GNoM: Graph Neural Network Enhanced Language Models for Disaster Related Multilingual Text Classification

Samujjwal Ghosh  
Indian Institute of Technology  
Hyderabad  
India

Subhadeep Maji  
Amazon  
India

Maunendra Sankar Desarkar  
Indian Institute of Technology  
Hyderabad  
India

ABSTRACT

Online social media works as a source of various valuable and actionable information during disasters. These information might be available in multiple languages due to the nature of user generated content. An effective system to automatically identify and categorize these actionable information should be capable to handle multiple languages and under limited supervision. However, existing works mostly focus on English language only with the assumption that sufficient labeled data is available. To overcome these challenges, we propose a multilingual disaster related text classification system which is capable to work under monolingual, cross-lingual and multilingual lingual scenarios and under limited supervision. Our end-to-end trainable framework combines the versatility of graph neural networks, by applying over the corpus, with the power of transformer based large language models, over examples, with the help of cross-attention between the two. We evaluate our framework over total nine English, Non-English and monolingual datasets in monolingual, cross-lingual and multilingual lingual classification scenarios. Our framework outperforms state-of-the-art models in disaster domain and multilingual BERT baseline in terms of Weighted F1 score. We also show the generalizability of the proposed model under limited supervision.

CCS CONCEPTS

- Information systems → Information extraction: Multilingual and cross-lingual retrieval; Social networks.

KEYWORDS

Multilingual Learning, Natural Language Processing, Graph Neural Networks, Text Classification, Disaster Management

1 INTRODUCTION

People affected by disasters turn to online social network to seek help and report actionable information. Identification and categorization of these actionable information can help in planning rescue and relief operations effectively. However, these user-generated contents, such as tweets, are generally in languages native to the location of the disaster. On the other hand, majority of works in the literature focus mostly on English language [1, 3, 9, 12, 17, 21, 22] only. Understanding and processing texts in multiple languages is of paramount importance for effective disaster mitigation. A multilingual system capable of working with various languages will expand the applicability of such systems towards rescue and relief operations. Because of this, there is a strong need for a multilingual text classification framework which can identify and categorize useful and actionable information generated during disasters. Another challenge in building such an automated system is the lack of sufficient labeled data [1, 12] in disaster mitigation domain. Labeling examples during an ongoing disaster is costly and might not be feasible. This bottleneck becomes even more prevalent in multilingual scenario.

Keeping the above-mentioned constraints in mind, we propose a Graph Neural network based Multilingual text classification framework (GNoM) which can work efficiently under limited labeled data in monolingual, cross-lingual and multilingual linguistic settings. Our proposed approach enhances the power of transformer based large language models with the help of a Graph Neural Network (GNN) based formulation which enables the model to work both in multilingual setting and under limited labeled data by utilizing a word graph constructed from available textual corpus. GNoM has three main components, a Text Representer (TR), a Graph Featurizer (GF) and an Importance Estimator (IE), all of which are trained jointly in end-to-end manner. The purpose of the TR is to represent mono and multi lingual texts effectively which captures example level context for better class separability. On the other hand, the GF captures corpus level context from multilingual data which enables the framework to work in both mono and cross lingual settings. Both of these components are agnostic to any specific architecture and can be realized using any transformer and GNN architectures for TR and GF respectively. This flexibility allows for easier incorporation of new and powerful architectures in future. The IE combines the two components by estimating cross attention between them.

Due to recent success of GNNs in multiple domains [30, 31], we explore GNN-based GF to encode relationships among words in the dataset. We construct a word graph by connecting words present in the whole corpus (i.e. labeled and unlabeled data available from
which explore disaster-related text classification by applying Man-
work with multilingual data under limited supervision by capturing
There are many studies focusing on disaster-related tweet classifi-
2 RELATED WORKS
There are some notable studies which look into the multilingual
There are a few approaches which tried to incorporate a graph
Our proposed framework outperforms state-of-the-art (SOTA)
methods in disaster domain in mono, cross and multi lingual experi-
In summary, our contributions are as follows:
• We propose a framework for disaster related text classifi-
cation which works across monolingual, cross-lingual and
multilingual settings.
• The proposed framework is effective in utilizing easily avail-
able unlabeled data. At the same time, flexible with the ar-
chitectures that can be used.
• We show significant improvement in total 9 English Non-
English and multilingual disaster related tweet classification
dataset. Additionally we show that our framework is able to
generalize under limited supervision.

2 RELATED WORKS
There are many studies focusing on disaster-related tweet classifi-
cation in both binary [3, 17, 21, 22] and multi-label [12, 32] settings.
However, much of the literature is focused on monolingual cor-
pora, particularly, English language only [3, 17, 21, 22]. Whereas,
in a real-world scenario, user generated information may come in
any language. We first explore the literature of multilingual learn-
ing in disaster response followed by approaches which focus on
incorporating an additional graph component.

There are some notable studies which look into the multilingual
direction. We highlight a few works which explore disaster-related
tweet classification in multilingual setting. One of the comprehen-
sive works in this area was done by Raychowdhury et al. in [25]
which explore disaster-related text classification by applying Man-
ifold Mixup [28] on mBERT. They aggregated multiple disaster
datasets containing tweets in multiple languages into a single large
dataset and performed their experiments on that dataset. Krish-
nan et al. [14] explored classification of crisis related tweets using
attention realignment by introducing a language classifier in addi-
tion to task classifier. They use XLM-Roberta architecture as the
multilingual featurizer. However, their approach is dependent on
availability of parallel corpora. Piscitelli et al. explored applica-
tion of context-independent multilingual word embeddings called
MUSE [6] to perform tweet classification during emergencies in
their work [24]. Similarly, Lorini et al. used context-independent
multilingual embeddings for their flood recognition system based
on online social media called European Flood Awareness System
(EFAS) in [16]. However, context-independent word embedding
typically fails to capture relevant information [2]. Torres et al. ex-
plored crisis-related conversations in a cross-lingual setting [26].
Their study was limited to Spanish and English tweets only. The
work [20] by Musaev et al., filters tweets relevant to landslides
using Wikipedia articles as knowledge repository. One limitation
of this approach is that the model needs the same Wikipedia article
in multiple languages to learn multilingual embeddings. In [13],
Khare et al. classify relevancy of tweets from 30 crisis events in 3
languages (i.e. English, Spanish, and Italian). The main drawback
in their approach is that it is limited to the languages present in
the training data only and does not generalize to other languages.

There are a few approaches which tried to incorporate a graph
component to the model in the domain of disaster management.
These approaches majorly differ in how the graph is formed and
what kind of information they are trying to capture. However, none
of the approaches explored the multilingual perspective. Alam et
al. proposed end-to-end approach based on adversarial learning in
their work [1] (DAAT). They employ a GNN based component to
construct a document-level graph by calculating k-nearest neigh-
bours using Word2Vec [19] vectors. Li et al. use a Domain Recon-
struction Classification Network (DRCN) in their work [15] for
disaster related text classification. DRCN reconstructs the target
domain data with an autoencoder to minimize domain shift. Both
DAAT and DRCN approaches are designed for domain adaptation
setting in English language only. Zahera et al. explored the com-
bination of GAT [27] with mBERT [8] encoder in [52]. However,
their graph formulation is heterogeneous and based on the im-
plex assumption that sufficient training data is available which
might not be true in disaster scenarios. They incorporate class la-
beis in addition to named entities as additional nodes during graph
construction. Additionally, their work is focused only on English
language. Ghosh et al. proposes a 2-part global and local graph
neural network based technique called GLEN [12] to utilize global
token graph to learn domain agnostic features. They used token as
nodes in the graph with token cooccurrence as edges. This graph
construction is somewhat similar to ours. However, their formul-
ationale is limited to monolingual (English) setting only. We compare
GNoM with GLEN in monolingual experiments (Ref. Section 4).

3 PROPOSED FRAMEWORK
Our method comprises of three main components which attempt
to capture the corpus level (e.g dataset) and example level (e.g
tweet) context information representing same word in two differ-
ent embedding spaces and an attention mechanism to decide the
importance of those embeddings. The corpus-level information
serves as prior for word representation by aggregating over many
contexts in which a word appears while the example-level infor-
mation captures the context specific to an example for better class
separation. We claim that modelling the two contexts explicitly
helps in better generalization in a low resource text classification
setting such as disaster response, as compared to using example
level context (by employing, for example, transformer based mod-
els) alone. We establish this claim empirically in our experiments
Figure 1: Overview of our proposed GNoM framework. The “Text Representer” (TR) (§ 3.1) takes multilingual examples as input and generates both example level and word level representation. It primarily captures example level context which aids in effective separation of classes. On the other hand, tokenized multilingual unique words are used to construct the word graph which passes through the “Graph Featurizer” (GF) (§ 3.2) and serves as a prior over the words. The “Important Estimator” (§ 3.3) estimates the importance of word priors with respect to the input example by taking example representation from TR and node (word) vectors from GF. Finally, these attention weighted node vectors are aggregated with word vectors generated by the TR and passed to the classifier.

(§ 4). We combine the two embedding spaces with a novel end-to-end scaled dot product cross attention mechanism which learns to attend on corpus level context information given an example, where the downstream task is text classification. In addition, we enable multilinguality in our model by making it applicable to realistic disaster situations while most existing works on disaster response domain are monolingual [1, 9, 12] only.

In § 3.1, we discuss our method for obtaining example level contextual word embeddings with recent transformer based models (e.g. BERT [8] or XLM-R [5]). Our method is naturally multilingual by virtue of using a multilingual transformer model as the Text Representer. In § 3.2, we discuss our method for obtaining corpus level word embeddings with a Graph Featurizer (GF). We base the featurizer on Graph Convolution Network (GCN) on a word graph constructed from available labeled and unlabeled data. Our word graph is multilingual containing nodes (tokens) from multiple languages, and defines edges using embedding similarity between cross-language words from pretrained multilingual transformer models. In § 3.3, we discuss our method for combining the corpus level and example level word embeddings using a scaled dot product attention scheme, called Importance Estimator, which uses similarity between example embedding and individual GCN node embeddings to the compute attention scores.

We will refer to examples (e.g. tweets) as $x = (x_1, x_2, \ldots, x_n)$ where $x_i$ is the $i^{th}$ word. We assume access to labeled dataset for every classification task; $L = \{(x, y)\}$, where $y$ is binary, multi-class or multi-label target depending on the task. In addition, for some tasks, we also utilize unlabeled data $U = \{(x)\}$ when available (details on the collection process is deferred to § 4) which we use in the construction of the word graph while learning corpus level context embeddings (§ 3.2). Focus of the current work is multilingual text classification in disaster domain which is typically low resource, therefore our labeled datasets are small (on average ≈ 5K labeled examples).

3.1 Example Level Context Embedding

We employ a transformer based model as the Text Representer to learn example level contextual word embeddings. Transformer models owing to their self-attention structure learn word embeddings for a word depending on its similarity to all the other words in the example. The main objective of the text representer is to represent multilingual text effectively, and at the same time learn an embedding space which increases the separation among the classes. In the monolingual setting we use BERT and in multilingual setting we use mBERT to represent examples. In both settings, the pooled token embedding (i.e. [CLS] for BERT) is considered as the example embedding. The [CLS] token embedding based text representation has been widely used for downstream classification tasks [29]. We would like to emphasize that our overall model is not tied to BERT architecture and can be replaced with any transformer based text representation model architecture, for example XLM-RoBERTa. In context of an example $x$, we will denote the embedding of $x$ as $[CLS]_x$ and individual words $x_i$ as $h_{T|x}(x_i)$.

3.2 Corpus Level Context Embedding

We propose a graph neural network based Graph Featurizer to learn corpus level context based word embeddings. We define a word graph whose vertices are unique words $x_i$ from examples $x \in L \cup U$. In some tasks there is no unlabeled dataset (i.e $U = \emptyset$).
Typically, edges in word graphs are defined purely in terms of co-occurrence (within a window) of words from examples of an underlying corpus [12]. However, this fails to capture multilinguality because words from different languages seldom co-occur in a example, which in turn will result in a word graph with disconnected components. To address this limitation, we obtain embedding similarity from embedding layer of the transformer based large multilingual language model. We will refer to co-occurrence (within a window in examples) based similarity as matrix $C_{i,j}$ and embedding based similarity as matrix $E_{i,j}$. Matrices $C$ and $E$ are row-normalized and added (i.e $S = C + E$) to obtain the combined measure of similarity. The similarity values above a threshold ($S_{i,j} > \tau$) are used to define edges in the graph. The threshold ($\tau$) is a hyperparameter in our model. As a pre-processing step, very infrequent words (minimum corpus frequency of 3) and high frequency stopwords are not considered as nodes in the word graph. Co-occurrence similarity helps expand context over words within a language, whereas embedding similarity captures relationships across words from multiple languages. We initialize the word graph’s initial embeddings with the word embedding layer word representation from the Text Representer. This initialization technique serves two advantages: (a) as the multilingual TR’s are generally pretrained with large corpus we are able to bring this prior information in the formulation of the word graph and (b) enables both the Text Representer and the Graph Featurizer to have the same vocabulary. On the word graph, we apply a $k$-hop GCN to obtain graph based token (node) embeddings. GF expands the context information present in immediate neighborhood of nodes (i.e frequently co-occurring words/high embedding similarity) in the graph to its $k$-hop neighborhoods by aggregating information over multiple hops. A high value of $k$ expands the context to a larger neighborhood but risks oversmoothing [23], whereas smaller $k$ will limit the context expansion. We set $k = 2$ in all our experiments. We will denote the graph based embedding of word (node) $v$ as $h_G (v)$. 

### 3.3 Scaled Dot Product Cross Attention

We now turn to the question of combining the above two embedding spaces to improve generalization of the overall model on downstream text classification task. The graph based embedding of a word $x_i$ ($h_G (x_i)$) in an example $x$ is independent of rest of words in $x$ and is based on information propagation over its $k$-hop graph neighborhood. Therefore, the graph based embedding serves as a prior for word representation. We propose to combine this prior information with example level embedding $h_{TF} (x_i)$ using a scaled dot product attention which chooses to (ignore) attend to a prior word representation basis how (dis)similar the prior is to the pooled example embedding $[CLS]_x$. These attention scores works as an Importance Estimator in context to the example. In the standard scaled dot-product attention notation, the query $(Q)$ is $[CLS]_x$ and keys $(K)$ and values $(V)$ are both node vectors corresponding to words $x_{i=1...n}$ in $x$ formally,

$$Q = [CLS]_x$$

$$K = V = (h_G (x_i))_{i=1...n}$$

$$A(Q, K, V; W) = \text{Softmax} \left( \frac{(W_Q Q)(W_K K^T)}{\sqrt{d}} \right)$$

Here $W_Q$, $W_K$ are parameters of the dot-product attention. The attention scores $A_{i=1...n}$ form a distribution over values $V$. We combine the two embeddings by concatenating the attention multiplied prior embedding with example level context embedding per word as follows [$A_i \ast h_G (x_i); h_{TF} (x_i)$]. We refer to this mechanism as scaled dot product cross attention because embeddings from one subspace (example level context) serve as query for computing attention on another subspace (corpus level context). The attention layer is learned end-to-end with a classification task. In our experiments, we show an ablation study against the naive strategy of simply concatenating the two embeddings and establishing the effectiveness of our scheme.

### 4 EXPERIMENTAL SETUP

Our goal is to build a disaster-related text classification system which works across monolingual, cross-lingual and multilingual linguistic settings. Particularly, we aim to answer the following research questions via our experiments:

- How does the performance of GNoM compare to state-of-the-art mono/cross/multi lingual models in disaster-related text (e.g., tweets) classification domain?
- Is GNoM capable of working when the amount of training data available is very limited?
- How does each component of GNoM impacts classification performance (i.e., Ablation Study)?

In disaster domain, it is imperative that the system works under limited supervision. To verify the effectiveness of GNoM in such scenarios, similar to [12], we reduce the training data to 50%, 25% and 10% of the original training set without changing the validation and test sets.

### 4.1 Datasets

We performed experiments on total 9 datasets, out of which 5 are in English, 3 are in Non-English (e.g. Spanish, Italian, etc.) language and 1 contains multilingual data. All the datasets are publicly available containing disaster related tweets. To perform experiments in both in-domain and cross-domain settings, we pair up datasets with same class labels.

#### 4.1.1 English Datasets

For experiments with English language, we used publicly available two binary datasets and two multi-label datasets of tweets generated during disasters. The binary datasets ’2013 Queensland Flood’ (QFL) and ’2015 Nepal Earthquake’ (NEQ) [1] are labeled with relevance of tweets as classes. We present the class specific details of these two datasets in Table 1. The unlabeled part of both the datasets were downloaded using Twitter’s public API. We obtained a total of 49,223 and 15,464 tweets from NEQ and QFL datasets respectively. We used the train, dev and test split provided by the authors as train, validation and test set.

We also experiment with two multi-label datasets, namely, ’Forum for Information Retrieval Evaluation 2016’ (FIRE16) [10] and ’Social Media for Emergency Relief and Preparedness’ (SMERP17) [11], containing tweets collected during Nepal 2015 earthquake and 2016 Italy earthquake respectively. Tweets in these datasets are labeled with multi-label annotation where each example may belong to
| Dataset       | Language | 1   | 0   | Train | Val | Test  |
|--------------|----------|-----|-----|-------|-----|-------|
| QFL          | English  | 5414| 4619| 6019  | 1003| 3011  |
| NEQ          | English  | 5527| 6141| 7000  | 1166| 3502  |
| ChileEQT1    | Spanish  | 928 | 1259| 1312  | 88  | 787   |
| SoSItalyT4   | Italian  | 4739| 903 | 3385  | 226 | 2031  |
| EcuadorS     | Spanish  | 2322| 1846| 2501  | 167 | 1500  |
| EcuadorE     | English  | 2249| 1946| 2515  | 180 | 1500  |

Table 1: Details of QFL, NEQ, ChileEQT1, SoSItalyT4 and Ecuador datasets. 1 and 0 indicate relevant and irrelevant classes.

| FIRE16 | SMERP17 |
|--------|---------|
| Resources Available | Resources Available |
| Medical Resources Available | Medical Resources Available |
| Resources Required | Resources Required |
| Medical Resources Required | Medical Resources Required |
| Resources Specific Locations | Resources Specific Locations |
| Infrastructure Damage & Restoration | Infrastructure Damage & Restoration |
| Activities NGOs / Government | Activities NGOs / Government |

Table 2: Class mapping from FIRE16 to SMERP17. Class 5 of FIRE16 was ignored.

4.1.2 Non-English Datasets. We use four datasets collected from different sources. Ray Chowdhury et al. curated a large multilingual dataset of 134420 tweets [25], annotated with five classes in multiclass setting, related to multiple disasters. They provided the train, validation and test split of the data in the form of tweet ids, due to Twitter’s policy on data sharing, which we tried to download. However, we could only download 46667, 4226 and 5928 tweets from the train, validation and test sets respectively. Table 3 contains details about the dataset, named ‘MixUp’. The dataset ‘Ecuador’ was collected by Torres et al. in [26]. The dataset contains tweets from four different natural disasters in Italy between 2009 and 2014. The tweets in the dataset are annotated with “damage”, “no damage”, or “not relevant”. However, similar to [13], we convert the annotations to binary relevance with “damage” and “no damage” both indicating relevance. We pair up ChileEQT1 and SOSItalyT4 for our cross lingual experiments.

4.2 Baselines

Our framework enhances the transformer based Text Representer by incorporating the Graph Featurizer and Importance Estimator. To verify the effectiveness of these components, we define the vanilla Text Representer as the baseline for our experiments. We also compare with SOTA methods from disaster related text classification domain. A GNN based SOTA method was applied on QFL, NEQ, FIRE16 and SMERP17 in paper [12] (GLEN) by Ghosh et al., we compare with this method in our experiments over those datasets. A few other SOTA methods presented in [1] (DAAT) by Alam et al. and [15] (DRCN) by Li et al. also experimented with QFL and NEQ datasets, we compare against them. Torres et al. in [26] applied their approach (CLP) in both mono and cross lingual setting for Ecuador dataset. We compare with CLP in addition to vanilla mBERT for experiments over Ecuador dataset.

Recall that GNoM is flexible with the transformer architecture in the TR component. We experiment with three realizations of TR using BERT (GNoMB) for English datasets, and using mBERT (GNoMM) and XLM-RoBERTa (GNoMX) architecture for Non-English or multilingual datasets. BERT-base-uncased, BERT-Base-Multilingual-Cased and XLM-RoBERTa-Base variant are used for experiments with GNoMB, GNoMM and GNoMX respectively. We initialize the word graph node vectors with the word embeddings of the corresponding TR.

For our ablation study, we report results on the following ablations:

- Only TR (Without GF and IE): This setting corresponds to training the TR only i.e. BERT for English and mBERT for other language datasets. Only word vectors are passed to the classifier without the node vectors in the Figure 1.
- TR+GF (Without IE): In this ablation, we estimate the need of Importance Estimator in our framework. Both TR and GF are trained but without the IE, i.e. vectors from TR and GF are simply concatenated directly without reweighting GF vectors.
- TR+GF-e+IE (Without embedding similarity edges): We construct the edges in the word graph using only cooccurrence for monolingual and both cooccurrence and embedding similarity for cross and multilingual settings. However, this ablation verifies the situation when only cooccurrence edges are used in cross and multi lingual settings.
- GNoM Framework (GNoM): This setting represents our framework GNoM. We argue that GNoM is flexible with various transformer architectures. We show two realisations of TR using (m)BERT [8] and XLM-RoBERTa [5] architectures for Non-English experiments.
Table 3: Details of FIRE16 (left), SMERP17 (middle) and MixUp (right) datasets. FIRE16 and SMERP17 contains tweets in English with multi-label annotation whereas MixUp contain tweets in multiple languages with only multi-class annotation. We provide example count for multi-label datasets as it may differ from the total number of annotations.

| Class | Train | Val | Test |
|-------|-------|-----|------|
| 1     | 498   | 55  | 237  |
| 2     | 217   | 24  | 104  |
| 3     | 367   | 41  | 175  |
| 4     | 302   | 34  | 144  |

Example Count 957 106 459

| Class | Train | Val | Test |
|-------|-------|-----|------|
| 1     | 184   | 22  | 76   |
| 2     | 105   | 15  | 46   |
| 3     | 774   | 141 | 393  |
| 4     | 212   | 25  | 80   |

Example Count 1159 189 548

| Class | Train | Val | Test |
|-------|-------|-----|------|
| 1     | 5933  | 544 | 861  |
| 2     | 3409  | 304 | 509  |
| 3     | 1328  | 144 | 240  |
| 4     | 18132 | 1635| 2197 |
| 5     | 17865 | 1599| 2121 |

Total 46667 4226 5928

4.3 Training Configuration

A 2-layer bi-directional LSTM (BiLSTM) network with a fully connected layer head is used as the classifier. Our framework was trained jointly with the classifier in an end-to-end manner. We update all the layers during training for both GNoM and the baselines. We ran each experiment 3 times and report the average of those runs. Weighted F₁ score is used as the evaluation metric as it is a commonly used metric in the literature.

A few datasets have unlabeled data available in addition to labeled data. GNoM is capable to incorporate such extra data during the construction of the word graph. For SoStitalyT4, ChileEQT1 and Ecuador datasets, we treat the target domain train data as the unlabeled data during cross domain experiments. Note that target domain class information is not used in any of our experiments. For in-domain monolingual experiments for SoStitalyT4, ChileEQT1, Ecuador and MixUp dataset, no unlabeled data was used. For monolingual experiments, we construct the word graph using only cooccurrence, similar to [12], as there is no need to model inter-language relations. However, for cross and monolingual experiments we use both cooccurrence and embedding similarity to construct the edges.

We tuned our hyperparameters such as embedding similarity threshold (Ref. 3.2, τ), learning rate and the number of epochs using the validation data. We searched the value of embedding similarity threshold based on performance on validation data and set the value to 0.5 across all experiments. We searched learning rate values with $10^{-i}$ where $i \in \{4, 5, 6\}; i = 5$ found to be most suitable in majority of the training scenarios.

5 RESULTS

GNoM utilizes corpus as well as example level context to capture relations across languages. We validate the effectiveness of GNoM through multiple experiments in mono, cross and multi-lingual settings.

5.1 Monolingual Classification

In this setting, we use data from a single language for both training and evaluation. We present our findings in Tables 4, 5, 6 and 7 for QFL, NEQ, FIRE16, SMERP17, SoStitalyT4, ChileEQT1 and Ecuador datasets respectively. For QFL and NEQ datasets (Table 4), we compare with GLEN, DRCN and DAAT from disaster related text classification literature and with BERT baseline. We perform experiments in both in and cross domain monolingual setting. GNoM is able to outperform GLEN (best performing among SOTA) by average 4% in F₁ score. We compare with GLEN and BERT for multi-label monolingual datasets FIRE16 and SMERP17 in Table 5. Our framework boosts F₁ significantly by as much as 6.42% on average.

In Non-English SoStitalyT4, ChileEQT1 and Ecuador datasets, bottom two rows signify monolingual setting in Tables 6 and 7. No unlabeled extra data was used for these experiments. We compare with BERT baseline for SoStitalyT4, ChileEQT1 datasets. In addition, we compare with CLP for Ecuador dataset. Our approach is able to outperform BERT baseline in all 4 scenarios. Although the performance improvement is marginal over GLEN but we achieve a significant improvement over BERT.

Table 4: Weighted F₁ scores over NEQ and QFL datasets. GNoM outperforms other SOTA methods in both cross and in domain setting.

| Source | Target | DAAT | DRCN | BERT | GLEN | GNoMB |
|--------|--------|------|------|------|------|-------|
| NEQ    | QFL    | 65.90| 81.18| 80.72| 83.42| 86.68 |
| QFL    | NEQ    | 59.50| 68.38| 67.22| 71.61| 71.73 |
| NEQ    | NEQ    | 65.11| -    | 76.39| 77.76| 78.95 |
| QFL    | QFL    | 93.54| -    | 96.24| 96.77| 96.26 |

Table 5: Scores (Weighted F₁) of FIRE16 and SMERP17 datasets.
5.3 Multilingual Classification

Multilingual classification setting refers to the scenario when both train and test set contains data from a mixture of multiple languages. This setting is practical in disaster scenarios as user generated social network data may be available in multiple languages. We summarize our result for multilingual classification in Table 8 for the MixUp dataset. We only use the train set and do not use any extra data for construction of the word graph in this setting. However, we observe that explicit modelling of the inter-language relation (by constructing the inter-language word graph with initial embedding similarity scores as edges) help improve performance by 1.15 (GNoMX) and 1.29 (GNoMM). Unfortunately, our result can not directly be compared with [25] as they use a larger set of data which we could not collect as it was not available (Ref. 4.1.2).

| Source          | Target        | mBERT | CLP  | GNoMX | GNoMM |
|-----------------|---------------|-------|------|-------|-------|
| EcuadorE        | EcuadorS      | 77.93 | 77.49| 81.89 | 81.54 |
| EcuadorS        | EcuadorE      | 90.45 | 85.88| 91.72 | 91.45 |
| EcuadorE        | EcuadorE      | 94.23 | 94.05| 94.30 | 94.50 |
| EcuadorS        | EcuadorS      | 85.18 | 85.77| 86.79 | 86.86 |

Table 8: Weighted F₁ scores for MixUp (multilingual) dataset.

5.4 Limited Supervision

Due to lack of labeled data in disaster domain is a common phenomenon, an effective classification system should work when very limited amount of labeled data is available. We want to verify if GNOM is capable to capture appropriate context from the unlabeled corpus so that it perform considerably well under limited supervision. To verify this, we design an experiment to limit the availability of training data to 50%, 25% and 10% of the original size, similar to [12].

Tables 9 and 10 summarises our findings under limited supervision over the English and Non-English datasets. We utilized the unlabeled data available with QFL, NEQ, FIRE16 and SMERP17 datasets to construct the word graph (Ref. 4.1.1). For SoSItalyT4, ChileEQ1T1 and Ecuador datasets, we use the target domain data as the unlabeled data during crosslingual experiments. Note that target domain class information is not used in any of our experiments. For monolingual experiment, we do not use any unlabeled data.

GNOM outperforms both baseline BERT and SOTA method GLEN for English datasets by a large margin, Table 9. A vanilla BERT model overfits the small amount of training data, however, our formulation enables the model to capture larger context and overcome the overfitting problem. GLEN relies on word pair-wise contextual attention using a GAT [27] to capture class separability, whereas our formulation uses self-attention across all the words. Additionally, our IE (cross attention) component aids in filtering noisy priors out. These additions result in average absolute gain of 2.67%, 3.73% and 5.85% with 50%, 25% and 10% of training data respectively.

For Non-English datasets, we compare with mBERT baseline only, as GLEN does not have multilingual capability. We experiment with GNoMM, additionally, we also report results on GNoMX. We presents our experimental results over Non-English datasets in Table 10. Our framework GNoMM consistently outperforms vanilla mBERT baseline across all training data proportions with average
We proposed an multilingual disaster related text classification framework, called GNoM, which works across different languages. Explicit capturing of the corpus-level and example-level contexts enable GNoM to work under monolingual, cross-lingual and multi-lingual settings. Each component of GNoM plays a crucial role to make an effective classification system at the same time being flexible with the choice of architectures. The framework is also able to work under very limited supervision significantly outperforming baselines. Our experiments over 5 English, 3 Non-English and 1 multilingual datasets with binary, multi-class and multi-class multi-label settings show broader applicability of our framework in disaster related text classification. We argue that any GNN based graph featurizer can be applied in our framework. We plan to experiment and validate this in future. We also plan to explore the possibility of applying our framework in other short-text classification domains.

REFERENCES

[1] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain Adaptation with Adversarial Training and Graph Embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1087–1097.

[2] Ainiyra Barua, S. Thara, B. Premjith, and K. P. Somam. 2021. Analysis of Contextual and Non-contextual Word Embedding Models for Hindi NER with Web Application for Data Collection. In Advanced Computing, Deepak Garg, Kit Wong, Jagannathan Sarangapani, and Sunnet Kumar Gupta (Eds.). Springer Singapore, Singapore, 183–202.

[3] Cornelia Caragea, Adrian Silvescu, and Andrea H. Tapia. 2016. Identifying informative messages in disaster events using Convolutional Neural Networks. In ISCRAM 2016 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management (Proceedings of the International ISCRAM Conference). Pedro Antunes, Victor Amadeo Banels Silvey, Joao Porto de Albuquerue, Kathleen Ann Moore, and Andrea H. Tapia (Eds.). Information Systems for Crisis Response and Management, ISCRAM.

[4] Alfredo Cobo, Denis Farra, and Jaime Navón. 2015. Identifying Relevant Messages in a Twitter-Based Citizen Channel for Natural Disaster Situations. In Proceedings of the 24th International Conference on World Wide Web (Florence, Italy) (WWW ‘15 Companion). Association for Computing Machinery, New York, NY, USA, 1189–1194. https://doi.org/10.1145/2740908.2741719

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint arXiv:1912.02116 (2019).

[6] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. arXiv preprint arXiv:1710.04087 (2017).

[7] Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell’Orletta. 2015. A Linguistically-Driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages. In Proceedings of the 24th International Conference on World Wide Web (Florence, Italy) (WWW ’15 Companion). Association for Computing Machinery, New York, NY, USA, 1195–1200. https://doi.org/10.1145/2740908.2741722

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://arxiv.org/abs/1810.04805

[9] Samujjal Ghosh and Maunendra Sankar Desarkar. 2020. Semi-Supervised Granular Classification Framework for Resource Constrained Short-texts: Towards Retrieving Situational Information During Disaster Events. In 12th ACM Conference on Web Science. 29–38.

[10] Saptarshi Ghosh and Kripabandhu Ghosh. 2016. Overview of the FIRE 2016 Microblog track: Information Extraction from Microblogs Posted during Disasters. In Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-16, 2016 (CEUR Workshop Proceedings, Vol. 1737). Prasenjit Majumder, Mandar Mitra, Parth Mehta, Jainisha Sankhavara, and Kripabandhu Ghosh (Eds.). CEUR-WS.org, 56–61. http://ceur-ws.org/Vol-1737/37-2-1.pdf

[11] Saptarshi Ghosh, Kripabandhu Ghosh, Tammy Charkraborty, Debasish Ganguly, Gareth Jones, and Marie-Francine Moens. 2017. First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP). ADVANCES IN INFORMATION RETRIEVAL, ECIR 2017 10193 (2017), 779–783.

[12] Saptarshi Ghosh, Subhadeep Maji, and Maunendra Sankar Desarker. 2021. Unsupervised Domain Adaptation with Global and Local Graph Neural Networks in Limited Labeled Data Scenario: Application to Disaster Management. arXiv:2104.01436 [cs.CL]

[13] Prashant Khare, Grégory Burel, Diana Maynard, and Harith Alani. 2018. Cross-Lingual Classification of Crisis Data. In The Semantic Web – ISWC 2018. Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl (Eds.). Springer International Publishing, Cham, 617–633.
Table 9: Weighted $F_1$ scores for NEQ, QFL, FIRE16, SMERP17, datasets under 50%, 25% and 10% of train set. We compare with baseline BERT and SOTA method GLEN. GNoM is able to outperform both.

| Source | Target | 50% | 25% | 10% |
|--------|--------|-----|-----|-----|
| NEQ    | QFL    | 84.99 | 83.86 | 86.62 |
| QFL    | NEQ    | 65.78 | 70.33 | 70.45 |
| NEQ    | NEQ    | 74.07 | 75.82 | 77.77 |
| QFL    | QFL    | 94.73 | 96.23 | 96.54 |
| FIRE16 | SMERP17| 72.64 | 76.77 | 78.01 |
| SMERP17| FIRE16 | 37.73 | 47.33 | 56.82 |
| FIRE16 | FIRE16 | 70.44 | 78.68 | 81.19 |
| SMERP17| SMERP17| 91.24 | 93.49 | 95.68 |

Average Gain (%) 2.67 3.73 5.85

Table 10: Weighted $F_1$ scores for ChileEQT1, SoSItalyT4, Ecuador and MixUp datasets under 50%, 25% and 10% of train set. We compare with multilingual BERT baseline. We use two realizations for our TR component using XLM-RoBERTa (GNoMX) and multilingual BERT (GNoMM).

| Source | Target | 50% | 25% | 10% |
|--------|--------|-----|-----|-----|
| EcuadorE | EcuadorS | 77.29 | 81.47 | 81.22 |
| EcuadorS | EcuadorE | 88.86 | 91.41 | 91.30 |
| EcuadorE | EcuadorE | 93.84 | 94.23 | 94.24 |
| EcuadorS | EcuadorS | 83.12 | 84.36 | 84.56 |
| ChileEQT1 | SoSItalyT4 | 42.50 | 51.61 | 48.32 |
| SoSItalyT4 | ChileEQT1 | 53.38 | 66.26 | 62.52 |
| ChileEQT1 | ChileEQT1 | 83.81 | 85.49 | 85.59 |
| SoSItalyT4 | SoSItalyT4 | 84.26 | 85.15 | 85.03 |
| MixUp | MixUp | 69.49 | 71.12 | 71.15 |

Average Gain (%) 3.11 3.08 4.62

Figure 2: UMAP projections of tokens from different languages (color-coded) before and after training. Figures (a) and (b) show for SoSItalyT4-ChileEQT1 datasets. Similarly, Fig. (c) and (d) show the plots for the EcuadorE-EcuadorS datasets.

[14] Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. 2020. Attention Realignment and Pseudo-Labeling for Interpretable Cross-Lingual Classification of Crisis Tweets. In KML@KDD.

[15] Xukun Li and Doina Caragea. 2020. Domain Adaptation with Reconstruction for Disaster Tweet Classification. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1561–1564.

[16] V. Lorini, C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salamon. 2019. Integrating Social Media into a Pan-European Flood Awareness System: A Multilingual Approach. arXiv:1904.10876 [cs.IR]

[17] Reza Mazloom, Hongming Li, Doina Caragea, Muhammad Imran, and Cornelia Caragea. 2018. Classification of Twitter Disaster Data Using a Hybrid Feature-Instance Adaptation Approach. In ISCRAM.
Tweets during Disasters. In ISCRAM.

[22] Tien Dat Nguyen, Kamila Al-Mannai, Shafiq R. Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks. In IJWSSN.

[23] Kenta Oono and Taiji Suzuki. 2020. Graph Neural Networks Exponentially Lose Expressive Power for Node Classification. In International Conference on Learning Representations. https://openreview.net/forum?id=S1ldO2EFr

[24] Sara Fiscelli, Edoardo Arnaldo, and Claudio Ross. 2021. Multilingual Text Classification from Twitter during Emergencies. In 2021 IEEE International Conference on Consumer Electronics (ICCE). 1–6. https://doi.org/10.1109/ICCE50685.2021.9427581

[25] Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020. Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Association for Computational Linguistics, Online, 292–298. https://doi.org/10.18653/v1/2020.acl-sw.39

[26] Johnny Torres, Carmen Vaca. 2019. Cross-Lingual Perspectives about Crisis-Related Conversations on Twitter. In Companion of The 2019 World Wide Web Conference (San Francisco, USA) / WWW ’19. Association for Computing Machinery, New York, NY, USA, 255–261. https://doi.org/10.1145/3308560.3316799

[27] Petar Velickovic, Guillem Cusculli, Aranzka Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In International Conference on Learning Representations (ICLR ’18).

[28] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold Mixup: Better Representations by Interpolating Hidden States. In Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97), Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 6438–6447. http://proceedings.mlr.press/v97/verma19a.html

[29] Congcong Wang, Paul Nulty, and David Lillis. 2021. Transformer-based Multi-task Learning for Disaster Tweet Categorisation. arXiv preprint arXiv:2110.08010 (2021).

[30] Yongji Wu, Defu Lian, Yiheng Xu, Le Wu, and Enhong Chen. 2020. Graph Convolutional Networks with Markov Random Field Reasoning for Social Spammer Detection. Proceedings of the AAAI Conference on Artificial Intelligence 34, 01 (Apr. 2020), 1054–1061. https://doi.org/10.1609/aaai.v34i01.5455

[31] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 7370–7377.

[32] Hamada M. Zahera, Richa Jalota, Mohamed Ahmed Sherif, and Axel-Cyrille Ngonga Ngomo. 2021. I-AID: Identifying Actionable Information From Related Conversations on Twitter. In Proceedings of ISCRAM.

Table 11: Ablation with weighted F1 scores over all the datasets. Both components (i.e. GF and IE) contribute to the improvement of performance. We experiment with TR+GF-e+IE on cross and multi lingual settings only (Ref. 4.3). We use GNoMB for English and GNoMM for Non-English datasets.

| Source Target | TR | TR+GF | TR+GF-e+IE | GNoMB(B|M) |
|---------------|----|-------|------------|-----------|
| NEQ QFL       | 80.72 | 84.87 | -          | 86.68     |
| QFL NEQ       | 67.22 | 68.42 | -          | 71.13     |
| FIRE16 SMERP17| 76.21 | 79.37 | -          | 79.49     |
| SMERP17 FIRE16| 55.52 | 58.90 | -          | 62.57     |
| EcuadorE EcuadorS | 77.93 | 81.37 | 81.46 | 81.54     |
| EcuadorS EcuadorE | 90.45 | 91.25 | 91.34 | 91.45     |
| ChileEQTI SoStalyTy | 43.17 | 46.97 | 47.38 | 49.14     |
| SoStalyTy ChileEQTI | 54.46 | 56.27 | 60.36 | 63.20     |
| MixUp MixUp   | 70.16 | 70.47 | 70.68 | 71.45     |
| Average       | 68.08 | 70.21 | 70.24 | 72.96     |

Table 12: Word and its five neighbors based on similarity of transformer word embeddings.

| Word Neighbors                                                                                     | Word Neighbors                                                                                     | Word Neighbors                                                                                     |
|---------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| report, Report, report, and, port                                                               | information, información, info, datos, informacion                                               | emergencia, vivienda, desastre, emergency, situación                                              |
| everyone, Everyone, everything, anyone, people                                                   | emergencia, vivienda, desastre, emergency, situación                                              | emergencia, vivienda, desastre, emergency, situación                                              |
| información, información, info, datos, informacion                                               | emergencia, vivienda, desastre, emergency, situación                                              | emergencia, vivienda, desastre, emergency, situación                                              |
| make it’s time we form a group of digital disaster responders who mobilize                         | varro entre los es que genera, los más, los más sobre                                 | vive en el terremoto en Ecuador                                                                  |
| Deadly 7 Eight earthquake rocks Ecuador                                                         | terremoto, que, la, han, de varios, cosas, personas                                              | Perdi col treno a Piacenza, causo, terremoto                                                    |
| ChiliEQTI SoStalyTy                                                                              | #luna di #un #a di #s #o #e #i #a #et #e #o #o #SA # #R | es # #de #res #mu

Figure 3: Visualizations of cross attention scores over a few examples estimated by the IE component.