ATTACKING NEURAL TEXT DETECTORS

Max Wolff
Viewpoint School
Calabasas, CA 91302, USA
m.wolff20@viewpoint.org

ABSTRACT

Machine learning based language models have recently made significant progress, which introduces a danger to spread misinformation. To combat this potential danger, several methods have been proposed for detecting text written by these language models. This paper presents two classes of black-box attacks on these detectors, one which randomly replaces characters with homoglyphs, and the other a simple scheme to purposefully misspell words. The homoglyph and misspelling attacks decrease a popular neural text detector's recall on neural text from 97.44% to 0.26% and 22.68%, respectively. Results also indicate that the attacks are transferable to other neural text detectors.

1 INTRODUCTION

Contemporary state of the art language models such as GPT-2 (Radford et al., 2019) are rapidly improving, as they are being trained on increasingly large datasets and defined using billions of parameters. Language models are currently able to generate coherent text that humans can identify as machine-written text (neural text) with approximately 54% accuracy. (Gehrmann et al., 2019)– close to random guessing. With this increasing power, language models provide bad actors with the potential to spread misinformation on an unprecedented scale (Solaiman et al., 2019) and undermine clear authorship.

To reduce the spread of misinformation via language models and give readers a better sense of what entity (machine or human) may have actually written a piece of text, multiple neural text detection methods have been proposed. Two automatic neural text detectors are considered in this work, RoBERTa$_{\text{LARGE}}$ (Solaiman et al., 2019; Liu et al., 2019) and GROVER (Zellers et al., 2019), which are 95% and 92% accurate in discriminating neural text from human-written text, respectively. Another tool, GLTR (Gehrmann et al., 2019), is designed to assist humans in detecting neural text, increasing humans' ability to correctly distinguish neural text and human-written text from 54% to 72%. Fundamentally, these detectors are based on the fact that neural text follows predictable patterns based on the neural text's underlying language model generator.

Attacks on machine learning models, called adversarial attacks (Szegedy et al., 2014; Athalye et al., 2018; Sharif et al., 2016; Ebrahimi et al., 2018), have been studied in depth and used to expose both security holes and understand how machine learning models function by purposefully causing machine learning models to make mistakes.

Historically, homoglyph attacks[1] have been used to direct victims to malicious websites by replacing characters in a trusted URL with similar looking ones, called homoglyphs. Part of this work seeks to test whether homoglyph attacks can also be used to create effective black-box adversarial attacks on neural text detectors.

2 THREAT MODEL AND PROPOSED ATTACKS

In this paper, two classes of attacks on neural text detectors are proposed. Both of these attacks attempt to modify neural text in ways that are relatively visually imperceptible to humans, but will cause a neural text detector to misclassify the text as human-written. Specifically, these attacks

[1] https://en.wikipedia.org/wiki/IDN_homograph_attack

change the underlying distribution of neural text so that it diverges from that of the language model which generated it.

The first class of attacks are non human-like attacks, which imperceptibly (according to humans) change neural text in a way that humans normally would not. This class of attack shifts the modified text's distribution away from its original one. In this work, the non-human like attacks are realized by swapping selected characters with Unicode homoglyphs (e.g. changing English "a"s to Cyrillic "a"s throughout a neural text sample). Homoglyphs are chosen because they appear visually similar to their counterparts, but get tokenized differently by neural text detectors.

The second class of attacks are human-like attacks, which imperceptibly (according to humans) change neural text in a way that humans normally would. In this paper, this class of attack is realized by randomly swapping correctly spelled words with common human misspellings throughout a neural next sample–which from here onward is referred to as a "misspelling attack." However, this is not the only way human-like attacks may be implemented. This class of attack may also target word-choice, grammar, or punctuation. Misspelling attacks are simply a proof-of-concept for this larger umbrella of human-like attacks.

3 Experiments

A neural text dataset containing 5,000 text samples generated by GPT-2 1.5B using top-k 40 sampling was used to evaluate attacks in all experiments. This dataset was taken from a GitHub repository.[2] In all experiments, except for the transferability tests, an open source implementation of the automatic RoBERTa$_{\text{LARGE}}$ neural text detector[3] was used. Before the attacks, RoBERTa$_{LARGE}$'s recall on neural text was 97.44%. In this paper, five experiments testing homoglyph attacks were conducted, and two were conducted for misspelling attacks.

The first homoglyph experiment in this paper was designed to test the effectiveness of different homoglyph pairs in lowering detector recall on neural text. In this experiment, all attacks were restricted to randomly replacing 1.5% of all the characters in a given neural text sample to homoglyphs. If there were not enough of the character(s) being replaced in a neural text sample to meet this 1.5% quota, the text sample was thrown out and the result of the attack not considered. Even so, every attack in experiments conducted under these conditions was run on at least 2,500 neural text samples.

The second homoglyph experiment took the most effective homoglyph pair found in the first experiment and tested the effectiveness of the homoglyph attack when it was allowed to replace every occurrence of the target character(s).

The third homoglyph experiment was designed to take the most effective homoglyph pair and test how varying frequencies of replacement may affect detector recall on neural text.

The fourth homoglyph experiment was designed to test the transferability of the homoglyph attacks to the GROVER and GLTR online demos. [4] [5] In this experiment, 20 samples of neural text were randomly selected from the neural text dataset. Then, the most effective homoglyph attack (found in the first homoglyph experiment) was applied to the samples. GROVER's predictions on the original neural text and modified neural text were then recorded. The online demo for GROVER outputs "We are quite sure this was written by a machine," (Machine++) "We think this was written by a machine (but we're not sure),"(Machine+) "We think this was written by a human (but we're not sure)," (Human+) or "We are quite sure this was written by a human" (Human++). A similar experiment was performed on the GLTR demo. The most successful homoglyph attack was applied to 10 samples of text taken randomly from the neural text dataset.[6] Screenshots of GLTR's graphical interface were then taken before and after the attack, and patterns were observed.

[2] https://github.com/openai/gpt-2-output-dataset

[3] https://github.com/openai/gpt-2-output-dataset/tree/master/detector

[4] https://grover.allenai.org/detect

[5] http://gltr.io/dist/index.html

[6] The GLTR interface does not allow many Unicode characters, including Cyrillic ones. Thus, the homoglyph attack used for the GLTR experiments was the most successful, GLTR-allowed homoglyph attack.

Original	Homoglyph	Detector Recall	Average Confidence
		97.44%	5.29%
a (U+0061), e (U+0065)	а (U+0430), е (U+0435)	13.57%	81.61%
e (U+0065)	е (U+0435)	16.11%	79.43%
e (U+0065)	é (U+00E9)	18.11%	77.42%
a (U+0061), c (U+0063)	а (U+0430), с (U+0441)	19.96%	75.98%
a (U+0061)	а (U+0430)	20.40%	75.55%
c (U+0063)	с (U+0441)	36.94%	61.78%
p (U+0070)	р (U+0440)	42.25%	56.99%

Table 1: RoBERTA$_{LARGE}$ recall on neural text and average confidence RoBERTA$_{LARGE}$ predicted human-written with using various homoglyph pairs (with corresponding Unicode codes in parentheses). First row contains results on unaltered neural text. Unfortunately, default LaTeXdoes not support all Unicode characters. However, a package was used to render the Cyrillic characters in this table. Examples of actual Cyrillic Unicode characters displayed can be seen in Appendix B Figure 4.

For the misspelling attack experiments, words were randomly misspelled throughout a text sample using a Wikipedia list[7] of commonly misspelled words (by humans) in the English language. The attack was restricted to randomly misspelling 5% of the words in each neural text sample in the dataset. The same transferability experiments used in the homoglyph attacks were used for the misspelling attacks, except instead of characters being replaced with homoglyphs, a random 5% of the words in neural text samples were misspelled.

Code to reproduce results found in this paper can be found at `https://github.com/mwolff31/attacking_neural_text_detectors`.

4 Results

Results for the first homoglyph experiment can be seen in Table 1. Interestingly, replacing vowels with homoglyphs was a much more effective attack, even when the frequency of replacement was the same as that of consonants. Additionally, attacks using multiple homoglyph pairs were more effective than those which used only one.

For the second homoglyph experiment, according to Table 1, the most successful homoglyph pair was English “e” and English “a” to Cyrillic “Ye” and Cyrillic “a”, respectively. When this homoglyph attack was allowed to replace all of the English “e”s and English “a”s in the neural text dataset, RoBERTa$_{LARGE}$’s recall on neural text dropped to 0.26%.

The results of the third homoglyph experiment can be seen in Figure 1. The most successful single character homoglyph attack was used. Neural text detector recall on neural text was inversely proportional to the amount of characters a homoglyph attack was allowed to replace.

The results of the fourth homoglyph experiment indicate that the homoglyph attacks are transferable to other neural text detectors. Before the English “e” and English “a” to Cyrillic “Ye” and Cyrillic “a” attack was implemented, GROVER predicted Machine++ for 19 of the 20 samples, and predicted Human++ for 1 of the 20 samples. After the homoglyph attack, GROVER predicted Machine++ for 3 of the 20 samples, Machine+ for 1 of the 20 samples, Human+ for 1 of the 20 samples, and Human++ for the remaining 15 samples. In an experiment testing the transferability of the homoglyph attack to GLTR, replacing all English “e”s with Latin “é”s across 10 neural text samples consistently shifted histograms and the way GLTR colored the given text in the online demo towards GLTR behavior characteristic of human writing. Graphical results can be seen in Appendix B.

The results of the second misspelling experiment indicate that the misspelling attack is transferable to other neural text detectors as well. Before the misspelling attack was implemented, GROVER predicted Machine++ for 19 of the 20 samples, and predicted Human++ for 1 of the 20 samples.

[7] `https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines`

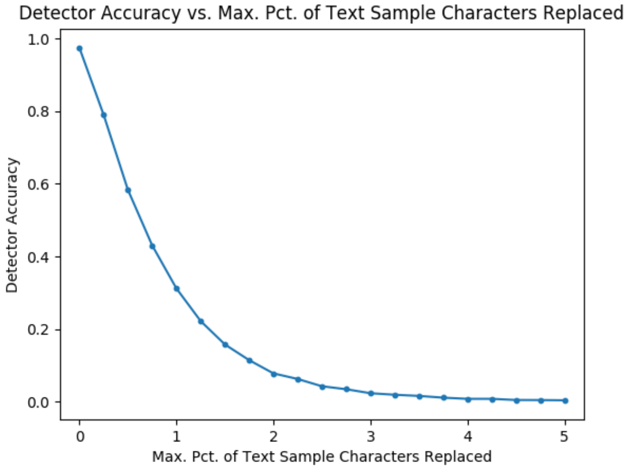


Figure 1: RoBERTa$_{LARGE}$ neural text recall on neural text as a function of the maximum percentage of neural text sample characters a random homoglyph English "e" to Cyrillic "Ye" attack was allowed to replace.

Note that random samples different from the ones used for the homoglyph transferability attack were used. After the homoglyph attack, GROVER predicted Machine++ for 8 of the 20 samples, Machine+ for 2 of the 20 samples, Human+ for 1 of the 20 samples, and Human++ for the remaining 9 samples. Similarly, the misspelling attack was also able to shift GLTR behavior towards that characteristic of humans across 10 neural text samples. An example of this can be seen in Appendix B.

5 Discussion

It is interesting to note that the non-human like attacks were effective because they are not characteristic of human-written nor neural text, yet the neural text detectors predicted the text was human-written–just because the modified neural text wasn't characteristic of neural text. Clearly, automatic neural text detectors are trained not to discriminate between neural text and human-written text, but rather decide what is characteristic and uncharacteristic of neural text. As seen by the success of the homoglyph attacks presented in this paper, this creates a vulnerability for neural text detectors in which an adversary can change neural text to be characteristic of neither language models nor humans (e.g. mixing English and Cyrillic alphabets), yet have the modified neural text be classified as human-written text.

While homoglyph attacks may be defended against with similar tactics such as those employed by modern web-browsers and spell-check, human-like attacks will ultimately be much more difficult to defend against, especially as they increase in complexity and employ methods which create not just spelling errors, but also grammatical errors or different sampling mechanisms to encourage different word-choice. Such attacks will force neural text detectors to increasingly deepen their understanding of not only what constitutes neural text, but also what constitutes human-written text.

6 Conclusion

This work defines two classes of attacks on neural text detectors: non human-like and human-like. Both proved to be very effective in disrupting neural text detectors' ability to classify neural text accurately. Additionally, this paper sheds some light on what kinds of methods neural text detectors employ, and how these may be exploited. Future work should focus on making neural text detectors robust against the attacks presented in this work, and further explore the extent to which the attacks presented in this paper, particularly human-like attacks, may be deployed on neural text detectors.

References

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 284–293, 2018.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36. Association for Computational Linguistics, 2018.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116. Association for Computational Linguistics, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1528–1540, 2016.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language models, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*, 2019.

A Shifting Neural Text's Distribution

This experiment, similar to the ones performed by the GLTR authors, was designed to quantify the extent to which a homoglyph attack could shift the distribution of neural text away from that of text produced by a language model. The GPT-2 117M language model[8] was used to generate predictions for each token in a text sample. The token's position within GPT-2 117M's predictions, or rank, was then recorded. Lower ranks indicate an alignment with GPT-2 117M's predictions. 50 randomly chosen text samples from the WebText dataset made available in the same GitHub repository that provided the neural text dataset[9] was used for human evaluation. 50 text samples were randomly chosen from the neural text dataset, which then had the English "e" and English "a" to Cryllic "Ye" and Cyrillic "a" homoglyph attack with no maximum character replacement restriction applied to them. The results for the this experiment are displayed in Table 2. Overall, the homoglyph attack was successful in shifting neural text's distribution away from that of a language model.

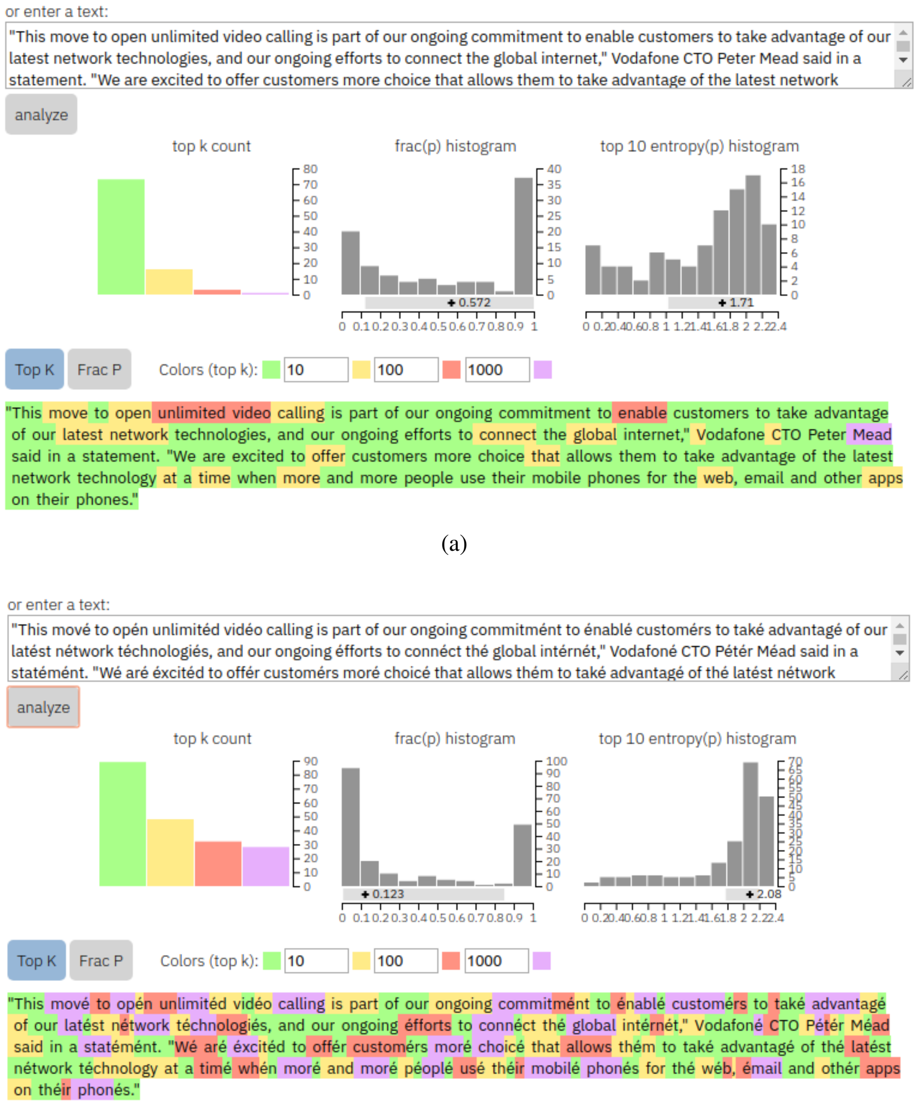
[8] An open-sourced GPT-2 117M model was taken from `https://github.com/huggingface/transformers`

[9] `https://github.com/openai/gpt-2-output-dataset`

	Rank
Human	128.26
Neural	14.05
Homoglyph Neural	371.41

Table 2: Average rank of each word in GPT-2 117M predictions across 50 randomly chosen text samples for each category.

B EXAMPLES



"This move to open unlimited video calling is part of our ongoing commitment to enable customers to take advantage of our latest network technologies, and our ongoing efforts to connect the global internet," Vodafone CTO Peter Mead said in a statement. "We are excited to offer customers more choice that allows them to take advantage of the latest network

Colors (top k): 🟩 10 🟨 100 🟥 1000 🟪

"This move to open unlimited video calling is part of our ongoing commitment to enable customers to take advantage of our latest network technologies, and our ongoing efforts to connect the global internet," Vodafone CTO Peter Mead said in a statement. "We are excited to offer customers more choice that allows them to take advantage of the latest network technology at a time when more and more people use their mobile phones for the web, email and other apps on their phones."

(a)

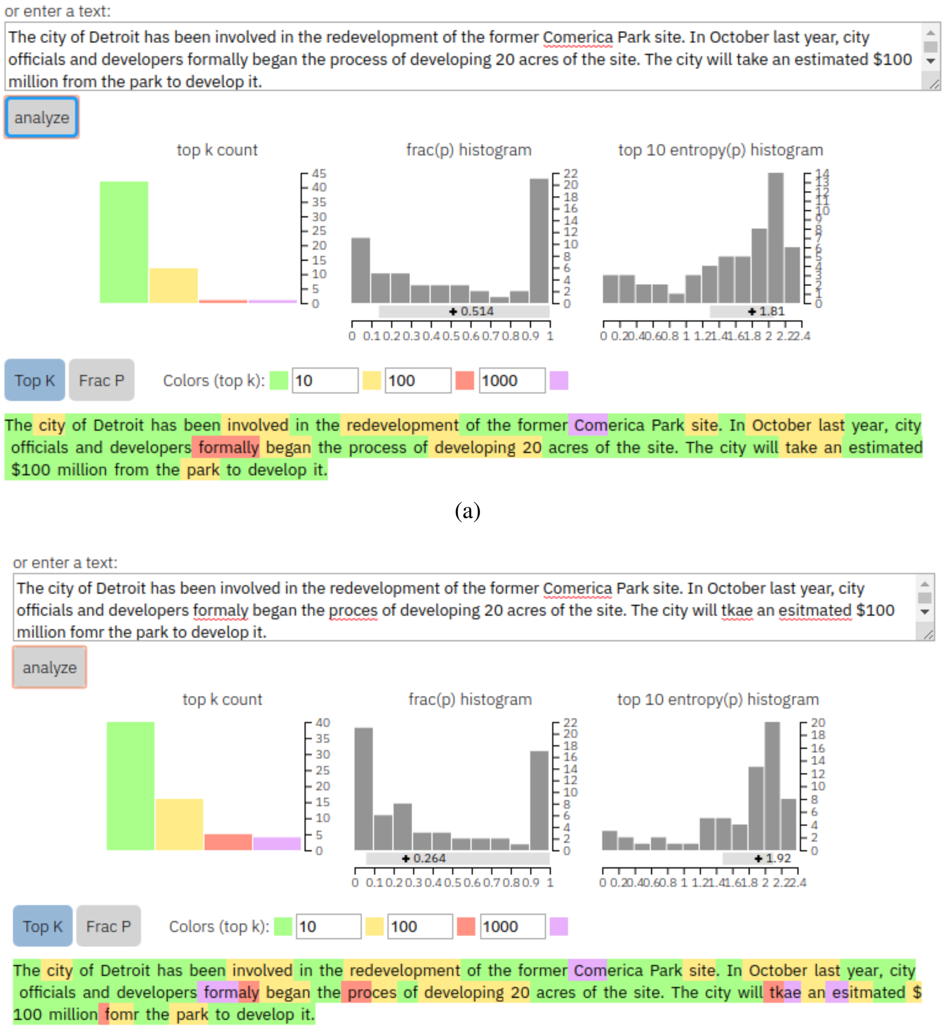
"This mové to opén unlimitéd vidéo calling is part of our ongoing commitmént to énablé customérs to také advantagé of our latést nétwork téchnologiés, and our ongoing éfforts to connéct thé global intérnét," Vodafoné CTO Pétér Méad said in a statémént. "Wé aré éxcitéd to offér customérs moré choicé that allows thém to také advantagé of thé latést nétwork

Colors (top k): 🟩 10 🟨 100 🟥 1000 🟪

"This mové to opén unlimitéd vidéo calling is part of our ongoing commitmént to énablé customérs to také advantagé of our latést nétwork téchnologiés, and our ongoing éfforts to connéct thé global intérnét," Vodafoné CTO Pétér Méad said in a statémént. "Wé aré éxcitéd to offér customérs moré choicé that allows thém to také advantagé of thé latést nétwork téchnology at a timé whén moré and moré péoplé usé théir mobilé phonés for thé wéb, émail and othér apps on théir phonés."

(b)

Figure 2: Top: GLTR output before homoglyph attack. Bottom: GLTR output after homoglyph attack. The presence of red and purple highlighted words indicates that GPT-2 117M had a difficult time predicting the word being highlighted, which helps human readers decide whether text was written by a language model or human.



or enter a text:

The city of Detroit has been involved in the redevelopment of the former Comerica Park site. In October last year, city officials and developers formally began the process of developing 20 acres of the site. The city will take an estimated $100 million from the park to develop it.

Colors (top k): 🟩 10 🟨 100 🟥 1000 🟪

The city of Detroit has been involved in the redevelopment of the former Comerica Park site. In October last year, city
officials and developers formally began the process of developing 20 acres of the site. The city will take an estimated
$100 million from the park to develop it.

(a)

or enter a text:

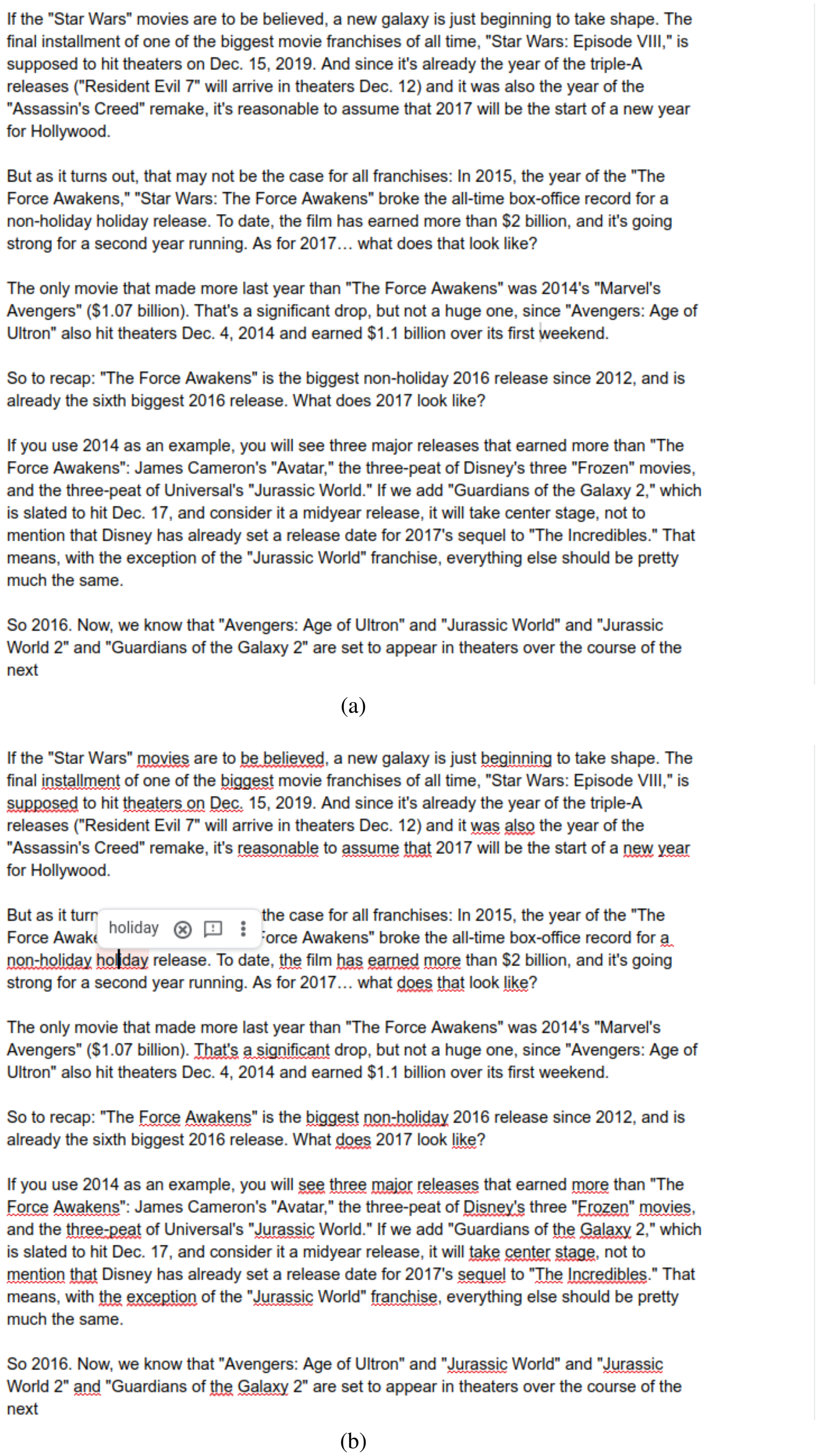
The city of Detroit has been involved in the redevelopment of the former Comerica Park site. In October last year, city officials and developers formaly began the proces of developing 20 acres of the site. The city will tkae an esitmated $100 million fomr the park to develop it.

Colors (top k): 🟩 10 🟨 100 🟥 1000 🟪

The city of Detroit has been involved in the redevelopment of the former Comerica Park site. In October last year, city
officials and developers formaly began the proces of developing 20 acres of the site. The city will tkae an esitmated $
100 million fomr the park to develop it.

(b)

Figure 3: GLTR before (Subfigure (a)) and after (Subfigure (b)) the misspelling attack was applied.



If the "Star Wars" movies are to be believed, a new galaxy is just beginning to take shape. The final installment of one of the biggest movie franchises of all time, "Star Wars: Episode VIII," is supposed to hit theaters on Dec. 15, 2019. And since it's already the year of the triple-A releases ("Resident Evil 7" will arrive in theaters Dec. 12) and it was also the year of the "Assassin's Creed" remake, it's reasonable to assume that 2017 will be the start of a new year for Hollywood.

But as it turns out, that may not be the case for all franchises: In 2015, the year of the "The Force Awakens," "Star Wars: The Force Awakens" broke the all-time box-office record for a non-holiday holiday release. To date, the film has earned more than $2 billion, and it's going strong for a second year running. As for 2017… what does that look like?

The only movie that made more last year than "The Force Awakens" was 2014's "Marvel's Avengers" ($1.07 billion). That's a significant drop, but not a huge one, since "Avengers: Age of Ultron" also hit theaters Dec. 4, 2014 and earned $1.1 billion over its first weekend.

So to recap: "The Force Awakens" is the biggest non-holiday 2016 release since 2012, and is already the sixth biggest 2016 release. What does 2017 look like?

If you use 2014 as an example, you will see three major releases that earned more than "The Force Awakens": James Cameron's "Avatar," the three-peat of Disney's three "Frozen" movies, and the three-peat of Universal's "Jurassic World." If we add "Guardians of the Galaxy 2," which is slated to hit Dec. 17, and consider it a midyear release, it will take center stage, not to mention that Disney has already set a release date for 2017's sequel to "The Incredibles." That means, with the exception of the "Jurassic World" franchise, everything else should be pretty much the same.

So 2016. Now, we know that "Avengers: Age of Ultron" and "Jurassic World" and "Jurassic World 2" and "Guardians of the Galaxy 2" are set to appear in theaters over the course of the next

(a)

If the "Star Wars" movies are to be believed, a new galaxy is just beginning to take shape. The final installment of one of the biggest movie franchises of all time, "Star Wars: Episode VIII," is supposed to hit theaters on Dec. 15, 2019. And since it's already the year of the triple-A releases ("Resident Evil 7" will arrive in theaters Dec. 12) and it was also the year of the "Assassin's Creed" remake, it's reasonable to assume that 2017 will be the start of a new year for Hollywood.

But as it turn the case for all franchises: In 2015, the year of the "The Force Awake orce Awakens" broke the all-time box-office record for a non-holiday holiday release. To date, the film has earned more than $2 billion, and it's going strong for a second year running. As for 2017… what does that look like?

The only movie that made more last year than "The Force Awakens" was 2014's "Marvel's Avengers" ($1.07 billion). That's a significant drop, but not a huge one, since "Avengers: Age of Ultron" also hit theaters Dec. 4, 2014 and earned $1.1 billion over its first weekend.

So to recap: "The Force Awakens" is the biggest non-holiday 2016 release since 2012, and is already the sixth biggest 2016 release. What does 2017 look like?

If you use 2014 as an example, you will see three major releases that earned more than "The Force Awakens": James Cameron's "Avatar," the three-peat of Disney's three "Frozen" movies, and the three-peat of Universal's "Jurassic World." If we add "Guardians of the Galaxy 2," which is slated to hit Dec. 17, and consider it a midyear release, it will take center stage, not to mention that Disney has already set a release date for 2017's sequel to "The Incredibles." That means, with the exception of the "Jurassic World" franchise, everything else should be pretty much the same.

So 2016. Now, we know that "Avengers: Age of Ultron" and "Jurassic World" and "Jurassic World 2" and "Guardians of the Galaxy 2" are set to appear in theaters over the course of the next

(b)

Figure 4: The text in Subfigure (a) was classified as neural text by the neural text detector with 99.94% confidence. Then, the English "e" and English "a" to Cyrillic "Ye" and Cyrillic "a" attack was applied to it. After this modification, the text sample, shown in Subfigure (b) was classified by RoBERTa$_{LARGE}$ as human-written with 98.50% confidence. Both text samples are rendered in Google Docs. Interestingly, Google Docs's spell-correct program provides suggestions (marked by red underlining) to change SOME but not ALL of the words the attack modified back to their unaltered state.