Ai for Detecting and Addressing Misinformation

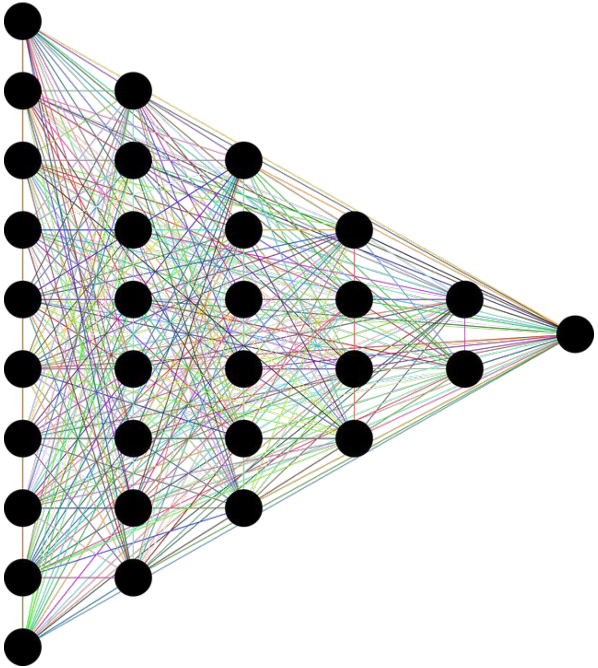
Presented By: Gabriel Stewart, Harsh Gupta, Natan Berehe, Tejas Desale

What is Misinformation?

- false or inaccurate information, especially that which is deliberately intended to deceive.
- Affecting major areas of our society:
    - Social Media
    - News
    - Various
- Mostly done using AI, however, there exists multiple Ai tools that can help as well

AI Text Generation

- AI text generation has become smarter and harder to detect
- This can be used to make fake links that can mirage as real media



Source: Google Images Fair Use

Social Media

- 2016 election had many posts that were later flagged as misinformation
- Bots can spread misinformation fast enough for the algorithm to pick it up

Source: Google Images Fair Use

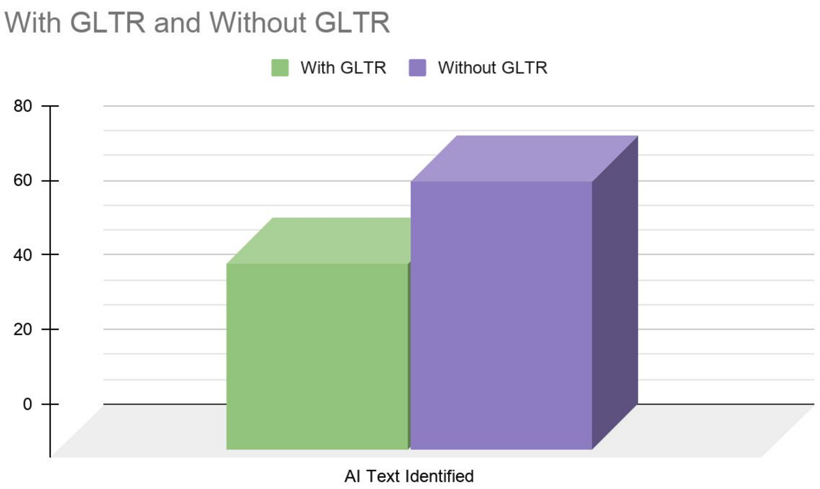
Deep Fakes

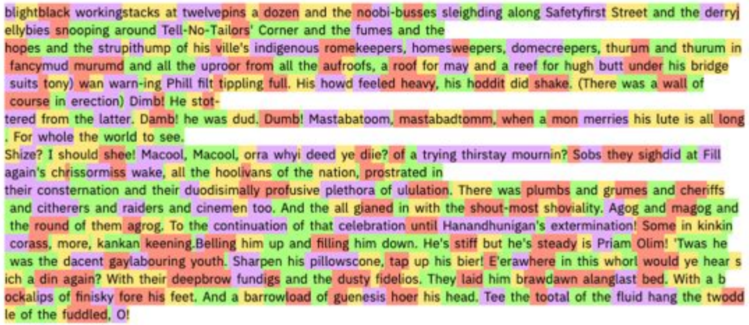
- DeepFakes has been used to alter images and videos
- DeepFakes are getting harder and harder to differentiate
- DeepFakes are getting easier to make

Source

Giant Language model Test Room

- Talk to Transformer – InferKit
- Giant Language model Test Room or “glitter”
- GLTR (glitter) v0.5
- GLTR is a tool not an autonomous solution





blightblack workingstacks at twelvepins a dozen and the noobi-busses sleighding along Safetyfirst Street and the derryj
ellybies snooping around Tell-No-Tailors' Corner and the fumes and the
hopes and the strupithump of his ville's indigenous romekeepers, homesweepers, domecreepers, thurum and thurum in
fancymud murumd and all the uproor from all the aufroofs, a roof for may and a reef for hugh butt under his bridge
suits tony) wan warn-ing Phill filt tippling full. His howd feeled heavy, his hoddit did shake. (There was a wall of
course in erection) Dimb! He stot-
tered from the latter. Damb! he was dud. Dumb! Mastabatoom, mastabadtomm, when a mon merries his lute is all long
. For whole the world to see.
Shize? I should shee! Macool, Macool, orra whyi deed ye diie? of a trying thirstay mournin? Sobs they sighdid at Fill
again's chrissormiss wake, all the hoolivans of the nation, prostrated in
their consternation and their duodisimally profusive plethora of ululation. There was plumbs and grumes and cheriffs
and citherers and raiders and cinemen too. And the all gianed in with the shout-most shoviality. Agog and magog and
the round of them agrog. To the continuation of that celebration until Hanandhunigan's extermination! Some in kinkin
corass, more, kankan keening.Belling him up and filling him down. He's stiff but he's steady is Priam Olim! 'Twas he
was the dacent gaylabouring youth. Sharpen his pillowscone, tap up his bier! E'erawhere in this whorl would ye hear s
ich a din again? With their deepbrow fundigs and the dusty fidelios. They laid him brawdawn alanglast bed. With a b
ockalips of finisky fore his feet. And a barrowload of guenesis hoer his head. Tee the tootal of the fluid hang the twodd
le of the fuddled, O!

Source

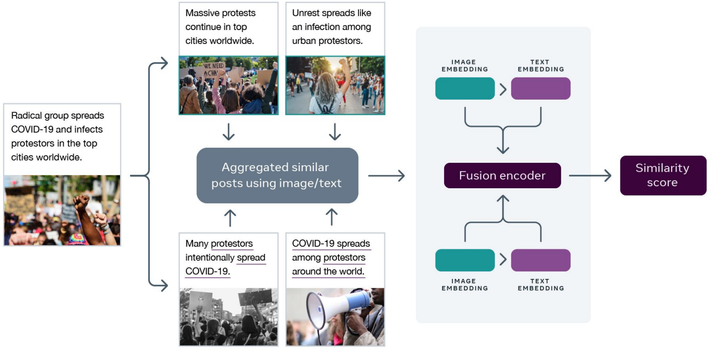
Detection of Fake News

- We believe the information sent from the people we trust.
  - Form biases and not verify it
  - Information Overload on Social Media, preventing verification
- Finite attention spans of social media
  - No discrimination on the basis of quality
  - Misinformation - Change Opinion & Disinformation - Put the wrong information
- Quick Spread
  - Fake news spread 70% faster than truth
  - Digital media is much quicker than print media
- Opinion based information instead of fact based
  - Demonstrate personal views
  - Multiple Variations

Source: Google Images Fair Use

Impact of AI on Social Media

- AI is challenged from emergence of IoT, Robotics, Augmented & Virtual Reality, etc.
- Dr. Huan Liu from ASU, has been working on creating a dataset “FakeNewsNet” storing fake news.
- Medical :- End to End attention neural network that detects fake information on Twitter posts.
- Several systems rely on LASER and ObjectDNA, based on Facebook AI research.



Radical group spreads COVID-19 and infects protestors in the top cities worldwide.

Massive protests continue in top cities worldwide.

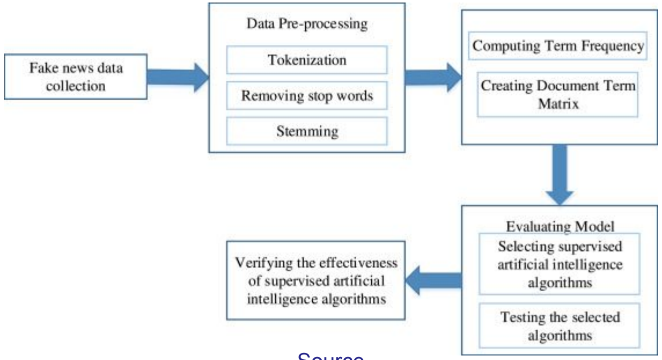
Unrest spreads like an infection among urban protestors.

Many protestors intentionally spread COVID-19.

COVID-19 spreads among protestors around the world.

Aggregated similar posts using image/text

Source



Fake news data collection

Data Pre-processing

Tokenization

Removing stop words

Stemming

Computing Term Frequency

Creating Document Term Matrix

Evaluating Model

Selecting supervised artificial intelligence algorithms

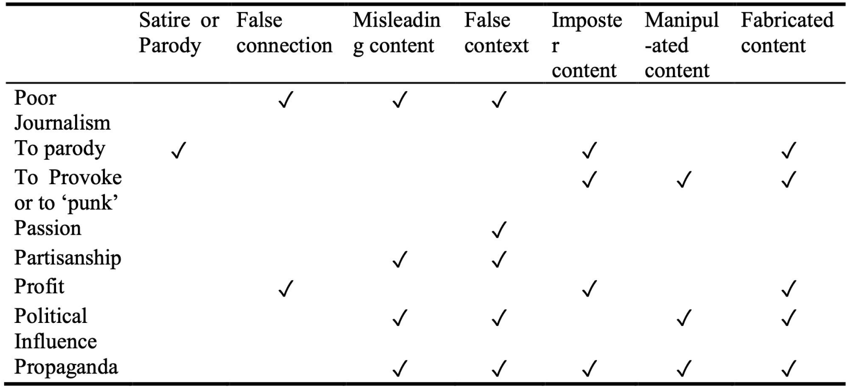
Testing the selected algorithms

Verifying the effectiveness of supervised artificial intelligence algorithms

Source

Results achieved

- The medical field is able to hunt down 95% of the fake information with regard to COVID-19. It is able to explain the misinformation as well.
- Several Machine Learning Libraries like Textbox to help run sentiment analysis to detect fake news with about 90-95% accuracy.
- Development of several AI powered analytics that include stance to correlate the heading with the content in the article
- Text processing to analyze the writing style and reliability.
- Development of matrices to build new NLP models



	Satire or Parody	False connection	Misleading content	False context	Imposter content	Manipul-ated content	Fabricated content
Poor Journalism		✓	✓	✓			
To parody	✓				✓		✓
To Provoke or to ‘punk’					✓	✓	✓
Passion				✓			
Partisanship			✓	✓			
Profit		✓			✓		✓
Political Influence			✓	✓		✓	✓
Propaganda			✓	✓	✓	✓	✓

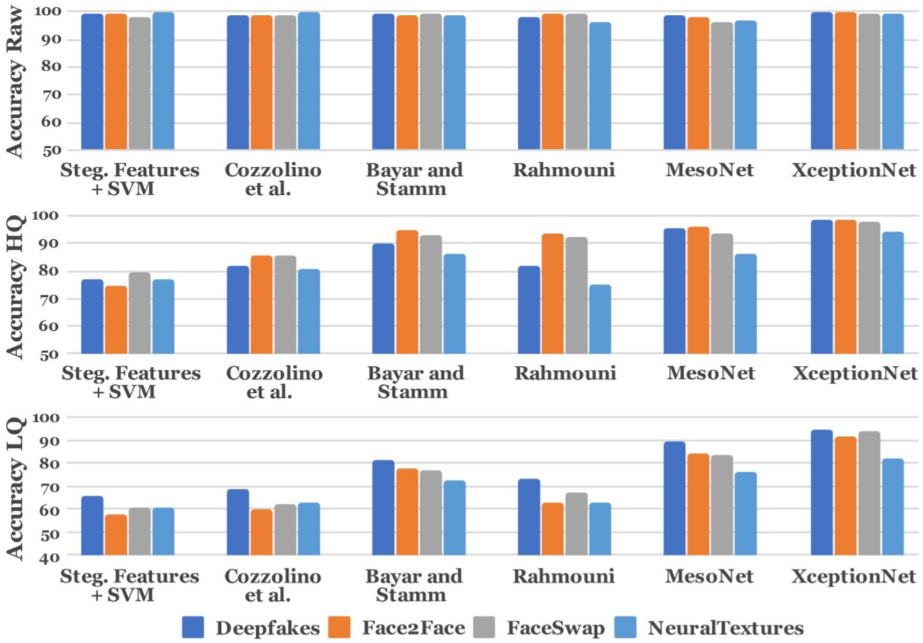
Source

Detection of Deepfakes

- Create forensic tools to detect deepfakes
  - Increasingly more difficult as they improve
  - Inconsities, blurring, blinking, and more
- Search for artifacts of other sources
  - Use method similar to creation
  - Algorithms designed to trick these detectors
- Non-AI approaches of combating deepfakes
  - Digital watermarks
  - New legislation
- Research and funding coming from government and private industry

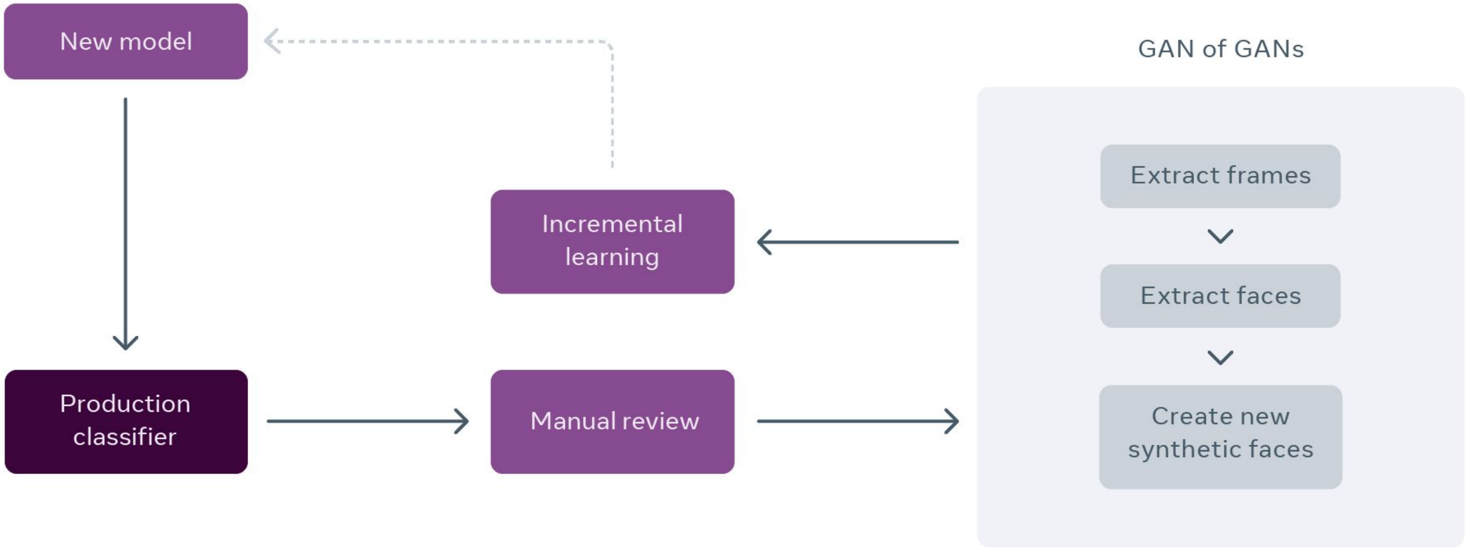
Source: Google Images Fair Use

Model Accuracy - Features and CNN



Source

Multiple Generative Adversarial Network



Source

The Deepfake Detection Challenge - Facebook AI Red Team

- Hosted by Facebook, Microsoft, Amazon, and the Partnership on AI
- Challenge open to public
- Best against public dataset, 82.56% success
- Best against black box dataset, 65.18% success

Source

THANKS!
ANY QUESTION?

References

Arnold, M. (2020, July 13). *An algorithm to detect fake news: A Q&A with Huan Liu and Kai Shu | Knowledge Enterprise*. Knowledge Enterprise. https://research.asu.edu/algorithm-detect-fake-news-qa-huan-liu-and-kai-shu

Bovet, A. (2019, January 2). *Influence of fake news in Twitter during the 2016 US presidential election.* Nature Communications. https://www.nature.com/articles/s41467-018-07761-2

Brown, S. (2020, October 5). *MIT Sloan research about social media, misinformation, and elections*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/mit-sloan-research-about-social-media-misinformation-and-elections

Burrows, L. (n.d.). *Visual forensics to detect fake text*. Harvard John A. Paulson School of Engineering and Applied Sciences. https://www.seas.harvard.edu/news/2019/07/visual-forensics-detect-fake-tex

CBS News. (2021, March 15). *Cheerleader's mom accused of making "deepfake" videos of daughter's rivals*. https://www.cbsnews.com/news/raffaela-spone-cheerleader-mom-deepfakes/

Ciobanu, M. (2018, April 14). *The challenges and opportunities of using artificial intelligence to tackle misinformation*. Media News. https://www.journalism.co.uk/news/the-challenges-and-opportunities-of-using-artificial-intelligence-to-tackle-misinformation/s2/a720411/

Cohen, I. (2020, October 8). *Can AI Analytics Stop Fake News?* Anodot. https://www.anodot.com/blog/ai-analytics-stops-fake-news/

Dodgson, N. (2018, February 21). *Face-swap on steroids: How 'deepfake' videos are messing with reality*. The Spinoff. https://thespinoff.co.nz/science/22-02-2018/face-swap-on-steroids-how-deepfake-videos-are-messing-with-reality/

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). *The deepfake detection challenge (dfdc) preview dataset*. arXiv preprint arXiv:1910.08854.

Edell, A. (2018, June 21). *I trained fake news detection AI with >95% accuracy, and almost went crazy*. Medium. https://towardsdatascience.com/i-trained-fake-news-detection-ai-with-95-accuracy-and-almost-went-crazy-d10589aa57c

Engler, A. (2019, November 14). *Fighting deepfakes when detection fails*. Brookings. https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/

Fake News on Social Media. (n.d.). Mailchimp. https://mailchimp.com/resources/fake-news-on-social-media/

Fano, A., & Sengupta, S. (2020, September 18). *Tackling medical misinformation in social media with AI*. Accenture. https://www.accenture.com/us-en/blogs/technology-innovation/sengupta-fano-tackling-medical-misinformation-in-social-media-with-ai

Ferrer, C. C., Dolhansky, B., Pflaum, B., Bitton, J., Pan, J., & Lu, J. (2020, June 12). *Deepfake Detection Challenge Results: An open initiative to advance AI*. Facebook AI. https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation

Here's how we're using AI to help detect misinformation. (2020, November 19). Facebook AI. https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation

Kertysova, K. (2018). *Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered*. Security and Human Rights, 29(1-4), 55-81.

Lim, H. (2020, November 18). *Three Types of Deepfake Detection*. Lionbridge AI. https://lionbridge.ai/articles/three-types-of-deepfake-detection/

Menczer, F. T. H. (2020, December 1). *Information Overload Helps Fake News Spread, and Social Media Knows It*. Scientific American. https://www.scientificamerican.com/article/information-overload-helps-fake-news-spread-and-social-media-knows-it/?error=cookies_not_supported&code=470872a4-7c5b-4038-a856-b161823ddb33

MIT Media Lab. (2020, May 4). *Detect DeepFakes: How to counteract misinformation created by AI*. Medium. https://medium.com/mit-media-lab/detect-deepfakes-how-to-counteract-misinformation-created-by-ai-4ea111251c5f

MIT-IBM Watson AI lab. (n.d.). *GLTR (glitter) v0.5*. Catching Unicorns with GLTR. https://gltr.io/dist/index.html

Ozbay, F. A., & Alata, B. (2020, February 15). *Fake news detection within online social media using supervised artificial intelligence algorithms*. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0378437119317546

Paur, J. (2020, December 19). *Nicolas Cage is Neo From THE MATRIX in This New Deepfake Video*. GeekTyrant. https://geektyrant.com/news/nicolas-cage-is-neo-from-the-matrix-in-this-new-deepfake-video

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceforensicOzbay, F. A., & Alata, B. (2020, February 15). Fake news detection within online social media using supervised artificial intelligence algorithms. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0378437119317546s++: Learning to detect manipulated facial images*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11).

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). *Fake News Detection: A Deep Learning Approach*. SMU Data Science Review. https://scholar.smu.edu/cgi/viewcontent.cgi?article=1036&context=datasciencereview

Woolley, S. (2020, April 2). *We're fighting fake news AI bots by using more AI. That's a mistake. MIT Technology Review*. https://www.technologyreview.com/2020/01/08/130983/were-fighting-fake-news-ai-bots-by-using-more-ai-thats-a-mistake/