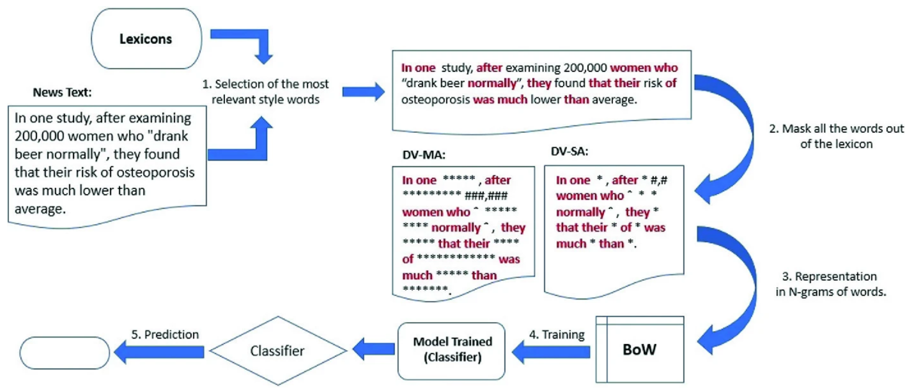
Text Masking (two strategies)

	Mask info related to news content while keeping writing style elements from previous section
	1. Distorted View with Multiple Asterisks (DV-MA)
	    a. Every word not part of the reference lexicon is hidden by having each **character of the word** replaced by an asterisk (*)
	    b. Each **digit** (of a number) is replaced with a pound symbol (#)
	2. Distorted View with Single Asterisks (DV-SA)
	    a. Every word not part of the reference lexicon is hidden by having each **singular word** replaced by an asterisk (*)
	    b. Each **number** (sequence of digits) is replaced by a pound sign (#)
	3. Other extra rules
	    a. Punctuations marks are kept (.,;:)
	    b. Smart quotes replaced with ^ symbol
	    c. Parentheses, braces, and brackets replaced with only open and closed parentheses
	        i. Aka ([{ replaced with just (,)]} replaced with just)
	    d. Exclamation and question marks (!?) replaced with mu (μ)
	    e. Mathematical signs like $%+= replaced with pi ($\pi$)
	Experiments
	Datasets (used in state-of-the-art works)
	1. Spanish datasets
	    a. MEX-A3T
	    b. RAW-CovidES
	2. English Datasets
	    a. LIAR
	    b. CoAID
	Observations of these datasets will be in Important Figures
	Experimental Setup
	1. Preprocessing
	    a. Text converted to lowercase, no characters removed
	2. Used parameters
	    a. k = # of words extracted from lexicon that will not be masked
	        i. Set as 100 to 1000 by increments of 100
	    b. n = length of n-grams of words
	        i. Set as 1, 2, or 3
	3. Text representation
	    a. TF-IDF weighting scheme
	4. Classifier
	    a. A Support Vector Machine (SVM) with linear kernel

	from Scikit-Learn library
	5. Training and Evaluation
	   a. LIAR
	      i. Already has preset partitions: Training, validation, and test
	   b. RAW-Covid and CoAID
	      i. CFV performed with 5-folds
	   c. MEX-A3T
	      i. 20% of test partition used of validation
	6. State of the Art
	   a. Results compared with best results from previous work for the 4 collections
	Results and Discussion
	- DV-MA always more effective than DV-SA
	   - Preserving length of words helped classifier
	- Best results occurred when # of words not masked (k) was above or equal to 500
	- Majority of state of the art uses neural networks while this work does not
	Discriminative Style Patterns
	- In spanish, long sequences of * had large GI values, mostly associated with adverbs ending in “mente”
	- In english, short strings are highlighted, mostly associated with abbreviations (“gov”, “rep”, “nov”, “dec”)
	- Numerical data in fake news mainly for dates and ages
	- Numerical data in real news used more for statistical data, percentages, and monetary amounts
	   - Writers of real news are not afraid to show data that verifies their information, unlike those of fake news
	- Real news is often longer than fake news, so punctuation is naturally used more in real news
	- Quotations used in both but for differing reasons
	   - Used more in fake news

Research Question/Problem/ Need	Can the written style of news be utilized for a text classification task to determine if it is real or fake?

Important Figures



Flowchart detailing the steps of the model.

Datasets	Domain	Language	Fake	True	Total
MEX-A3T	Multiple	Spanish	480	491	971
RAW-Covid	Health		95	105	200
LIAR	Politics	English	6889	8470	15359
CoAID	Health		185	3167	3352

Table detailing the datasets used. English datasets seem to be bigger than spanish ones at first glance. Two of the datasets are dedicated to COVID-19.
LIAR in particular only seems to contain statements from Politifact as opposed to full articles. Since LIAR doesn't use a binary classification of "true" or "false", it's made unclear how the authors split the data.
CoAID seems to have a very limited set of fake news articles. It seems to contain data for both full articles and claims from politifact or other fact checking sites.
MEX-A3T seems to be a dataset that contains articles in Mexican Spanish that are labeled either true or false from various web sources and were manually labeled. It covers 9 different topics: Science, Sport, Economy, Education, Entertainment, Politics, Health, Security, and Society.
RAW-Covid also contains full articles but specifically pertaining to health related issues. The name used in the article is misleading; not all the articles pertain to COVID-19 specifically.

Overall, the English dataset, though seemingly containing more data, do not have many full articles of various different niches. LIAR in particular is made up entirely out of claims. The spanish datasets, in contrast, are all entirely made up of full news articles.

Datasets	Model	F1-macro		
		Baseline	Result	SoA
MEX-A3T	Bigrams, *k* = 500	0.77	0.80	**0.85** [3]
RAW-Covid	Unigrams, *k* = 900	0.57	**0.89**	0.74 [4]
LIAR	Trigrams, *k* = 900	0.54	0.56	**0.62** [10]
CoAID	Unigrams, *k* = 900	0.64	**0.67**	0.58 [6]

Table depicting the F1 scores of the best performing models for each dataset. English datasets seem to have less accurate results than Spanish ones. Baseline is result from doing that task without masking

VOCAB: (w/definition)	**Masking (linguistics/NLP):** the action of hiding words in a text. Many definitions seem to see this in the context of predicting the masked word, but that clearly isn't the definition used here. **Smart quotes:** quotation marks that adjust based on if they start or end a set of quotation marks. Aka, the bane of my existence when I am copying code from a word document to an IDE. **Validation (machine learning):** "The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters." Seems to be a type of testing dataset using data from the training set? **BoW representation/model:** counting how many times a word (unigram) or set of words (bigram, trigram, any n-gram) occurs in a text
Cited references to follow up on	n/a
Follow up Questions	Does masking words provide an increase in accuracy compared to not doing so? Why would the number of characters in a masked work be helpful for evaluating if the article/text is true or not? What are sufficient datasets in English that are purely full articles, not claims? Would this method work for a more character-based language like Chinese?

Article #18 Notes: A Topic-Agnostic Approach for Identifying Fake News Pages

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	A Topic-Agnostic Approach for Identifying Fake News Pages
Source citation (APA Format)	Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., & Freire, J. (2019). A Topic-Agnostic Approach for Identifying Fake News Pages. *Companion Proceedings of The 2019 World Wide Web Conference*, 975–980. https://doi.org/10.1145/3308560.3316739
Original URL	https://dl.acm.org/doi/pdf/10.1145/3308560.3316739
Source type	Conference paper
Keywords	Misinformation; Fake News Detection; Classification; Online News

Summary of key points + notes (include methodology)	- Most approaches to fake news detection: using content of news
	- However, news topics and discourse are constantly changing so this approach cannot be used for a long term solution
	- Some studies also show that page content alone is not enough to classify the truthfulness of news
	- This paper's approach uses classification strategy that is topic-agnostic
	    - Alternative strategy to using bag of words
	    - Requires the use of a diverse dataset
	Related Work
	
	TOPIC-AGNOSTIC CLASSIFICATION
	Topic-Agnostic Features
	- Fake news pages have a lot of ads
	- Recent work proposes that fake news articles are designed to insight inflammatory emotions in readers
	- Fake news contains text patterns related to understandability that differs from that of real news
	- Fake news websites tend to have

	- Lots of ads - Polluted layouts - Sensationalist headlines designed to catch reader's attention ("Just in", "read this". "Breaking news") - Two broad class of features analyzed - Web-markup - Frequency of ads - Presence of an author name - Frequency of various tag groups - Linguistic based - Morphological features - Obtained through part-of-speech tagging - Each word assigned to category based on definition and context - Psychological Features - Percentage of total semantic words in text - Obtained by using dictionary with words that express physiological processes - Readability features - Show ease or difficulty of comprehending a text - Previous work has found that fake news often differs between its headline and body text - Body text of fake news tends to be less informative because main idea is already in title - Analysis of linguistic features can then be split into 3 categories - Only headline - Only content - Headline and content Feature Selection - Combination of 4 different methods - Shannon Entropy (SE) - Tree Based rule (TB) - L1 Regularization (L1) - Mutual Information (AI) - Outputs of these methods are combined and normalized and then applying the geometric mean $r(f_i) = \sqrt[4]{SE(f_i)^{-1} \times TB(f_i) \times L1(f_i) \times MI(f_i)}$ - - Features with a score of 0 are removed - "The sizes of the sets of topic-agnostic features are: (1) for headlines, 137 features; (2) for content, 148; and (3) headline+content, 145." 3.3 Classification - Two categories used: fake news and real news - 3 different learning methods used: Support Vector Machine (SVM),

	K-Nearest Neighbours (KNN), and Random Forest (RF)
	EXPERIMENTAL EVALUATION
	- Created a new dataset called PoliticalNews that contains a variety of new sources from 2013 to 2018
	- NLTK library used for Morphological Features
	- LIWC used for Psychological Features
	- Readability Features uses Textstat library
	- Web-markup features uses BeautifulSoup and Newspaper
	- This classifier compared with Fake News Detector (FNDetector)
	Effectiveness of Different Features
	- Combination of features obtained highest accuracies
	- Combining other features with LIWC obtained higher accuracies
	- Results for US-Election2016 and PoliticalNews are similar
	- For celebrity dataset better results obtained from content of articles
	    - Possibility because celebrity news all have similarity styled headlines by body text between real and fake has notable differences
	- Using web-markup data is effective
	Effectiveness over Time
	- Used news from one timeframe for testing and tested using news from different timeframe
	- TAG model always performed better than FNDetector
	- FNDetector dependent on textual content while TAG is not
	- Content based error detection have to be constantly retrained which is costly and prone to error
	Effectiveness for Different Domains
	- Two experiments
	    - Celebrity as training dataset and US-Election2016 as testing dataset
	    - The other way around
	    - TAG approach turns out to be more effective than FNDetector for both approaches
	CONCLUSIONS & FUTURE WORK
	- Approach is able to account for political news but also news for differing domains
	- Uses significantly fewer features and doesn’t need frequent retraining
	- “topic-agnostic features are effective for distinguishing between fake and real news”
	- “New corpus of over 14,000 political news pages drawn from 137 sites and spanning 6 years”
	- Future work
	    - Account for additional features like user engagement and network structure
	    - Find different strategies to expand fake news corpus like use of social media or a focused crawler

Research	How can we accurately detect fake news when the common news topics are

Question/Problem/ Need	changing constantly?

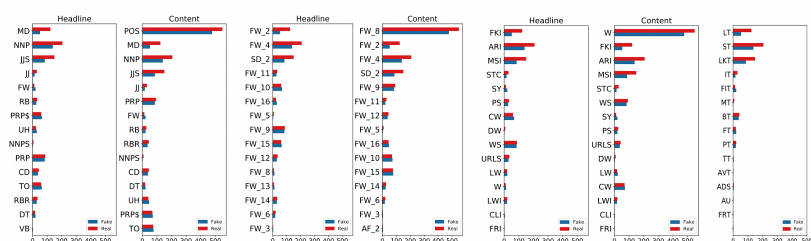
Important Figures

(a) Fake News: Clash Daily (b) Fake News: CDP (c) Real News: Reuters

Figure 1: Web pages from unreliable and reliable new sites.

Table 1: Linguistic and web-markup features used to represent news articles.

	Abbr.	Description	Abbr.	Description	Abbr.	Description	Abbr.	Description	Abbr.	Description
Morphological Features	WDT	WH-determiner	PDT	Pre-determiner	JJ	Adjective or numeral, ordinal	VB	Verb, base form	MD	Modal auxiliary
	CD	Numeral, cardinal	VBD	Verb, past tense	VBG	Verb, present participle or gerund	VBN	Verb, past participle	RP	Particle
	DT	Determiner	NNPS	Noun, proper, plural	NN	Noun, common, singular or mass	CC	Conjunction, coordinating	WRB	Wh-adverb
	FOW	Foreign word	NNS	Noun, common, plural	TO	"To" as preposition/infinitive, superlative	WP$	WH-pronoun, possessive	JJS	Adjective, superlative
	WP	WH-pronoun	POS	Genitive marker	VBP	Verb, present tense, not 3rd singular	RBR	Adverb, comparative	NNP	Noun, proper, singular
	UH	Interjection	PRP	Pronoun, personal	VBZ	Verb, present tense, 3rd singular	RBS	Adverb superlative	PRP$	Pronoun, possessive
	RB	Adverb	JJR	Adjective, comparative	IN	Preposition or conjunction, subordinating				
Psychological Features	SO	Social (family, friend)	SD	Summary Dimensions	BP	Biological Processes (ingest, health, sexual)	AF	Affect (anger, sad, anxiety)	PC	Personal Concerns
	RL	Relativity (space, time)	FW	Function Words	TR	Time Orientation (focuspast, focuspresent)	PP	Perceptual Process	IL	Informal Language
	DR	Drives (power, risk)	PM	Punctuation Marks	OG	Other Grammar (quantifiers, interrogatives)	CP	Cognitive Processes		
Readability Features	FRI	Flesch Reading Ease	WS	Words per sentence	LW	Long words	LWI	Linsear Write	CLI	Coleman-Liau
	FKI	Flesch Kincaid Grade	CW	Capitalized words	SY	Syllables	CW	Complex words	DW	Difficult words
	MSI	McLaughlinâĂŹs SMOG	LX	Lexicon	PS	Percentage of stop words	ARI	Automated Readability	W	Words
	GFI	Gunning Fog	URL	URLs	STC	Sentences	CH	Characters		
Web-markup features	AU	Author	IT	Images (e.g., img, canvas)	ST	Semantics (e.g., article, section)	FRT	Frames (e.g., frame, frameset)	LT	Lists (e.g., ul, ol, li)
	BT	Basic (e.g., title,h1, p)	FT	Formatting (e.g., acronym)	FIT	Forms and inputs (e.g., textarea, button)	MT	Metainfo (e.g., head, meta)	TT	Tables (e.g., tbody, tfoot)
	AVT	Audio and video	LKT	Links (e.g., a, nav, link)	PT	Programming (e.g., script, object)	ADS	Advertisements		



(a) Morphological (b) Psychological (c) Readability (d) Web-markup

Figure 2: Mean frequency distribution of features per class in the PoliticalNews dataset.

Table 2: Accuracy results for models that use different set of topic-agnostic features – where L is LIWC, N is NLTK, R is readability, and W is webmarkup features – over three different datasets: Celebrity, US-Election2016, and PoliticalNews. The best accuracies for each feature set are bold; the best accuracies for each news article's representations (H, C and HC) are underlined.

Dataset	**Celebrity**			**US-Election2016**			**PoliticalNews**		
Features	**H**	**C**	**HC**	**H**	**C**	**HC**	**H**	**C**	**HC**
W	**0.68**	**0.68**	**0.68**	**0.65**	**0.65**	**0.65**	**0.71**	**0.71**	**0.71**
L	0.69	**0.73**	**0.73**	0.77	0.81	**0.83**	0.71	0.75	**0.76**
N	0.58	**0.68**	0.66	**0.81**	0.75	0.76	**0.77**	0.66	0.67
R	0.57	**0.62**	0.57	**0.75**	0.73	0.73	**0.69**	0.62	0.64
L-R	0.65	**0.76**	0.74	0.79	0.82	**0.83**	0.74	0.75	**0.76**
N-R	0.65	**0.68**	0.67	**0.83**	0.78	0.78	**0.78**	0.71	0.72
N-W	0.68	**0.72**	**0.72**	0.79	0.79	**0.80**	**0.81**	0.79	0.79
L-W	0.70	**0.77**	0.75	0.79	0.82	**0.83**	0.78	**0.81**	**0.81**
R-W	0.67	**0.72**	0.67	**0.80**	0.76	0.76	**0.77**	0.76	0.76
N-R-W	0.67	**0.72**	0.71	**0.83**	0.79	0.80	**0.81**	0.78	0.79
L-R-W	0.71	**0.78**	0.71	0.79	0.83	**0.85**	0.80	0.80	**0.81**
L-N-R-W	**0.73**	**0.73**	0.71	0.83	0.82	**0.86**	**0.83**	0.81	0.82

Table 3: Classification results (accuracies) for three datasets.

Dataset	**Celebrity**	**US-Election2016**	**PoliticalNews**
FNDetector	0.73	0.81	0.76
TAG Model	**0.78**	**0.86**	**0.83**

Table 4: Cross-domain results (accuracies) between models.

Training	Test	Classifier	Accuracy FNDetector	Accuracy TAG Model
Celebrity	US-Election 2016	SVM	0.59	**0.70**
		KNN	0.59	**0.64**
		RF	0.56	**0.64**
US-Election 2016	Celebrity	SVM	0.59	**0.63**
		KNN	0.56	**0.60**
		RF	0.51	**0.60**

VOCAB: (w/definition)	Sensationalist: presenting stories in a way that is intended to provoke public interest or excitement, at the expense of accuracy. Morphological: relating to the form of words, in particular inflected forms Semantic: relating the meaning in language or logic Granularities: the level of detail in a data structure five-fold cross-validation: resampling procedure used to evaluate machine learning models on a limited data sample. Five-fold means the data is split into 5 parts Agnostic: Not dependent on
Cited references to follow up on	https://osf.io/3agmb (database) Nltk.org
Follow up Questions	In what cases is the use of vocabulary topic-agonistic? How were the factors related to web layout measured? How do you automate the analysis of the html and css elements of a webpage? Is there a way to predict the main news topics that will be prominent in the future?

Article #19 Notes: Python Machine Learning for Beginners

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	Python Machine Learning for Beginners: Learning from scratch NumPy, Pandas, Matplotlib, Seaborn, Scikitlearn, and TensorFlow for Machine Learning and Data Science
Source citation (APA Format)	Malik, U. (2020). *Python Machine Learning for Beginners: Learning from scratch NumPy, Pandas, Matplotlib, Seaborn, Scikitlearn, and TensorFlow for Machine Learning and Data Science*. AI Publishing LLC.
Original URL	n/a
Source type	Book (chapters 3 and 4)
Keywords	Misinformation; Fake News Detection; Classification; Online News
Summary of key points + notes (include methodology)	Chapter 3

```
[3]: #creating a numpy array
     import numpy as np
     nums_list = [10, 12, 14, 16, 18, 20]
     nums_array = np.array(nums_list)
     type(nums_array)

[3]: numpy.ndarray
```

```
[4]: #creating a 2D numpy array
     row1 = [1, 2, 3]
     row2 = [1, 2, 3]
     row3 = [1, 2, 3]

     nums_2d = np.array([row1, row2, row3]) #this is just creating a list of lists and then converting it into a numpy array lul
     nums_2d.shape

[4]: (3, 3)
```

```
[6]: #using arrange method
     nums_arr = np.arange(5, 11)
     print(nums_arr)

     [ 5  6  7  8  9 10]
```

```
[8]: #using arrange method with steps
     nums_arr = np.arange(5, 11, 2)
     print(nums_arr)

     [5 7 9]
```

[12]:
```
#creating an array of ones (why are there periods in the output?) oh the periods are decimal points
ones_array = np.ones(6)
print(ones_array)
ones_array = np.ones((6,4))
print(ones_array)
```

```
[1. 1. 1. 1. 1. 1.]
[[1. 1. 1. 1.]
 [1. 1. 1. 1.]
 [1. 1. 1. 1.]
 [1. 1. 1. 1.]
 [1. 1. 1. 1.]
 [1. 1. 1. 1.]]
```

[13]:
```
#creating an array of zeros (same as previous block but woah there's zeros)
ones_array = np.zeros(6)
print(ones_array)
ones_array = np.zeros((6,4))
print(ones_array)
```

```
[0. 0. 0. 0. 0. 0.]
[[0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]
 [0. 0. 0. 0.]]
```

[15]:
```
# eyes method: diagonal of ones, rest are zeros
eyes_array = np.eye(5)
print(eyes_array)
```

```
[[1. 0. 0. 0. 0.]
 [0. 1. 0. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 0. 0. 1. 0.]
 [0. 0. 0. 0. 1.]]
```

[16]:
```
#random values
uniform_random = np.random.rand(6,4)
print(uniform_random)
```

```
[[0.4051233  0.61784359 0.24114935 0.08200774]
 [0.78565382 0.30101406 0.4018976  0.50339418]
 [0.53837428 0.04589819 0.35204789 0.51099489]
 [0.24866094 0.27277775 0.2075006  0.97245423]
 [0.66833221 0.90935247 0.51064579 0.34755727]
 [0.18457834 0.17378026 0.19042301 0.46152971]]
```

[17]:
```
#index array and slicing: it's the same as normal lists in python
s = np.arange(1, 11)
print(s[1])
print(s[1:9])
print(s[5:])
print(s[:5])
```

```
2
[2 3 4 5 6 7 8 9]
[ 6  7  8  9 10]
[1 2 3 4 5]
```

[20]:
```
#2D arrays are the same for slicing, just put a comma between each corridinate
row1 = [1, 2, 3]
row2 = [1, 2, 3]
row3 = [1, 2, 3]

nums_2d = np.array([row1, row2, row3])
print(nums_2d[:,2:])
print(nums_2d[:,:2])
print(nums_2d[1:,1:])
```

```
[[3]
 [3]
 [3]]
[[1 2]
 [1 2]
 [1 2]]
[[2 3]
 [2 3]]
```

[23]:
```
#arithmetic operations
nums = [1, 4, 9, 16, 25]

np_sqrt = np.sqrt(nums)
print(np_sqrt)

np_log = np.log(nums)
print(np_log)

np_exp = np.exp(nums)
print(np_exp)

np_sin = np.sin(nums)
print(np_sin)

np_cos = np.cos(nums)
print(np_cos)
```

```
[1. 2. 3. 4. 5.]
[0.         1.38629436 2.19722458 2.77258872 3.21887582]
[2.71828183e+00 5.45981500e+01 8.10308393e+03 8.88611052e+06
 7.20048993e+10]
[ 0.84147098 -0.7568025   0.41211849 -0.28790332 -0.13235175]
[ 0.54030231 -0.65364362 -0.91113026 -0.95765948  0.99120281]
```

[25]:
```
#linear algebra operations
#dot product (matrix multiplication)
A = np.random.rand(5, 4)
B = np.random.rand(4, 5)
Z = np.dot(A, B)
print(Z)

# does not work if columns in 1st matrix and rows in 2nd matrix do not match
A = np.random.rand(5, 3)
B = np.random.rand(4, 5)
Z = np.dot(A, B)
```

```
[[0.74309163 1.33172954 0.82406029 1.4061441  0.68389378]
 [0.67143721 1.28918781 0.9808969  1.31312781 0.43265249]
 [0.9048389  1.61814579 0.99071828 1.65705632 1.00169375]
 [0.79498287 1.83939224 1.15715469 1.9443499  0.71848003]
 [0.35375623 1.13155944 0.85957336 1.12148459 0.28394801]]
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Input In [25], in <cell line: 11>()
      9 A = np.random.rand(5, 3)
     10 B = np.random.rand(4, 5)
---> 11 Z = np.dot(A, B)

File <__array_function__ internals>:5, in dot(*args, **kwargs)

ValueError: shapes (5,3) and (4,5) not aligned: 3 (dim 1) != 4 (dim 0)
```

[30]:
```
#element-wise matrix multiplication
#array1[1][1] * array2[1][1] = product[1][1]
#array1[2][1] * array2[2][1] = product[2][1]
#etc...
row1 = [1, 2, 3]
row2 = [4, 5, 6]
row3 = [7, 8, 9]

num_2d = np.array([row1, row2, row3])
multiply = np.multiply(num_2d, num_2d)
print(multiply)
```

```
[[ 1  4  9]
 [16 25 36]
 [49 64 81]]
```

[32]:
```
#Matrix inverse
#since you cannot divide matrixes, the alternative is to multiply a matrix by an inverse
row1 = [1,2,3]
row2 = [4,5,6]
row3 = [7,8,9]

nums_2d = np.array([row1, row2, row3])
inverse = np.linalg.inv(num_2d)
print(inverse) #why is the output different from the one in the book?
```

```
[[-4.50359963e+15  9.00719925e+15 -4.50359963e+15]
 [ 9.00719925e+15 -1.80143985e+16  9.00719925e+15]
 [-4.50359963e+15  9.00719925e+15 -4.50359963e+15]]
```

[34]:
```
#Exercise 3.2

rand = np.random.rand(5, 4)
print(rand[2:,1:])
```

```
[[0.99844116 0.1949089  0.489411  ]
 [0.38322582 0.99200532 0.02540675]
 [0.13210004 0.35842705 0.49025144]]
```

Chapter 4

[1]:
```
import pandas as pd
news_data = pd.read_csv("/Users/annewu/Desktop/STEM1 Project/Horne2017_FakeNewsData/Public Data/Buzzfeed Political News Datase
news_data.head()
```

[1]:

	Filename	Segment	WC	Analytic	Clout	Authentic	Tone	WPS	BigWords	Dic	...	assent	nonflu	filler	AllPunc	Period	Comma	QMark	Exc
0	44_Fake.txt	1	429	65.82	41.45	32.50	10.88	15.89	19.58	78.55	...	0.00	0.0	0.0	13.75	4.66	6.06	1.86	
1	45_Fake.txt	1	445	78.69	66.48	21.15	11.15	15.89	18.20	81.35	...	0.00	0.0	0.0	9.44	6.07	2.70	0.22	
2	33_Fake.txt	1	547	82.20	71.72	10.32	24.92	23.78	30.53	76.42	...	0.00	0.0	0.0	11.70	4.20	5.67	0.00	
3	32_Fake.txt	1	358	74.63	58.36	24.09	20.23	18.84	19.27	84.08	...	0.28	0.0	0.0	10.61	5.03	4.19	0.28	
4	39_Fake.txt	1	295	36.77	38.09	24.04	51.56	17.35	13.90	85.42	...	0.00	0.0	0.0	12.20	5.08	6.44	0.68	

5 rows × 119 columns

[2]:
```
titanic_data = pd.read_csv("/Users/annewu/Desktop/STEM1 Project/machine_learning_beginner/Data/titanic_data.csv")
titanic_data.head()
```

[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

[3]:
```
titanic_pclass1 = (titanic_data.Pclass == 1)
titanic_pclass1
```

[3]:
```
0      False
1       True
2      False
3       True
4      False
       ...
886    False
887     True
888    False
889     True
890    False
Name: Pclass, Length: 891, dtype: bool
```

[4]:
```
titanic_pclass1_data = titanic_data[titanic_pclass1]
titanic_pclass1_data.head()
```

[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
23	24	1	1	Sloper, Mr. William Thompson	male	28.0	0	0	113788	35.5000	A6	S

[6]:
```
#isin operator: takes list of values and returns rows that contain the list of values in a chosen column
ages = [20, 21, 22]
age_dataset= titanic_data[titanic_data["Age"].isin(ages)]
age_dataset.head()
```

[6]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	NaN	S
12	13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.05	NaN	S
37	38	0	3	Cann, Mr. Ernest Charles	male	21.0	0	0	A./5. 2152	8.05	NaN	S
51	52	0	3	Nosworthy, Mr. Richard Cater	male	21.0	0	0	A/4. 39886	7.80	NaN	S
56	57	1	2	Rugg, Miss. Emily	female	21.0	0	0	C.A. 31026	10.50	NaN	S

[8]:
```
# can use "and" and "or" operators to filter rows
# code shows condition where passenger is 20, 21, or 22 years old and is part of the 1st class
ages = [20, 21, 22]
ageclass_dataset = titanic_data[titanic_data["Age"].isin(ages) & (titanic_data["Pclass"] == 1)]
ageclass_dataset.head()
```

[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
102	103	0	1	White, Mr. Richard Frasar	male	21.0	0	1	35281	77.2875	D26	S
151	152	1	1	Pears, Mrs. Thomas (Edith Wearne)	female	22.0	1	0	113776	66.6000	C2	S
356	357	1	1	Bowerman, Miss. Elsie Edith	female	22.0	0	1	113505	55.0000	E33	S
373	374	0	1	Ringhini, Mr. Sante	male	22.0	0	0	PC 17760	135.6333	NaN	C
539	540	1	1	Frolicher, Miss. Hedwig Margaritha	female	22.0	0	2	13568	49.5000	B39	C

[9]:
```
#Filter columns
titanic_data_filter = titanic_data.filter(["Name","Sex","Age"]) # only display the Name, Sex and Age columns
titanic_data_filter.head()
```

[9]:

	Name	Sex	Age
0	Braund, Mr. Owen Harris	male	22.0
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0
2	Heikkinen, Miss. Laina	female	26.0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
4	Allen, Mr. William Henry	male	35.0

[10]:
```
#can also filter out columns using drop() methods
titanic_data_filter = titanic_data.drop(["Name", "Sex", "Age"], axis = 1) # display columns that are not Name, Sex, or Age
# what does "axis = 1" do? Ans: it is the axis that it wants to filter on. In this case, 1 is filter the columns
titanic_data_filter.head()
```

[10]:

	PassengerId	Survived	Pclass	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	1	0	PC 17599	71.2833	C85	C
2	3	1	3	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	1	0	113803	53.1000	C123	S
4	5	0	3	0	0	373450	8.0500	NaN	S

```
[13]: #Concatenation: joining multiple panda dataframes
#concatenate vertically
#concatenated datasets must have an equal number of columns
titanic_pclass1_data = titanic_data[titanic_data.Pclass == 1]
print(titanic_pclass1_data.shape)

titanic_pclass2_data = titanic_data[titanic_data.Pclass == 2]
print(titanic_pclass2_data.shape)

#first way: append pclass2 to pclass1 (or vice versa)
#apparently this is deprecated and will be removed later so just use concat
final_dataA = titanic_pclass1_data.append(titanic_pclass2_data, ignore_index = True)
#ignore_index = True resets data indexes
print(final_dataA.shape)

#second way: put both pclass1 and pclass2 as parameters in concat() method
final_dataB = pd.concat([titanic_pclass1_data, titanic_pclass2_data])
#print([titanic_pclass1_data, titanic_pclass2_data].shape)
print(final_dataB.shape)
```

```
(216, 12)
(184, 12)
(400, 12)
(400, 12)
/var/folders/bd/6xhn36wj7pz38wyn20d25vh00000gp/T/ipykernel_16347/2566370345.py:10: FutureWarning: The frame.append method is d
eprecated and will be removed from pandas in a future version. Use pandas.concat instead.
  final_dataA = titanic_pclass1_data.append(titanic_pclass2_data, ignore_index = True)
```

```
#concatenate horizontally (I don't exactly understand the difference)
#concatenated datasets must have an equal number of rows
#will need to change axis attribute to 1
#...what do these extra rows do

df1 = final_dataA[:200]
df2 = final_dataA[200:]

final_data2 = pd.concat([df1, df2], axis = 1, ignore_index = True)
print(final_data2.shape)
final_data2.head()
```

(400, 24)

	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21	22	23
0	2.0	1.0	1.0	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1.0	0.0	PC 17599	71.2833	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	4.0	1.0	1.0	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1.0	0.0	113803	53.1000	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	7.0	0.0	1.0	McCarthy, Mr. Timothy J	male	54.0	0.0	0.0	17463	51.8625	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	12.0	1.0	1.0	Bonnell, Miss. Elizabeth	female	58.0	0.0	0.0	113783	26.5500	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	24.0	1.0	1.0	Sloper, Mr. William Thompson	male	28.0	0.0	0.0	113788	35.5000	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 24 columns

```
#sorting dataframes

#sort passangers by age ascending
age_sorted_data = titanic_data.sort_values(by=["Age"])
age_sorted_data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
803	804	1	3	Thomas, Master. Assad Alexander	male	0.42	0	1	2625	8.5167	NaN	C
755	756	1	2	Hamalainen, Master. Viljo	male	0.67	1	1	250649	14.5000	NaN	S
644	645	1	3	Baclini, Miss. Eugenie	female	0.75	2	1	2666	19.2583	NaN	C
469	470	1	3	Baclini, Miss. Helene Barbara	female	0.75	2	1	2666	19.2583	NaN	C
78	79	1	2	Caldwell, Master. Alden Gates	male	0.83	0	2	248738	29.0000	NaN	S

```
#if descending, ascending has to be declared False
age_sorted_data = titanic_data.sort_values(by=["Age"], ascending = False)
age_sorted_data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0000	A23	S
851	852	0	3	Svensson, Mr. Johan	male	74.0	0	0	347060	7.7750	NaN	S
493	494	0	1	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	NaN	C
96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	A5	C
116	117	0	3	Connors, Mr. Patrick	male	70.5	0	0	370369	7.7500	NaN	Q

```
#if multiple columns to be sorted, data will be sorted by 1st column and then sorted by other columns
#if there are tied values for first column
age_sorted_data = titanic_data.sort_values(by=["Age"], ascending = False)
```

Research Question/Problem/ Need	How to use the NumPy and Pandas python libraries for data analysis
Important Figures	n/a
VOCAB: (w/definition)	Dot product: the sum of the product of corresponding entries for two

	sequences of numbers Concatenate: link things together in a chain of series
Cited references to follow up on	n/a
Follow up Questions	How can effective graphs be made using python? How do you make side by side box and whisker plots on python?