**Title:** Dockerized Knowledge-Oriented Multimodal Event Detection System

**Rationale:** As unstructured multimedia data rises exponentially, fast understanding of world events is essentially a more challenging task. Humans understand events by organizing them into frequently occurring narrative structures. These structures are abstracted as knowledge, the organized units of graphs that represent memory patterns used in human cognition. In general, the previous research work can be divided into two waves. Among them, the first wave of rule-based symbolic reasoning methods, such as conventional information extraction systems that use artificial feature extraction, cannot handle real-world event analysis[1-2]. At the same time, the second wave of machine learning or artificial intelligence-based systems requires too many manually generated annotated examples as training data to supervise machine learning so that it cannot meet the actual needs of event understanding[3-7]. (Note that because of page limitations, more references will be found in our paper.)

To solve these challenges, our research project aims to develop a knowledge-oriented Artificial Intelligence (AI) system that can recognize complex events and bring them to the attention of users to solve these challenges. ***Our research seeks to understand the complex events described in the multimedia input by developing a semi-automated system that identifies, links, and sorts its subsidiary elements, involved participants, and complex event types in chronological order.*** Inquisitive events can produce changes that significantly impact national security or participate in the causal chain that makes such impacts. ***In this ISEF research project, we propose a systematic analysis of world events: such as the Boston Marathon bombing, Capital Riots, Covid-19, etc.***

**Research Question(S), Hypothesis(Es), and Expected Outcomes:** There has been some initial progress in the event detection task. With the development of deep learning, multimedia information and natural language processing technologies have also made significant progress. For instance, Information Extraction (IE) has had some success with word2vector[3,4], transformers[6]. With the development of deep learning, multimedia information processing systems and text-based IE also have made huge progress in the past few years [7]. ***We hypothesize that knowledge-oriented multimedia data can help create an accurate analysis of events.*** ***Our research program analyzes complex events to understand social hotspots. The outcome of our research is to propose a Dockerized knowledge-oriented multimodal event detection system.*** (Note that because of page limitations, more references will be found in our paper.)

**Procedure:** In this paper, we will build a dockerized knowledge-oriented multimedia event detection system. There are four steps to achieving the following results and confirming our hypnosis. (Note that because of page limitations, more references will be found in our paper.)

1. ***Pre-processing:*** With the large amounts of unstructured multimodal data, pre-processing is a key step towards optimizing the outcome effectively. There are two general steps that are required for text, image, and video formats of the input.

* ​​Text pre-processing involves removing useless information from untouched text. The process eliminates information such as punctuation, stopwords, and frequent words, and rare words. It also uses the skill of lemminizing and stemming from converting words into their base form, such as changing the word “geese” to “gees” using lemminization and changing “gees” to “goose” using stemming.
* Keyframe extraction is a pre-processing step used on multimedia inputs with an MP4 format to extract the most important frames from the video and convert them into a JPG format. Because most videos have around 60 frames each second, we must eliminate useless information from the video by only detecting large changes within each frame and extracting the significant ones out. Here we use FFmpeg1 to extract and upload the JPGs. Cropping or trimming the video down may also be thumbnails, and audio may also be used here.

1. ***Information Extraction:*** After the preprocessing step, we move onto the step of Information Extraction (IE), which uses logical reasoning to analyze unstructured/semi-structured multimedia data to understand and summarize events. There are three steps in this process: Extract Entity, Events and Relations model; Named-Entity Recognition (NER); and Object Detection/Facial Recognition.

* We will try to extract Entity, Events, and Relations in this procedure as this is a subsection of Information Extraction work. It turns unstructured text in different domains into structured knowledge. A few challenges to this are overlapping and nested entities and long-ranged dependencies. This is usually done with pre-trained BERT with a layer of LSTM or fine-tuned BERT which will be explained later.
* Name-entity recognition is similar to Entity, Events, and Relations as it classifies the named entities in the unstructured text and turns it into predefined categories. Most approaches do it with conditional random fields. Here, we will identify, link, and sort its subsidiary elements, involved participants, and complex event types in chronological order.
* Object detection (image localization & image classification) and facial recognition are used for identifying objects and familiar faces. The first usually uses pre-trained CNN and either Tensorflow, PyTorch, or Keras. Facial recognition uses mostly similar styles as object detection, with the main difference being that there is holistic processing to identify a face in a certain image.

1. ***Multimodal Embedding***: NPL tasks have reached a higher level due to Google’s BERT (Bidirectional Encoder Representations from Transformers). It uses encoders and masked language models to become state-of-the-art-level models. Before moving onto inputting results into the neural network, we first have to put it through a multimodal embedding system, which typically relates a word in text/image/audio to a specific vector that has the meaning of the word so that similar words should be closer together. This involves three steps: Token Embedding, Segment Embedding, and Position Embeddings. Token embeddings are just vector representations of words. Segment embeddings are vector representations to show whether the vectors are similar or not. Position embedding will help BERT understand the difference between visually similar but semantically different words.
2. ***Knowledge-Oriented Event Detection***: Building a Knowledge Graph from text data uses sentence segmentation, dependency parsing, parts of speech tagging, and entity recognition to help both machine and us to understand the relationships between entities. After feeding inputs into the neural network, we build knowledge graphs. Then we have a data visualization framework that allows us to rapidly develop queries and visualizations of the data in the knowledge graphs. A barebones structure consists of two nodes and a relation (edge) in between. In the end, the output result should be a dockerized knowledge graph consisting of world events that could be easily understood by both the machine and humans.

**Risk and Safety:** The only risk and safety that should be considered in research are bias resulting from misinterpretation of analytical results. ***Using the wrong results may affect the understanding of events, leading to unreasonable decisions to believe in a particular person/idea.***

**Data Analysis:** Our research aims to understand the complex events in multimedia information by developing a system that identifies, links, and sorts its subsidiary elements, involved participants, and event types in chronological order. ***To reasonably evaluate the performance of our proposed model, followd by the previous work[7] we collected Wikipedia articles describing complex events on the Boston Marathon bombing, Capital Riots, and Covid-19 to evaluate our system in event extraction and tracking tasks.***

**Bibliography:**

[1] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004, May). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec* (Vol. 2, No. 1, pp. 837-840).

[2] Ji, H., & Grishman, R. (2008, June). Refining event extraction through cross-document inference. In *Proceedings of ACL-08: Hlt* (pp. 254-262).

[3] Li, Q., Ji, H., & Huang, L. (2013, August). Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 73-82).

[4] Nguyen, T. H., Cho, K., & Grishman, R. (2016, June). Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 300-309).

[5] Liu, S., Chen, Y., Liu, K., & Zhao, J. (2017, July). Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1789-1798).

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[7] Fung, Y., Thomas, C., Reddy, R. G., Polisetty, S., Ji, H., Chang, S. F., ... & Sil, A. (2021, August). Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1683-1698).