Key contributions of the project paper:

1. Generating annotated NER dataset for disaster management (1st of its kind)
2. Constructing word2vec-inspired contextual word embedding model – contextual vector representation of word tokens capturing semantic relationship between words in disaster domain.
3. Develop custom character embedding model using CNN to extract character-level features to recognise patterns and structures within words.
4. Create an optimised self-attention based Bi-LSTM CRF model
5. Evaluate performance of BiLSTM-ATTN-CRF model and investigate effectiveness of contextual word embeddings.
   1. Investigation – compare the use of context-specific word embedding features vs general word embedding features in recognising disaster related named entities.

Related work:

1. Recurrent Neural Networks (RNN) – text classification, named entity extraction, language modelling.
2. Active learning and CNN-CNN-LSTM model for NER task – model has high efficiency (computational & exceptional performance on standard dataset)
3. Bi-LSTM-CRF model -and transition stack LSTM model – did not rely on language specific resources or features.
4. Pooled-GRU model for NER in the Arabic language combined with Multilingual Universal Sentence Encoder – compared to Bi-LSTM-CRF model
5. Machine learning classifiers: Naïve Bayes, logistic regression, SVM, XGBoost, random forest used to classify news into relevant vs irrelevant categories.
6. LDA topic modelling combined with deep learning at both word and character level embeddings.
7. Thai NER using BiLSTM-CNN-CRF model.

Research Gaps:

1. Existing research focuses on using social media data as unstructured text for NER in disaster management context, while online news articles are considered reliable sources of information.
2. Analysis of context-specific embedding features on the performance of NER tasks is largely absent.

How are these gaps addressed?

1. Consider online news articles as semi-structured texts for extracting disaster-related named entities.
2. Construct context-specific word embedding model based on word2vec algorithm using CNN (the news network) news articles on disaster events.
3. Develop a novel BiLSTM network with self-attention and CRF layers, incorporating contextual word embedding, character embedding, and other NLP features.
4. Analyse effectiveness of contextual word embedding features in extracting disaster-specific named entities vs general embedding features.

Method used

Data prep:

1. Dataset - [HadeelAlzoubi/Disaster\_NER\_Attention\_BiLSTM\_CRF](https://github.com/HadeelAlzoubi/Disaster_NER_Attention_BiLSTM_CRF)
2. Noise removal – preprocessing to remove blank or unnecessary rows, identify and eliminate duplicated articles, removing extra spaces.
3. Word tokenisation of training sentences, including punctuation marks. Divide text into individual words or tokens – allows for further processing: parsing, POS tagging, etc.
4. Dataset annotation - gold standard corpus by manual annotation, and Automated extraction of named entities.
5. Disaster class description - dataset annotation produced 32 747 entries representing 13 different classes from the news article dataset.
6. Comprehensive word embedding features
   1. syntactic-level embedding: grammatical features like Parts of Speech (POS) tagging and Casing (capitalising words) information used for NER modelling to improve performance. NLTK python library used for these tasks.
   2. Word-level embedding : using word2vec
   3. Character-level embedding: using CNN
7. Bi-LSTM augmented with self-attention and CRF layers – the aim is to assign appropriate disaster class to each word token within a sentence (sequence labelling problem)
   1. LSTM – effectively address sequencing labelling challenges, particularly in NER tasks.
   2. Bi-LSTM can learn from sequence data by using hidden neurons to capture both past and future contextual information.
   3. Self-attention RNN can effectively capture long-range dependencies among words within a sentence.
   4. CRF layer – uses the Viterbi algorithm (probabilistic graphical model) for probabilistic inference of the entity labels. CRF layer outcome provides final tagging score for each entity.

Modelling BiLSTM-Attn-CRF:

Achieved through integration of the BiLSTM network, multi-head self-attention, and CRF layer.

Table 6 give detailed hyperparameters used.

Modelling time: Approximately 2 hours

Results

Table 7

Table 8

Table 9

Named Entity Extraction also known as entity recognition (NER) is a NLP technique that identifies and extracts named entities from any given text and classifies them into predefined categories.

Semantic annotation can be defined as the process of combining various pieces of information to concepts such as people, places, and things. It involves text identification & analysis, concept extraction, relationship extraction, and indexing. Named entity extraction is the part of semantic annotation and helps analyse the data.

BiLSTM is bidirectional long-short term memory. It can leverage information of past and future moment for better performance.

Conditional random field (CRF) layer is added to the model in order to add the constraint relationship between the labels. It ensures that the predicted labels are valid and finds an optimal label sequence.

Attention mechanism ([What are Attention Mechanisms in Deep Learning?](https://www.freecodecamp.org/news/what-are-attention-mechanisms-in-deep-learning/)) allows the model to selectively focus on specific areas of the input data when making predictions. It was originally employed in neural machine translation to assist the model in focusing on the most significant words or phrases in a sentence when translating it into another language.

Word2Vec ([Word2Vec: NLP with Contextual Understanding | Machine Learning Archive](https://mlarchive.com/natural-language-processing/word2vec-nlp-with-contextual-understanding/))

Related Work:

BiLSTM-CRF Chinese Named Entity Recognition Model with Attention Mechanism ([pdf](https://iopscience.iop.org/article/10.1088/1742-6596/1302/3/032056/pdf))

BiLSTM-Attention-CRF model for entity extraction in internet recruitment data ([main.pdf](https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050921X00063/1-s2.0-S1877050921005949/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEOb%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJGMEQCIAVmunQlxQbUtmuJMdUPwPBPzHhpTo%2Bg2bWH8GNb2bpLAiBa4Ew8XSaq5p%2FP0dRg6ZYOBFcCGkfF99U1fMuO%2Bcbk6iqzBQheEAUaDDA1OTAwMzU0Njg2NSIMGk9bZYWt9T54qIBiKpAFT59TK3riadSX8%2Bsibqq4KbexWvEUEg7EzhTq0%2BqRHQhv0Jli7fMZl3DQ89gzoXfNTRDx32Jw1uNzSVL7KNPxIhMFAJgZuVVeKbl0pvODMOjA6jpHCUvOjIIZwVHCiw4kem0lNDAwl7i%2FaQY4a9gjpqT5t53lc22A%2FIes1ayJ7QqGRLJhTMNLXW6BemlI956Hd3ovr3%2B9d4xDR964%2B55cuBe0rV0TIvnuJiJ%2BuWLWfjI6HDgPnW50MZ3N9fB2xlqChyIsZSYucpL4ETyy3vDQZBXWE6gIpoSaUE6G34IeGOr9GoBlDBhhdvs9qWkBHdXD8btZQG9NT2QwCZq7MIiUSVsHUtaRXjwPRWAVGwNvQDni2akV5HZx82GjaLuIeE6y0aXLSuy%2FfHUkrhgZePaKyhowzH2J5uLllUIQShwxZFcQVu5kbtLEmEuT8LDN619JdCtoRdKl9TiKwNDYnlyr2dGMjcmLkd3vbGk3%2FZdQ01XFjGLAWnuW2Qb%2F3qIjHEEx2iPf8Hw77BKtHpqxNiqRZYtx0ly7bi0u2rew9d10jdjZNKH1Pl%2FpBbeMbtQKshnMkMgSQxlCDB6i88rYkSTFkB4OG5ZZIcLQs4%2Fsjow6gVCExl8t9xaSrDNbB3A5iGQGnfTILqMtAkfyK2quupzLmbqLvcqOTXQ1SLh%2FL7t6MXdQyj4Lu15JKqU1ylP5dPdtNh3la%2FTrOruCRbrXDW2UReKVYHH4g50I3k%2BsIbwe%2BEs%2FRMvik8NquRJONvVzdId3gKqWOrbf1W1hOOZE%2FaRLdF6lEJ%2FsUCeU%2FfnsUnDYW8%2F6zFRBZoYI2vObZi1fyFB8xBgls623zW2BJUaH0HCjH2c9wE8n9Xu0nTsgomMpAUowmYKbxgY6sgFBsqW38ImmSnppvlqvWxZsKUnF5PX7S53f9oC5nQ%2BBNK3xvtEkkYt3xvc0HU0cItWYmDoLZn0HIqcUhA%2BGgqVKlWd7yDWuf1rO5cMbXDOnJQ9uTBAqaHKXqAJMGOoUJP2Grh%2Fq31d4Wkv4wLqz7nduCaYi2KsqTc4La8m4OeloSTDrHbi4WKYxNXtbIpDfMjhnVjbYCNTVmmDiWUIqvNuPIJ1SVA6CDt1W4ae%2B2frjEfBz&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20250914T141809Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYXAQY455I%2F20250914%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=ad69fa251eff889a670e461aa6f33ae9a4922aecd39225ba42e71fcc40b2acea&hash=800fd121dd039941ee5360ece392c454bb4ce1c92cef61d2e22cad1031c7fda8&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050921005949&tid=spdf-cd32fd75-ef0c-446f-9f27-cffce52be4c3&sid=3a6655cd7ce213495d-b2ab-55c621be53e2gxrqb&type=client&tsoh=d3d3LnNjaWVuY2VkaXJlY3QuY29t&rh=d3d3LnNjaWVuY2VkaXJlY3QuY29t&ua=1c005e56025058525751&rr=97f083128e1873a4&cc=za))

An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition ([attention-based BiLSTM-CRF approach to document-level chemical named entity recognition | Bioinformatics | Oxford Academic](https://academic.oup.com/bioinformatics/article/34/8/1381/4657076))

To answer the proposal question:

What method or algorithm are you proposing?

* We are proposing the BiLSTM combined with Self-Attention and CRF layers.

If there are existing implementations, will you use them and how?

* Yes, We will apply the same methods applied in the paper and then experiment with FastText instead of word2vec for word-level embeddings, and GPT instead of BERT for the contextualised word embeddings

How do you plan to improve or modify such implementations?

* They used word2vec for the word-level embeddings. We can try FastText – and extension of word2vec ([word\_embeddings](https://web.engr.oregonstate.edu/~huanlian/teaching/ML/2024fall/unit4/word_embeddings.html)).
* For contextualised word embeddings, we can try (OpenAI) GPT – transformer based contextualised word embeddings. BERT (used in the paper – table 11).

[overleaf.com/9578628825hrxywkpbfbhh#1c55c2](https://www.overleaf.com/9578628825hrxywkpbfbhh#1c55c2)