Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

GENERATING ADVERSARIAL EXAMPLES FOR MISINFORMATION DETECTION USING AUTOMATED TEXT SIMPLIFICATION

Martina Gómez Martín

Tutor: Piotr Przybyla

Text Classification Algorithms

- Widely used for many tasks (information retrieval, sentiment analysis, ...)
- Reduced the workload and increased agility.
- In misinformation detection:
  - Online platforms to filter user-generated content.
  - Government Units to avoid the spread of misinformation during crises.
- Algorithms present vulnerabilities.

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Adverarial Examples

Inputs to machine learning models that are intentionally designed to cause the model to make a mistake

Text Simplification as a **novel** approach for generating Adversarial Examples

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Summary

1. **Choosing simplification technique**

   Study Automated Text Simplification techniques and choose a suitable tool for our experiments.

2. **Manual evaluation**

   Simplify texts for different misinformation detection tasks and manually analyze the results.

3. **Performing adversarial attack**

   Perform attacks on two victim models using the simplification tool selected for different misinformation tasks.

4. **Evaluation of attack results**

   Study the numerical results and manually analyze successful cases.

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Tools

MUSS

- Multilingual Unsupervised Learning by Mining Paraphrases.
- Trains models using sentence-level paraphrase data.
- No reliance on labeled simplification data.
- Uses control tokens to focus on specific features (output length, similarity, word rank, tree depth).
- Text Preprocessing
  - Segmentation with LAMBO.

BODEGA

- Framework for evaluating attacks on Missinformation Detection scenarios.
- Based on OpenAttack.
- Aims to standardize attack evaluation.

Misinformation tasks

HN	PR	FC	RD
Hyperpartisan News Detection	Propaganda detection	Fact checking	Rumor detection

• Detect fake news based on writing style.	• Identify written content in which the writer attempts to influence the reader's opinion.	• Input text consists of a pair of texts:
		  ◦ Claim
		  ◦ Evidence
		• Identify if evidence corroborates or contradicts claim
			• Detect information that is spreading between people despite not having a reliable source.

Manual Evaluation

Simplicity

Degree of simplification in relation to the original.

Fluency

Degree of grammatical and syntactic correctness.

Meaning Preservation

Level of retention of the content.

Some examples

Original	Simplified	S	F	MP
Progressively-minded people do not make political stances because they're young and that's what young people do, it's because of an overarching desire to better our nation by providing the social, economic, and political liberties that our current incumbents have repeatedly failed to do. Wanting the right for every single person to go to school regardless .... but because they are decent human beings who value people over profit.	Progressively-minded people don't take political positions because they're young, because that's what young people do. They do it because they want to better our country by giving social, economic, and political rights that our current leaders have failed to do. Wanting the right for every person to go to school no matter .... It is a choice they make because they are good people.	3	3	3

- Example of simplified text of HN task.
- Good word substitution and sentence splitting.

Some examples

Original	Simplified	S	F	MP
Whataburger told Fox News that it is cooperating with the police investigation.	Whataburger told Fox News that it is helping police with their investigation.	3	3	3

- Example of simplified text of PR task.
- As it is a short sentence, little changes are needed.
- Good word substitution.

Some examples

Original	Simplified	S	F	MP
One can condemn the #CharlieHebdo killings, while conjointly condemning **racism, xenophobia, and Islamophobia** framed as harmless cartoons.	One can condemn the #CharlieHebdo killings while simultaneously condemning **Islamophobia, racism and xenophobia** framed as harmless cartoons.	1	3	3

- Example of simplified text of RD task.
- Changes do not make the text simpler.

Some examples

Original	Simplified	S	F	MP
The Guardian has been vocal in the so-called “fake news” hysteria, decrying the influence of social media, the only place where leftwing dissidents have managed to find a small foothold to promote their politics and counter the corporate media narrative.	The Guardian has been vocal in its opposition to the so-called “fake news” hysteria, decrying the influence of social media. The only place leftwing dissidents have found a small foothold is on social media to promote their politics.	2	3	2

- Example of simplified text of PR task.
- Some changes are made such as sentence splitting.
- Not enough modifications are done as some words are too complex.

Some examples

Original	Simplified	S	F	MP
Why was Cardinal Mahony allowed to retire in good standing?	Why was Cardinal Mahony allowed to stay in office?	3	3	1

- Example of simplified text of PR task.
- Changes alter the intended meaning of the original sentence.

Some examples

Original	Simplified	S	F	MP
Martin Van Buren. He then moved on Washington where he did more than anyone to construct the modern Democratic Party which dominated American politics down to the American Civil War. ~ Martin Van Buren was exclusively unelected	Martin Van Buren. He then moved on Washington where he did more than anyone to build the modern Democratic Party that dominated American politics until the American Civil War. ~ Martin Van Buren was elected only once.	3	3	1

- Example of simplified text of FC task.
- Some good modifications are made.
- Changes alter the intended meaning of the original sentence.

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Remarks on Manual Evaluatioan

- Tool effectively improves text comprehensibility by:
  - Substituting complex words.
  - Restructuring and breaking down sentences.
  - Eliminating unnecessary parts.

- Challenges:
  - In some cases alterations made are insufficient.
  - In other cases changes might alter text's intended meaning.

- Alterations to the evidence of Fact Checking texts modify the original label.

Attack scenario

- 2 victim models
  - Fine-tuned language model: BERT
  - Recurrent neural network: BiLSTM
- 4 misinformation detection tasks.
- Grey-box scenario.
- Untargeted and targeted.
- Infinite number of queries are assumed.

Evaluation Metrics

SUCCESS SCORE	SEMANTIC SCORE	CHARACTER SCORE	BODEGA SCORE	NUMBER OF QUERIES
• Proportion of successful attacks.	• Measure of meaning preservation. • Based on BLEURT.	• Expresses how different one string of characters is from another. • Uses Levenshtein distance.	• Measures the quality of adversarial modifications in successful attacks.	• It is an average of the number of queries made to the model per example.

Results

- **SUCCESS score**

	HN	RD	FC	PR
	TARGETED			
BERT	0.13	0.10	0.19	**0.38**
BiLSTM	0.15	0.10	0.24	**0.33**
	UNTARGETED			
BERT	0.17	0.06	0.18	0.08
BILSTM	0.20	0.06	0.24	0.16

- **BODEGA score**

	HN	RD	FC	PR
	TARGETED			
BERT	0.07	0.03	0.08	**0.15**
BiLSTM	0.07	0.04	0.112	**0.16**
	UNTARGETED			
BERT	0.08	0.02	0.07	0.04
BILSTM	0.09	0.07	0.11	0.08

- Results consistent with BODEGA framework paper.
- Fine-tune BERT is more robust than BiLSTM.
- Difficulty of the attack is similar in both targeted and untargeted scenarios.
  - When the victim performs poorly, targeted attack becomes easier.
- Task with shorter input text are more vulnerable to modifications.

Results

Results of adversarial attacks on the BiLSTM attack with PR task.

	Untargeted					Targeted				
Method	**B.**	**con**	**sem**	**char**	**Q.**	**B.**	**succ**	**sem**	**char**	**Q.**
PR2										
BAE	0.15	0.23	0.72	0.94	32.94	0.26	0.38	0.71	0.94	38.72
BERT-ATTACK	0.53	0.80	0.72	0.91	61.41	**0.66**	0.94	0.74	0.94	50.14
DeepWordBug	0.29	0.38	**0.79**	**0.96**	27.45	0.56	0.72	**0.81**	**0.96**	35.30
Genetic	**0.54**	**0.88**	0.67	0.89	782.15	0.62	0.94	0.71	0.93	802.20
SememePSO	0.47	0.76	0.68	0.89	85.34	0.60	0.92	0.71	0.92	69.62
PWWS	0.53	0.84	0.69	0.90	130.85	0.63	0.92	0.73	0.94	168.60
SCPN	0.12	0.55	0.39	0.50	11.55	0.20	**0.98**	0.37	0.48	11.98
TextFooler	0.51	0.85	0.67	0.88	52.59	0.63	0.94	0.72	0.92	54.62
MUSS	0.16	0.33	0.70	0.69	2.16	0.08	0.16	0.72	0.69	2.16

- MUSS does not require many queries to come up with AE.
- Semantic and character scores have average values with respect to the rest of the attacks.

Study of successful adversarial examples

Evaluation of Successful Cases

Original	Simplified	S	F	MP
Farrell's **graciously offers notorious praise** for Martin's new book, Building a Bridge (to hell, ed).	Farrell has also **given good reviews** to Martin's new book, Building a Bridge (to hell, ed).	3	3	3

- Task PR
- Original label 1 --> predicted 0
- Good simplification / Good adversarial attack

Evaluation of Successful Cases

Original	Simplified	S	F	MP
Antisemitism at McGill does not emerge **from a vacuum**.	Antisemitism at McGill does not happen **by accident**.	3	3	3

- Task PR
- Original label 0 --> predicted 1
- Good simplification / Good adversarial attack

Evaluation of Successful Cases

Original	Simplified	S	F	MP
Mani Ratnam. Cited by the media as one of India's **influential** filmmakers, Mani Ratnam is widely credited with **revolutionising** the Tamil film industry and altering the profile of Indian cinema. Mani Ratnam is widely credited with **altering** the profile of Indian cinema.	Mani Ratnam. Cited by the media as one of India's most **important** filmmakers, Mani Ratnam is widely credited with **changing** the Tamil film industry **for the better**. Mani Ratnam's films have **changed** the profile of Indian cinema.	3	3	3

- Task FC.
- Original label 0 --> predicted 1.
- Good simplification / Good adversarial attack.

Evaluation of Successful Cases

Original	Simplified	S	F	MP
Janet Leigh. Janet Leigh (born Jeanette Helen Morrison, July 6, 1927 – October 3, 2004) was an American actress, singer, dancer and author. Janet Leigh was **Canadian.**	Janet Leigh. Janet Leigh (born Jeanette Helen Morrison, July 6, 1927 – October 3, 2004) was an American movie and television actress. Janet Leigh is **American.**	2	3	1

- Task FC.
- Original label 1 --> predicted 0.
- Evidence is incorrectly simplified, making it very easy to misclassify.
- correct response does not remain the same

Evaluation of Successful Cases

Original	Simplified	S	F	MP
Let the hostages go first, but OK. RT jaketapper: **Charlie Hebdo suspects tell police** they want to die as martyrs **http://t.co/cb9zSckTZ2** @JasStanford @WhitneyNeal @jaketapper Only if those Lady Kurd soldiers can do it, **with bacon-laced ammunition.** @JasStanford @WhitneyNeal @jaketapper Just saying, Danielle Mitterrand was a huge fan of the Lady Kurds. Bet **they'd volunteer.**	Let the hostages go first, but OK. RT @jaketapper: **Police tell Charlie Hebdo suspects** they want to die as martyrs. @JasStanford @WhitneyNeal @jaketapper Only if those Lady Kurd soldiers can do it, right? @JasStanford @WhitneyNeal @jaketapper Just saying, Danielle Mitterrand was a Lady Kurds fan. **I'd volunteer.**	2	3	1

- Task RD.
- Original label 1 --> predicted 0.
- Modifications change text intended meaning.

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

SUMMARY RESULTS

1. **BERT is more robust than BiLSTM**

   Attack results to BERT model are worse

2. **Shorter texts easily confuse the victim model**

   Modifications to texts that are shorter easily lead to a different prediction

3. **Bad simplifications easily mislead the algorithm's decision.**

   If the meaning is not preserved, the algorithm will most likely change its prediction.

4. **Number of queries needed is low**

   In contrast with other attacks studied the number of queries required for successful attacks is very low around 2-3.

Conclusions

Discussion

- Limitations with MUSS due to lack of output variation.
- Inaccuracies with the simplification tool might lead to incorrect adversarial examples.

Future Work

- Simplification has potential as a technique for generating adversarial examples.
- Use LLM to improve simplification results.
- Expand scope of the tasks for validating the generalizability of the proposed techniques

Universitat Pompeu Fabra Barcelona

Escola d'Enginyeria

FAQ

Thank You!

E-mail

martina.gomez01@estudiant.upf.edu