wahid_2022_topic2labels_a_framework_to_annotate_and_classify_the_social_media_data_through_lda_topics_and_deep_learning_models_for_crisis_response

Year

2022

Author(s)

Junaid Abdul Wahid and Lei Shi and Yufei Gao and Bei Yang and Lin Wei and Yongcai Tao and Shabir Hussain and Muhammad Ayoub and Imam Yagoub

Title

Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response

Venue

Expert Systems With Applications

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

\

Label generation

”topic 3 has words that depict information about the emotional support and help seeking during the pandemic, people talk about health care workers, giving emotional support to them, and talking about help seeking of different items needed during the pandemic. Similarly when you look at terms used in topic 0, it shows that this topic is about announcements of coronavirus, notification about the measures taken by the government and public health authorities, talks on medical equipment reserves and hospital conditions etc.

Similarly information extracted through topics from the disaster dataset is also relatable to different types of categories. As shown in Table 3, topic 3 contains information about death tolls, weather updates, about specific area which is Queensland and about specific type of crisis which is flood. Topic 1 shows the public reaction which they showed in their anger, attitudes, and feelings in a negative way towards government policies in crisis situations, therefore it categorizes as ‘‘People criticism in crisis situations’’. Information about injured, people death counts, updated news about the crisis is all contained in topic 0, therefore we categorized it as News about deaths and injured people.”

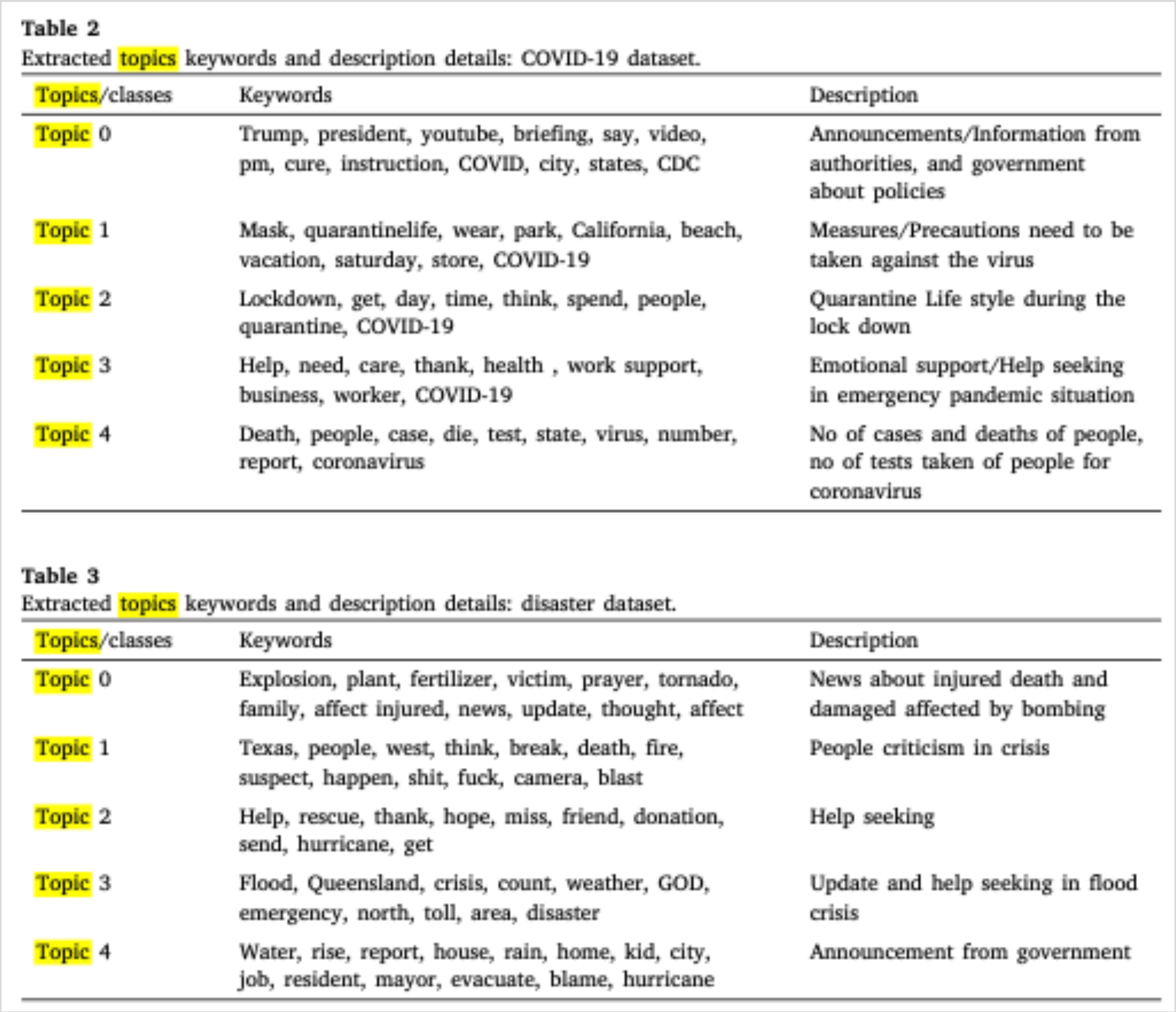

Table 2
Extracted topics keywords and description details: COVID-19 dataset.

Topics/classes	Keywords	Description
Topic 0	Trump, president, youtube, briefing, say, video, pm, cure, instruction, COVID, city, states, CDC	Announcements/Information from authorities, and government about policies
Topic 1	Mask, quarantinelife, wear, park, California, beach, vacation, saturday, store, COVID-19	Measures/Precautions need to be taken against the virus
Topic 2	Lockdown, get, day, time, think, spend, people, quarantine, COVID-19	Quarantine Life style during the lock down
Topic 3	Help, need, care, thank, health , work support, business, worker, COVID-19	Emotional support/Help seeking in emergency pandemic situation
Topic 4	Death, people, case, die, test, state, virus, number, report, coronavirus	No of cases and deaths of people, no of tests taken of people for coronavirus

Table 3
Extracted topics keywords and description details: disaster dataset.

Topics/classes	Keywords	Description
Topic 0	Explosion, plant, fertilizer, victim, prayer, tornado, family, affect injured, news, update, thought, affect	News about injured death and damaged affected by bombing
Topic 1	Texas, people, west, think, break, death, fire, suspect, happen, shit, fuck, camera, blast	People criticism in crisis
Topic 2	Help, rescue, thank, hope, miss, friend, donation, send, hurricane, get	Help seeking
Topic 3	Flood, Queensland, crisis, count, weather, GOD, emergency, north, toll, area, disaster	Update and help seeking in flood crisis
Topic 4	Water, rise, report, house, rain, home, kid, city, job, resident, mayor, evacuate, blame, hurricane	Announcement from government

Motivation

\

Topic modeling

LDA

Topic modeling parameters

Nr of topics: {5, 10, 15, 20}

Nr. of topics

10 (5 per dataset)

Label

Manually assigned full sentence descriptors

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper:
Dataset: Social media (Twitter)

Problem statement

In this study, we propose Topic2labels (T2L) framework which provides an automated way of labeling the data through LDA (latent dirichlet allocation) topic modeling approach and utilize Bert (the bidirectional encoder representation from transformer) embeddings for construction of feature vector to be employed to classify the data contextually.
Our framework consists of three layers. In the first layer, we adopt LDA to generate the topics from the data, and develop a new algorithm to rank the topics, and map the highest ranked dominant topic into label to annotate the data.
In the second layer, we transform the labeled text into feature representation through Bert embeddings and in the third layer we leveraged deep learning models as classifiers to classify the textual data into multiple categories. Experimental results on crisis-related datasets show that our framework performs better in terms of classification performance and yields improvement as compared to other baseline approaches.

Corpus

Dataset 1

Origin: Twitter

Nr. of documents: 10M

Details:

- tweets about COVID-19 pandemic

Dataset 2

Origin: Twitter

Nr. of documents: 70.000

Details:

- disaster related tweets

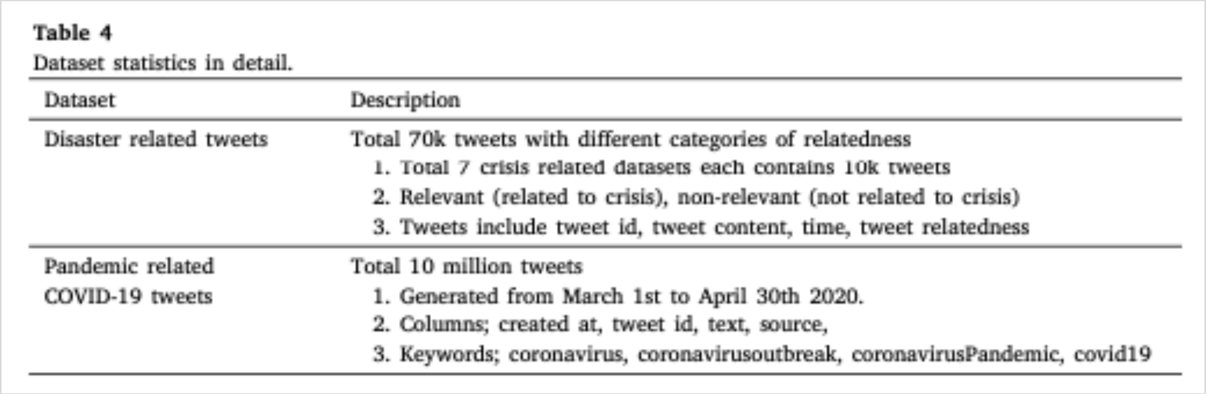

Table 4
Dataset statistics in detail.

Dataset	Description
Disaster related tweets	Total 70k tweets with different categories of relatedness 1. Total 7 crisis related datasets each contains 10k tweets 2. Relevant (related to crisis), non-relevant (not related to crisis) 3. Tweets include tweet id, tweet content, time, tweet relatedness
Pandemic related COVID-19 tweets	Total 10 million tweets 1. Generated from March 1st to April 30th 2020. 2. Columns; created at, tweet id, text, source, 3. Keywords; coronavirus, coronavirusoutbreak, coronavirusPandemic, covid19

Document

A tweet content together with:

- created at field, id and source (D1)
- id, time, tweet relatedness (D2)

Pre-processing

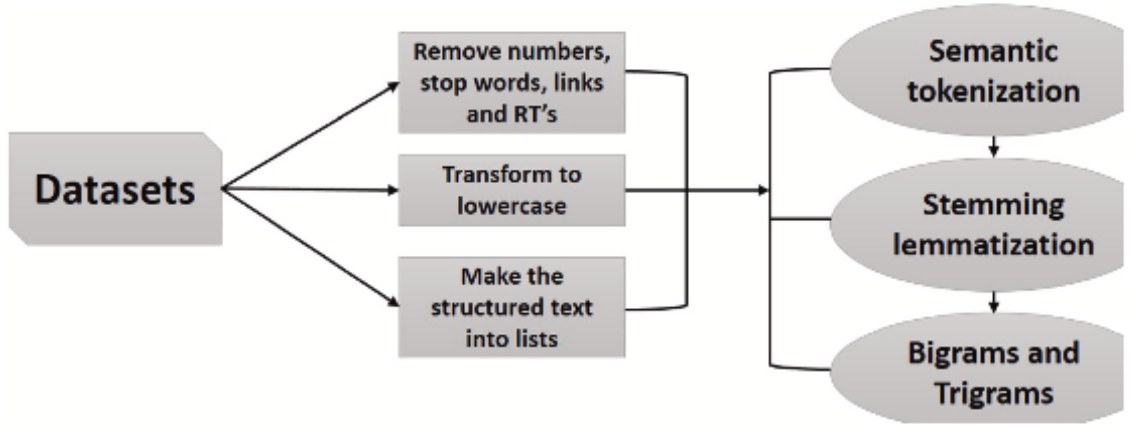

Fig. 3. Pre-processing steps involved in data pre-processing.

```
@article{wahid_2022_topic2labels_a_framework_to_annotate_and_classify_the_socia
l_media_data_through_lda_topics_and_deep_learning_models_for_crisis_response,
  abstract = {The abundant use of social media impacts every aspect of life,
including crisis management. Disaster management needs real-time data to be
used in machine learning and deep learning models to aid their decision making.
Mostly the data that is newly generated from social media is unstructured and
unlabeled. Current text classification models based on supervised deep learning
models heavily rely on human-labeled data that very small size and imbalanced
in the context of disasters, ultimately affecting the generalization of models.
In this study, we propose Topic2labels (T2L) framework which provides an
automated way of labeling the data through LDA (latent dirichlet allocation)
topic modeling approach and utilize Bert (the bidirectional encoder
representation from transformer) embeddings for construction of feature vector
to be employed to classify the data contextually. Our framework consists of
three layers. In the first layer, we adopt LDA to generate the topics from the
data, and develop a new algorithm to rank the topics, and map the highest
ranked dominant topic into label to annotate the data. In the second layer, we
transform the labeled text into feature representation through Bert embeddings
and in the third layer we leveraged deep learning models as classifiers to
classify the textual data into multiple categories. Experimental results on
crisis-related datasets show that our framework performs better in terms of
classification performance and yields improvement as compared to other baseline
approaches.},
  author = {Junaid Abdul Wahid and Lei Shi and Yufei Gao and Bei Yang and Lin
Wei and Yongcai Tao and Shabir Hussain and Muhammad Ayoub and Imam Yagoub},
  date-added = {2023-03-25 17:43:52 +0100},
  date-modified = {2023-03-25 17:43:52 +0100},
  doi = {https://doi.org/10.1016/j.eswa.2022.116562},
  issn = {0957-4174},
  journal = {Expert Systems with Applications},
  keywords = {Social media, Natural language processing, Neural network, Topic
modeling, Annotation, Classification, Transformer, Crisis response},
  pages = {116562},
  title = {Topic2Labels: A framework to annotate and classify the social media
data through LDA topics and deep learning models for crisis response},
  url = {https://www.sciencedirect.com/science/article/pii/S0957417422000604},
  volume = {195},
  year = {2022}}
```

#Thesis/Papers/Initial