IS-7033-Topics: Natural Language Processing (NLP) –Proposal and Literature Review

John Rodriguez and **Ian Scarff**
University of Texas at San Antonio
One UTSA Circle, San Antonio, TX 78249

Abstract

This project proposal and literature review describes and outlines the Natural Language Processing (NLP) initiative as part of the IS-7033 course. The project proposal will cover the following topics: the research question to be examined, project motivation and rational as an interesting question worth asking, and assess its potential to contribute to new knowledge by situating it within related literature in the scientific community.

1 Introduction

With the democratization of the public forum as a result of the proliferation of the internet starting in the 1990s, the control of the traditional media (i.e. television, radio, publishing companies, etc) was slowly loosened up over the next decade. The initial years brought information sharing to the masses and, as a principal, this is a win for free speech and democracies globally. However, with zero controls in place, due to the spirit of openness and good intentions, it was only a matter of time before anti-democratic agents found a way to use this spirit of openness to create chaos and lawlessness.

One medium commonly used for 'fake-news' / misinformation is the internet meme or meme for short. A meme is defined as "an amusing or interesting item (such as a captioned picture or video) or genre of items that is spread widely online especially through social media." This media can be easily transmitted over social media networks and reach vast numbers of users. This type of information dissemination can be classified as propaganda. The goal of propaganda is to spread misinformation to achieve a certain goal. Propaganda publishers use psychological (emotional appeal) and rhetorical techniques (logical fallacies) to capture their audiences and illicit certain responses and behaviors. By appealing to audience emotions and by misusing logical rules, propagandist can induce audiences to form an emotional bond thus suspending rational analysis and come to the wrong conclusion. Figure 1 shows a comparison between a relatively innocent meme and a more targeted meme. Deceptive memes efficiently enable propagandists by using an image that is blended with text to "reinforce/complement a technique in the text" to enable one or several "persuasive techniques."

Figure 1: The image on the left (taken from krqe.com) shows a playful take on Bernie Sanders during the 2021 presidential inauguration. The image on the right (taken from the Semeval 2021 Task 6 development images) shows a more targeted meme meant to illicit negative responses or behaviors.

The goal of the SemEval 2021 Task 6 and the focus of our project is to perform the following empirical research:

- Task 1 - Multi-label Classification Problem - Using only the text component of a meme, identify which of the 20 techniques are utilized.
- Task 2 - Multi-label Sequence Tagging - Using only the text component of the meme, identify which of the 20 techniques are utilized and also the spans of text covered by each technique. - **Note: This task will not be covered by this project but it was included for completeness.**

- Task 3 - Multi-modal Problem - Using both image and text component of a meme, identify which 22 techniques are utilized.

2 Research Question

Machine learning engineers are tasked with training computers to learn and understand various domains of the human experience. A sub-field of linguistics, computer science and artificial intelligence (AI), natural language processing (NLP) is the computer understanding and manipulation of human based languages. Because human language can be conveyed across multiple human senses (sight, sound, and touch through Braille), there are many avenues to pursue for research. With regards to this project, the research proposed is as follows: Classify the various techniques of propaganda in memes using multi-modal AI methods. These methods will include models and algorithms used in NLP and computer vision.

3 Literature Review

There have been previous studies regarding the multi-modal use of online text and imagery data. (Kruk et al., 2019) demonstrated the use of multi-modal architectures on Instagram images and their associated captions to classify them based on three taxonomies: intent, contextual, and semiotic. They also showed that using both the image data and the text data together to make predictions leads to better accuracy when compared to using the data separately. Additionally, memes present a challenge for NLP due to their multi-modality nature in conjunction with the usage of humor and sarcasm (Suryawanshi et al., 2020).

Memes specifically were the subject of study in Facebook's *Hateful Memes Challenge* (Kiela et al., 2020). With the vastness of the internet, it is practically impossible for humans to sift through countless of memes to identify hate speech. Thus, there is the need for automation through AI. The goal of this challenge was to construct and implement multi-modal models for classifying hate speech. The authors of the challenge designed their dataset such that uni-modal models would under perform multi-modal models. Even then, as a baseline, they showed that even state-of-the-art models struggled compared to humans. Over 3,000 people and teams participated with (Lippe et al., 2020; Velioglu and Rose, 2020; Muennighoff, 2020; Zhong, 2020) placing 4th, 3rd, 2nd, and 1st in the challenge, respectively.

In addition to the challenge, memes are still a very open research topic. (Afridi et al., 2020) conducted a survey to offer a generalized view of the challenges modeling memes present, what the current advanced techniques for classification are, and discuss ongoing research into memes classification, memes reasoning, memes semantic entailment, and multi-modal fusion and co-learning. Furthermore, other research such as *Detecting Hate Speech in multi-modal Memes* (Das et al., 2020) and *Detecting Hateful Memes Using a Multi-modal Deep Ensemble* (Sandulescu, 2020) continue to prove that memes and multi-modal classification are in further need of innovation and automation due to the complexity of problem. A major driving force is to realize human-like level accuracy for this problem while also maintaining performance of the implementation system.

4 Project Motivation

Given the proliferation of propaganda using modern channels such as the internet and various subchannels as social media, 4chan, dark web, etc., a need has arisen to identify memes that contain harmful messages that spread false information, drive violent behavior, and lead to lawlessness. The goal of this project is to utilize NLP and computer vision to efficiently and accurately classify memes that are considered to be propaganda.

The focus of this project will be multi-modal multi-label classification of structured representation of images and text. Multi-modal can be defined as communication that uses multiple semiotic systems. Semiotic is the method of how meaning is communicated. In the case of this project, images and text. The task is to model and classify them according to a particular set of domain values (listing of propaganda techniques). The images and text can potentially have multiple domain values assigned (multi-label classification). The data provided by Semeval 2021 is collected and stored as as a json structured object. Each entry consists of the text in a meme, the various propaganda technique labels assigned to the meme, and an associated image file name.

This project is considered the end of the line as it the last part of a bigger effort for propaganda identification, which (in our view), is as follows:

1. Gather memes from source (Facebook, Twitter, Instagram, 4chan, Dark Web, Other).

2. Determine meme type.
    (a) non-propaganda (i.e. funny, an exaggeration, truthful/factual, etc.)
    (b) ambiguous (i.e. it might might be heading to a bad place)
    (c) propaganda (i.e. false information, elicits violence, leads to lawlessness, etc)
3. Classification of ambiguous and propaganda type memes.

The overall motivation is to assist with identifying and labeling of propaganda to prevent misinformation, violence and lawlessness.

5 Contribution to Scientific Community

Human beings can process and understand vast amounts of data in a multi-modal format. Computers tend to do well at solving single modal learning problems; however there is a need to also solve multi-modal problems. This has been and continues to be a hot topic of current exploration. Text classification (language) is a very difficult problem in NLP to overcome due to various rules of the different human languages. When combined with image (vision) data, an even more difficult and computer resource intensive problem results. The goal of this project is to build a system that can be efficient while maintaining the required accuracy needed for multi-modal multi-label classification that can be used to determine if a meme is propaganda.

6 Conclusion

In this proposal, we have discussed the goal of this project, its motivation, and its contribution to the scientific community. Using natural language processing and computer vision methodologies, we plan to take on the SemEval 2021 Task 6 challenge. Our work will demonstrate the use of mutli-modal models for multi-label classification of memes. This project will contribute to NLP with respect to the ongoing study of meme classification and modeling.

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2020. A multimodal memes classification: A survey and open research issues.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. *CoRR*, abs/1904.09073.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.

Xiayu Zhong. 2020. Classification of multimodal hate speech – the winning solution of hateful memes challenge.