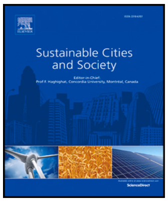
Sustainable Cities and Society 80 (2022) 103706

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs



CitEnergy : A BERT based model to analyse Citizens' Energy-Tweets

Jatin Bedi [a,*], Durga Toshniwal [b]

[a] *Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab, India*
[b] *Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee (IITR), Uttarakhand, India*

ARTICLE INFO

Keywords:
Deep learning
Twitter
BERT
Complaints classification
Sentiment analysis
Energy

ABSTRACT

Micro-blogging social site Twitter has emerged as a rich source of unstructured text information which could be processed and analysed to extract people's opinions about several topics and events including natural hazards, energy, sports, transportation, elections etc. The present research study adds a novel perspective in this dimension by extracting citizens' opinions on several electricity-related issues. Recent research studies in this domain have employed Bag-of-Words (BoW) model for the numerical representation of tweets. However, the BoW model suffers from several issues such as incapable of handling the semantic relationship between words, sparse and high dimensional representation. The present research work overcomes all aforementioned shortcomings by integrating popular word-embeddings with deep learning models for the classification tasks. The study harnesses social media data for two classification tasks: Sentiment classification task and Complaints classification task. Firstly, a series of preprocessing steps are applied on tweets extracted from Twitter streaming API. Subsequently, different word-embeddings models are employed to generate a numerical representation of tweets while capturing the semantic relationship among words. Several deep learning-based sentiment classification models are then deployed on top of generated word-embeddings for identifying/classifying citizens' sentiments from the tweets. Lastly, the tweets associated with negative sentiment class (identified by the sentiment classification model) are further processed and analysed for building the complaints classification model. The complaints classification model prioritize and assign negative sentiment tweets into one of the two target classes depending on the target issue raised in the tweets (Community level or recurring complaint and Individual level complaint). In addition to this, the current study also proposes Bi-directional Encoded Representations for Transformers (BERT) based Sentiment classification and Complaints classification models for achieving the improved classification accuracy. Experimental evaluation of the proposed BERT based models is done by comparing prediction results with several benchmark deep learning models.

1. Introduction

Energy contributes an essential factor to the technological, economic, and social development of a society (Sethi et al., 2020; Toman & Jemelkova, 2003). It is related to every domain of development including healthcare, transportation, national security, education, infrastructure and even with the living standard. Hence, every country is striving to provide an uninterrupted power supply to the people. The increasing population and recent developments in the Information and Communication Technology (ICT) including smart devices, sensors etc. have significantly increased the energy demand of country India (Iniyan et al., 1998; Shahbaz et al., 2016). In past years, enormous progress has been made by the country to improve energy reliability and efficacy. The Government of India (GOI) has launched several schemes to improve energy infrastructure for achieving 100% electrification and increased energy availability (of Power, 2012). To further plan, maintain and improve the energy system infrastructure and efficiency, it is necessary to know people's opinions on several energy-related issues including energy availability, infrastructure and other complaints. Collecting such information through the manual surveys is very expensive and time-consuming.

A compelling alternate solution is to make use of the latest technologies to know people's opinions on various topics and events. Several micro-blogging sites like Twitter, Facebook have emerged as a powerful source of informative text that could be analysed to extract citizens' viewpoints on any topic or event. Moreover, with the advent of big data technology and powerful tools to handle it, the extraction of useful information from text data has attained significant consideration from the research community. Sentiment analysis/Opinion mining has been widely adopted by researchers in numerous domains to monitor customers opinions. It is defined as the process of determining whether a piece of input text is positive, negative or neutral. In a similar

* Corresponding author.
E-mail addresses: jatin.bedi@thapar.edu, jbedi@cs.iitr.ac.in (J. Bedi).

https://doi.org/10.1016/j.scs.2022.103706
Received 20 October 2021; Received in revised form 14 January 2022; Accepted 15 January 2022
Available online 29 January 2022

context, one of the primary goal of the current research study is to contribute a novel perspective in this dimension by analysing citizens' sentiments or opinions on energy-related aspects (Liu, 2012). In the past decades, several dictionary-based approaches (Feldman, 2013a; Neviarouskaya et al., 2011; Nielsen, 2011) have been successfully utilized by the researchers for the task of sentiment analysis. These approaches calculate sentiment score by comparing the words in a text with a pre-defined dictionary of words. The dictionary-based approaches are broadly focused on sentiment or polarity of a word in the English language. So, these approaches raise several issues while adapting to a specific domain. For an instance, the polarity of words '*power*', '*energy*', '*light*' and '*smart*' is positive in context of the English language. However, the polarity of these words in the specific energy domain is neutral. To resolve these issues, building machine learning and deep learning models on a manually annotated domain-specific tweets dataset have gained popularity for improving the sentiment classification task. These models are based on utilizing the numerical representation of tweets for building the target solutions. In the earlier studies, Bag of Words (BoW) (Raju & Sridhar, 2020; Zhang et al., 2010) and TF–IDF (Domeniconi et al., 2015; Li & Shen, 2017) models have been extensively utilized for vector representation of text. BoW model performed well on classification task but handling high dimensional and sparse representation generated from the model is a great difficulty. Moreover, these models are not capable of dealing with the semantic relationship among the words. These problems get resolved with the introduction of word-embeddings which generates dense vector representation of text while preserving the semantic relationships among words (Giatsoglou et al., 2017; Onan, 2021; Rudkowsky et al., 2018; Salur & Aydin, 2020). However, to the best of our knowledge and extensive literature survey, no research studies have employed word-embeddings for sentiment classification in the energy-specific domain. Hence, in the present research study, we augment different word-embeddings (*Word2Vec, GloVe* and *FastText*) with several deep learning models (*DNN, LSTM, Bi-LSTM* and *CNN*) for the sentiment classification task.

Another critical aspect that has not received significant attention in the energy domain is to identify and categorize citizens' energy-related complaints from the tweets. The citizens' energy complaints can be categorized or classified and prioritized on the basis of several factors such as type and frequency of the issue, target sector etc. In this context, the present study proposes a deep learning-based solution for the categorization and prioritization of the complaints. The study integrates word-embeddings with deep learning models for this task. Furthermore, in the current study, we also propose a state-of-the-art Bi-directional Encoded Representation for Transformers (BERT) based solution for both Sentiment classification and Complaint classification tasks. The proposed BERT based solution helps to improve the accuracy of the classification tasks. The key research contributions of the present study are summarized as follows:

- The impact and reliability of social media (Twitter and Facebook) at capturing the citizens' behaviour/opinions on several issues have been well established in the literature (Ilieva & McPhearson, 2018). The current study harnesses Twitter data for two important tasks: (a) Sentiment Classification task (b) Complaints Classification task.
- For the first time in this domain, the amalgamation of word-embeddings (*Word2Vec, fastText and Glove*) and deep neural network models (*DNN, LSTM, Bi-LSTM and CNN*) are applied for Sentiments and Complaints identification from energy related tweets.
- A novel approach integrating Twitter and deep learning for building the Complaints classification model is proposed. The proposed approach prioritizes electricity-related complaints/tweets based on the target issues faced by the citizens' such as nature of the complaint, frequent or recurring issues and timestamp. Furthermore, a list of the ten most relevant identified keywords related to high priority complaints is also included.
- Bi-directional Encoded Representations for Transformers (BERT) model is employed to achieve state-of-the-art prediction accuracy for both Sentiment and Complaints classification tasks. The performance comparison of the proposed BERT based classification models is made with the existing benchmark models in terms of several widely adopted performance metrics.
- To demonstrate the applicability of the current work, BERT based model and other existing benchmark models are applied for both Sentiment classification and Complaints classification tasks of two Tier-1 cities of India (Delhi and Bangalore).

The rest of this paper is organized as follows: An extensive literature survey on text classification approaches is presented in Section 2. The dataset description and data preprocessing are discussed in Sections 3 and 4, respectively. A detailed methodology of the proposed Sentiment classification model and Complaints classification model is explained in Section 5. Experimental details and comparative results are discussed in Section 6. Lastly, the conclusion is stated in Section 7.

2. Literature survey

This section presents an extensive analysis of existing research studies in the field of sentiment/text classification. The overall discussion is partitioned into two sections depending on the type of techniques, namely Lexicon based approaches and Deep learning-based approaches.

2.1. Lexicon based classification approaches

In the past two decades, Lexicon or dictionary-based approaches have gained wide popularity for sentiment analysis task. Taboada et al. (2011) created a Semantic Orientation Calculator (SO-CAL) dictionary of annotated words for determining the polarity of the input text. The applications of the created dictionary to different domains are used for performance validation. Several research studies have widely adopted another popular lexicon AFINN introduced by Finn Nielsen for the sentiment analysis task. Feldman (2013b) presented a review of different levels for sentiment analysis, including document level, sentence level, aspect level, comparative sentiment etc. The applications and research issues associated with these different levels are also well-explained by the authors.

Later in the year 2018, Ikoro et al. (2018) presented the results of sentiment analysis on tweets posted by UK Energy consumers. The study explained several domain-specific issues while analysing the energy-related tweets using a sentiment analysis tool. Lastly, the study also compared the sentiment results of different energy entrants. Liu and Na (2018) propose a variant of the sentiment analysis task. The authors implemented a neural network to perform aspect level sentiment analysis. The proposed approach aims to analyse public opinions towards energy-related things. In the year 2019, Majumdar and Bose (2019) presented a study to analyse the impact of Twitter content on the target firms. A text mining approach is adopted to identify significant topics from the tweets. From the experimental results, the authors stated the existence of a positive association between twitter posts and target firm generated contents. Mogaji et al. (2020) performed lexicon-based sentiment analysis of consumers tweets for 82 companies in the United Kingdom. Furthermore, the study also investigated the role of qualitative factors shaping the consumers' behaviour towards target brands. Jain and Jain (2020) presented an approach for opinion analysis of citizens on renewable energy resources. The study performed a comparative analysis of citizens' tweets on various renewable resources like solar energy, bioenergy, wind power, hydro power and geothermal energy. The analysis results show that people are in more favour of renewable energy resources. In a similar context, Abdar et al. (2020) proposed an approach for sentiment analysis of energy-related topics of Alaskans' tweets. The authors presented the opinions and perception of citizens on energy-related topics to identify the problems faced by them.

In the past years, social media sites have gained exponential interest from people to express their views and opinions on a topic. In this context, Ilieva and McPhearson (2018) investigated the impact of social media (Twitter and flickr) data in assessing the sustainability-related topics of society. The authors presented the key challenges and opportunities for evaluating human behaviour using social media content. From the evidential results, the authors concluded that social media has an important role in understanding the various social aspects of society. Few other studies carried out in the same context include: (Liu & Hu, 2019) proposed an approach to analyse the attitude and sentiment of the Chinese public towards green buildings. The public comments/post were crawled from the weibo for the analysis task. Digitization is expected to impact the cities, jobs and lifestyle in a varied way. Balogun et al. (2020) have developed an approach to assess the potential of digitization in minimizing the impact of climate change-related shocks and stress. Rodrigues et al. (2020) analysed user engagement in managing and reducing energy use in a community. The authors collected information using several means, namely, social media, websites and other online platforms, to estimate a true picture of their perspectives.

2.2. Deep learning based classification approaches

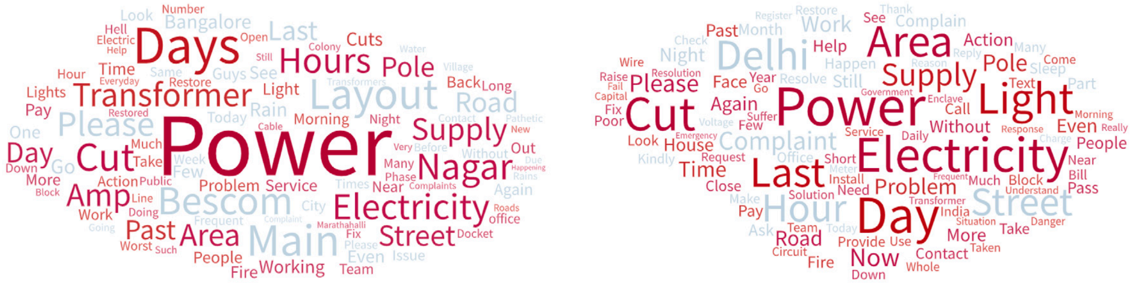
Deep learning has achieved wide popularity and success in nearly all domains, including computer vision, sensors, and natural language processing with real-world applications in healthcare, utility services, energy storage, transportation, and agriculture (Bedi, 2020, 2021; Fu et al., 2020; Tan et al., 2021). Feature extraction and representation play a critical role in building accurate models for several Natural Language Processing (NLP) related tasks such as text summarization, text classification and so on. Earlier studies in the NLP domain have utilized Bag of Words (BoW) model for feature extraction from text data. This model is based on the frequency of occurrence of words. Raju and Sridhar (2020) proposed a BoW model based approach for sentiment rating prediction from users reviews. From the analysis of experimental results, the authors stated that around 60% of rating could be estimated from users reviews. Wang, Cao et al. (2014) introduced a cross-media bag of words model for sentiment analysis. In addition to text, the approach also takes care of images posted by users. The comparative analysis of the approach is done by comparing results with Support Vector Machines (SVM) and Naive Bayes model. Da Silva et al. (2014) performed a comparative evaluation of the ensemble models for the sentiment classification task. Furthermore, the authors also compared the performance of BoW and feature hashing techniques at representation of tweets. In a similar context, Wang, Sun et al. (2014) evaluated the classification performance of three ensemble methods which were based on combining five popular machine learning algorithms (SVM, k-nearest neighbours, Entropy, Decision tree and Naive Bayes). From the performance evaluation done on ten public sentiment analysis datasets, the authors stated that the random subspace method works better than Bagging and Boosting ensemble methods. Although the BoW model is widely adopted for the classification tasks, there are several issues associated with this model such as high-dimensional output vector representation, increased computational cost & parameters of the learning model and many more. Different word-embeddings or vector representation approaches have been combined with deep learning models for providing solutions to the vector representation and text classification tasks in various domains, including healthcare, transportation, utility services and others. Some important research studies which have achieved promising results in the listed application domains are discussed as follows:

Huang et al. (2017) integrated word embeddings with CNN and LSTM model for the classification task. The features extracted from CNN are fed as input to the LSTM model for classification. The experimental results show that the integrated model has achieved good accuracy. Zhang et al. (2018) implemented deep learning models for traffic incident detection from social media. Deep belief networks (DBN) and LSTM models are implemented for this task. The classification results show that the belief network achieved highest accuracy and outperformed Support Vector Machines (SVM) and Latent Dirichlet allocation. Mauri et al. (2018) proposed an approach for extracting energy-related information from user-posted contents. The extracted social metring contents are classified into four categories (*dwelling, food, leisure and mobility*) using a logistic classification model. Dabiri and Heaslip (2019) proposed a deep learning approach for traffic events detection from social media data. The authors implemented CNN and LSTM model on top of FastText and Word2Vec embeddings for traffic event identification task. Stein et al. (2019) integrated prominent machine and deep learning algorithms including XGBoost, SVM and CNN with word-embeddings for the hierarchical text classification task. The performance evaluation on the public dataset indicated that word-embeddings contributes towards improved classification accuracy. Agarwal et al. (2020) investigated the reliability of Twitter contents for election results prediction. The authors combined word-embeddings with LSTM and Bi-LSTM models for election results prediction task. The experimental results demonstrate that the amalgamation of FastText and LSTM model provides the highest prediction accuracy (*87*%). Jain and Jain (2020) proposed a machine learning-based approach for renewable energy-related tweets classification. The authors integrated feature selection techniques with machine learning algorithms (SVM, K-nearest neighbour, AdaBoost, Naive Bayes and Bagging) for text classification. Salur and Aydin (2020) introduced a hybrid approach for the sentiment classification task. The author integrated deep neural models with different word embedding to leverage the benefits of both algorithms. Seo et al. (2020) performed a comparative analysis of eight deep neural network models (five based on CNN and three based on RNN model) for the sentiment classification task. Kim et al. (2021) proposed a deep learning based approach for sentiment classification of solar energy tweets. The approach results are validated on the dataset of United States. Alamoudi and Alghamdi (2021) proposed machine learning, deep learning and transfer learning based models for the sentiment classification task and achieved 83% accuracy using the proposed aspect extraction method. In the year 2021, Kim and Hong (2021) presented an approach for automatic identification of transportation complaints from the unstructured text. The authors achieved 90% classification accuracy. In the similar context, Blümel and Zaki (2021) proposed an approach integrating TF–IDF and word2vec with machine/deep learning models for automatic identification of customer complaints.

The existing research studies have well investigated the need and reliability of social media text at identifying and characterizing several sustainability-related topics (Balogun et al., 2020; Ilieva & McPhearson, 2018; Rodrigues et al., 2020). In the earlier studies, several lexicon-based approaches (AFINN, SO-CAL) have been widely adopted for the polarity estimation task (Ikoro et al., 2018; Jain & Jain, 2020; Taboada et al., 2011). However, these approaches lack capturing the contextual dependencies present in the data. Later, with advancements in the natural language processing domain, several feature extraction techniques were introduced to capture contextual/semantic information present in the input text. The combined benefits/applications of feature extraction techniques with neural network models have shown progressive results in several different application domains (Agarwal et al., 2020; Dabiri & Heaslip, 2019; Zhang et al., 2018). However, the applicability of these techniques in the targeted research domain (energy sentiments and complaints identification) has not been investigated in the literature. In this context, the current research provides a novel framework to identify, characterize and prioritize citizens' complaints from their tweets post.

3. Dataset description

- *Tweets Collection:* The present work aims to harness social media data to identify real-time electricity related issues. The first step



(a) Bangalore (b) Delhi

Fig. 1. Word cloud representation.

towards this goal is to collect social media data related to the problem under study. For this purpose, we have employed the Twitter Streaming API[1] to collect real-time tweets. The streaming API supports crawling tweets in two ways, either by using the hashtags or by geographical location. In the present case, we have initially collected the tweets based on geographical location for the country India. Subsequently, the desired domain-specific electricity related tweets are extracted from the keywords matching hashtags. The keywords/hashtags used for the target electricity related dataset extraction are listed in Table 1 The tweets are crawled for a period of six months, and we have collected around 20k tweets related to these keywords. Each tweet crawled using the Twitter streaming API contains a number of attributes which disseminate important information about the tweets. In the current study, we have stored only the following relevant information related to each tweet: {*text* (actual text of the tweet), *location*, *tweet_id* (unique identifier of the tweet), *userid* (unique identifier of the user), *created_at* (date and time when the tweet is created), *retweet_count* (number of time the original tweet is retweeted), *hashtags* and *mentions*}. The current study focuses on demonstrating the applicability of the work to two tier-1 cities (Delhi and Bangalore) of Country India. So, lastly, we have segregated the collected tweets with respect to location information for the city Delhi (2k) and Bangalore (1.7k). Fig. 1 shows respective Word Clouds for the city Delhi and Bangalore, i.e., a visual representation of the hundred most frequent keywords used by people in the electricity-related tweets.

4. Data preprocessing

Data accumulated from the Twitter Streaming API is unstructured text with some noise, metadata and irrelevant information. The reliability and accuracy of the learning models are closely related to the quality of data available for analysis. So, the collected must pass through several preprocessing steps to make it suitable for developing the deep learning models. The present study employs following filtering steps (with example demonstrated in Table 1) on the collected domain-specific tweets before labelling/assigning them to target sentiment classes:

- *Misspelling and Redundant Characters*: Some of the collected tweets contain words like '*energyyyyy*'. We have used regular expressions to remove repeating characters in a word.
- *Mentions and Hashtags Removal*: Hashtag ('#') symbol before a word or phrase is used for characterizing, categorizing and improving the search performance of a tweet. Mentions ('@') symbol is used to refer or tag another Twitter user in a tweet post. These symbols are removed from the collected tweets as they do not add any information to the learning model.
- Remove all URLs, special characters, numbers and punctuation marks (like a comma, full stop and brackets) from the tweets.
- Remove all tweets with less than five words from the collected dataset. Convert remaining tweets to the lowercase.
- *Stop words removal*: Stop words are the set of commonly used words in any language. These words do not add any value to the machine understanding of a text and can be removed. In the present case, we have collected the tweets corresponding to English language only. So, the stop word corresponding to English language (for example: '*a*', '*an*', '*in*', '*he*', '*the*' and many more) are removed from the tweets.
- *Contractions Handling*: Due to the restriction imposed by twitter on a post length that is characters limit, people commonly use contractions (for example: *'do not', 'cannot'* and so on) to write more. We have changed the contractions to actual words.
- *Stemming and Lemmatization* : aims to transform the derivational forms of the word to a common base form. This helps in reducing the corpus size and generating better word embedding.

5. Methodology

This section provides an in-depth understanding of the proposed sentiments and complaints classification models. Fig. 2 represents a multilevel architectural overview of the proposed framework.

A. Sentiment Classification Model

One of the major goals of the present study is to build an efficient deep learning based sentiment analysis model for the classification of electricity related tweets. The overall process of building sentiment classification model consists of following steps:

- *Tweets labelling:* For training and building the deep learning based sentiment classification models, we require labelled dataset. The segregated tweets corresponding to each selected geographical location (Delhi and Bangalore) are manually labelled into two sentiment classes, namely, class '*P*' and Class '*N*' (Step 2a: Fig. 2). These two classes represent the sentiment corresponding to each tweet. The Class '*P*' denotes that the tweet has associated '*positive sentiment*' and Class '*N*' represents that the tweet has associated '*negative sentiment*'.
- *Dataset Preparation*: The set of labelled tweets are divided into three parts: Train, Validation and Test parts. Train part is used to build a deep learning based sentiment classification model. The validation part is used to generate an unbiased evaluation of the trained model while tuning hyper-parameters. Lastly, the performance evaluation

[1] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.

Table 1
Tweet description.

Keywords/Hashtags			
{‘noelectricity’, ‘bsesdelhi’, ‘minofpower’, ‘electricityboard’, ‘powercut’, ‘powercuts’, ‘meterinstallationfraud’, ‘uhbvn’, ‘dhbvn’, ‘hvpnl’, ‘electricityindia’, ‘mpeb’, ‘msedcl’, ‘cea_india, ‘pvvnl’, ‘tneb’, ‘jvvnl’, ‘bses’, ‘mahadiscom’, ‘electricitycomplaint’, ‘sndl’, ‘mppkvvcl’, ‘mnreindia’, ‘onlinebescom’, ‘tssspdcl’, ‘nammabescom’, ‘bescom’, ‘tatapower’}			
Tweet attributes			
created_at	Date and time at when the tweet is created	*location*	Location information
text	Actual text of the tweet	*tweet_id*	Unique identifier of the tweet
rt_count	Number of times the original tweet is retweeted	*user_id*	Unique identifier of the user
hashtags	Tags to easily define and categorize the tweets	*mentions*	To refer to someone other in a tweet
Tweet preprocessing example			
Original tweet	This reply does not help. why have you guys written that the power will be back by 1 am and hello .. do you guys even look at the time. It is 3 am already. @NammaBESCOM		
Hashtags & mentions removal	This reply does not help. why have you guys written that the power will be back by 1 am and hello .. do you guys even look at the time. It is 3 am already.		
Contractions handling	This reply does not help. why have you guys written that the power will be back by 1 am and hello.. do you guys even look at the time. It is 3 am already.		
Lowercase conversion, punctuations and numbers handling	This reply does not help why have you guys written that the power will be back by am and hello do you guys even look at the time it is am already		

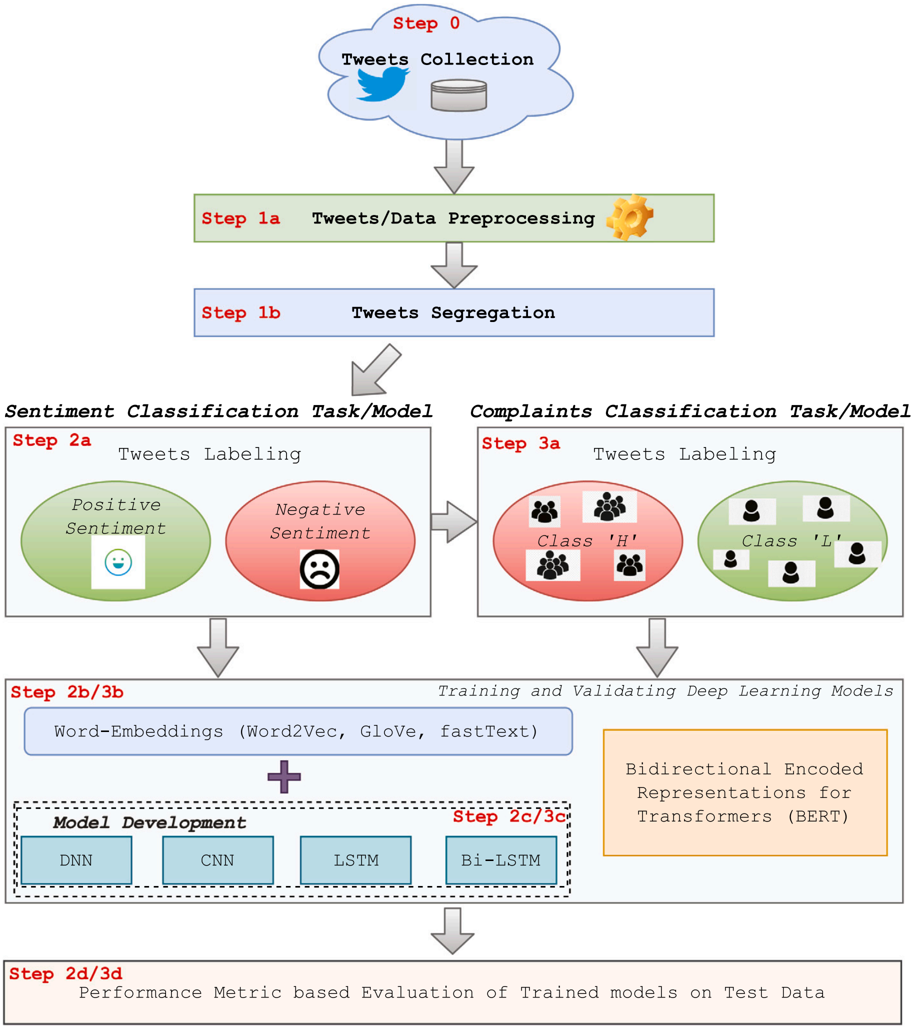

Step 0
Tweets Collection
Step 1a
Tweets/Data Preprocessing
Step 1b
Tweets Segregation
Sentiment Classification Task/Model
Complaints Classification Task/Model
Step 2a
Tweets Labeling
Positive Sentiment
Negative Sentiment
Step 3a
Tweets Labeling
Class 'H'
Class 'L'
Step 2b/3b
Training and Validating Deep Learning Models
Word-Embeddings (Word2Vec, GloVe, fastText)
Model Development
Step 2c/3c
DNN
CNN
LSTM
Bi-LSTM
Bidirectional Encoded Representations for Transformers (BERT)
Step 2d/3d
Performance Metric based Evaluation of Trained models on Test Data

Fig. 2. In-depth methodology of the proposed complaints classification framework.

of the proposed classification model is done on the test (unseen) part.

- *Building Sentiment Classification Model*: Deep neural network models have shown great success at providing accurate solutions to various NLP tasks including question answering systems, text classification & summarization, sentence matching and other machine translation tasks. In this study, we implement three supervised deep learning models namely, LSTM, Bi-LSTM and CNN for sentiment classification tasks. These deep neural network models are built on the top of well-known and efficient word-embedding models namely Word2Vec, FastText and GloVe. The integration of word embeddings with deep neural network models (LSTM, Bi-LSTM and CNN) is explained in the subsequent subsections. Furthermore, we also employ state-of-the-art BERT model for the classification task to achieve best prediction performance.

5.1. Word embeddings

In contrast to other artificial intelligence application domains/tasks where inputs are given in the form of vector space, the neural network learning models faces challenges while dealing with the text inputs. The possible solution in this context is to map words vocabulary to a vector representation form (known as word embedding). One commonly used representation is "One hot encoding" in which each word is represented with a $|W|$ dimensional vector space where $|W|$ denotes the size of the vocabulary. All other entries in the $|W|$ dimensional vector are set to zero except one which defines the word. Since each word is represented with a high dimensional and sparse vector, this encoding scheme has some major shortcomings such as incapable of handling context similarity, high computational requirement while dealing with learning models etc. An alternate solution is to generate dense word vector embeddings of the words to capture both semantic and syntactic context of the words (Liu & Zhang, 2018; Rudkowsky et al., 2018). The dense vector mapping should be done in a way that the words with similar meaning should appear close in the vector form. In recent years, several dense vector embeddings are proposed to improve the representation, generalization and computational powers of the neural network based text classification models. In this section, several popular or well known word embeddings and their interaction mechanism with neural network models are discussed (Step 2b: Fig. 2).

5.1.1. Word2Vec

Word2Vec (Goldberg & Levy, 2014; Mikolov et al., 2013) is an unsupervised learning algorithm to generate high dimensional distributed vector representation of all words in a corpus. It is a two layer neural network model which takes text as an input and generate vector representation for each word as output. There are two possible architecture of the Word2Vec model, namely, Skip-gram model and Continuous Bag of Words (CBOW) model. The architectural representation of Word2vec Skip gram model is given in Fig. 3. The Skip-gram model aims at predicting the context words for a given input word. It works by initially generating a V-dimensional "*One hot encoding*" representation of each word in the corpus. Subsequently, the generated word representation is given as input to the Skip-gram model with a single hidden layer. The number of neurons in the input layer and hidden layer are set equal to the vocabulary size and dimensions of the word vectors respectively. The weights between the input and hidden layer are denoted as $W_{V\times N}$, where V denotes the vocabulary size and N represents the hidden layer size. Each row in the input weight matrix denotes the N-dimensional vector representation of a word in the corpus. Similarly, the weights matrix between the hidden layer and output layer are denoted as $W'_{N\times V}$. This weight matrix is used to compute the score for each word in the vocabulary for a given input word. The output from the network model is a probability vector with values denoting the probability of occurrence of each word in context of the given input word. The loss function at the output computes the error values and back-propagation is utilized to update the weights (word embeddings). In this way, model helps to learn the relationship between different words in a vocabulary. Another Word2Vec architecture CBOW, is reverse to the Skip-gram model, takes the context of the word as an input and tries to predict the word corresponding to the context (Mikolov et al., 2013).

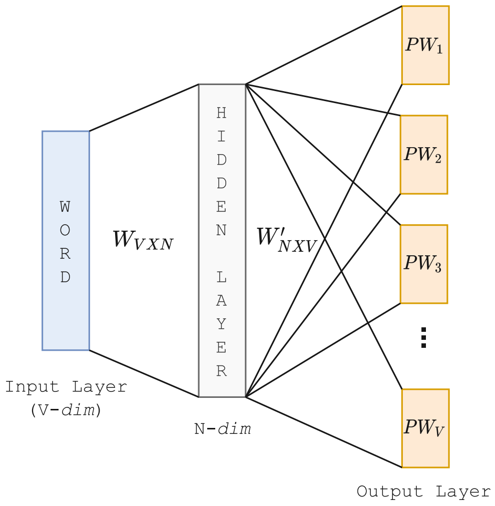


Fig. 3. *Word2Vec*: Skip-gram model architecture.

5.1.2. GloVe

GloVe stands for "*Global Vectors*" is an unsupervised learning algorithm based on utilizing the global frequency co-occurrence ratio of words together within in a window (Pennington et al., 2014). It is a combination of two methods which are matrix factorization approach and local context window. Matrix factorization approach provides a way to reduce the large term-frequency matrices and is termed as Latent Semantic Analysis (LSA). The method generate efficient vector space substructure, but does not work well on capturing the words analogy. In contrast to this, the local context window (Skip-gram) method works well on analogy task but fails to properly utilize the corpus statistics. The GloVe algorithm combines both by analysing local and global statistics of a corpus (Shi & Liu, 2014). The method aims at generating word vectors in a way that their dot product equals to the log of words co-occurrence probability.

5.1.3. FastText

FastText (Joulin et al., 2016) is an extension of the Word2Vec model and is proposed by Facebook in the year 2016. An important shortcoming of the Word2Vec and GloVe algorithms is that these models are not capable of efficiently generating the embeddings for rare words in the corpus. FastText overcomes this by working at a more deeper level. Instead of feeding individual words to the neural model, FastText model works by representing each word as n-grams of characters. For example, the tri-grams corresponding to word '*electricity*' are [*'el', 'ele', 'lec', 'ect', 'ctr', 'tri', 'ric', 'ici', 'cit', 'ity', 'ty'*]. After representing the input word using characters n-grams, a Skip-gram model is trained to learn the embedding for each n-gram. Finally, the word embedding for the word '*electricity*' is given by the summation of the embeddings for all these n-grams. In this way, even if a word has not appeared in the training vocabulary, its n-grams can be used to generate word embeddings as it is likely that some of its n-grams might appear in other words. Lastly, an important observation about FastText model is that it has high memory and system requirements since it generates word embeddings for each character n-grams of the words in the vocabulary (Xu & Du, 2019).

5.2. Model development

5.2.1. Embedding layer

The embedding layer is the first layer of a deep neural network model. It aims to generate continuous and distributed vector representation of the words in a corpus. There are four important arguments to the embedding layer. The first argument is size of the vocabulary (equal to the number of unique words in a text corpus). The second argument specifies the dimensions of the target vector embedding space (output dimensions = *300*) in which words are to be embedded. The third argument '*L*', defines the length of the input sequence. As tweets belonging to a corpus may have varying number of words, this argument is specified on the basis of percentile distribution of the number of words in tweets. The encoded representation of tweets with less than '*L*' words are padded with zeros and the tweets with more than '*L*' words are truncated.

The final argument for the embedding layer is embedding matrix. The embedding matrix is created from the word embeddings explained in the previous subsection. To create an embedding matrix, the '*d*' dimensional vector representation of each unique word in the corpus is retrieved from one of the pre-trained embedding models (*Word2Vec*, *GloVe* and *FastText*). The concatenation of these vector representation will form a matrix of shape [*L*×*d*] and is termed as *word-embedding matrix*. During training, for each of the tokenized input word in a tweet, the embedding layer will learn an optimal mapping of words to a vector of real numbers with size equal to the output dimensions. After proper training, the embedding matrix will have a similar multi-dimensional representation for the contextually similar meaning words of the corpus. The output from the embedding layer will be fed as input to the deep neural network models (*CNN, LSTM and BiLSTM*) for the sentiment classification task (Step 2c: Fig. 2).

5.2.2. Deep Neural Network (DNN)

A deep neural network (Goodfellow et al., 2016) is an artificial neural network with multiple hidden layers between the input and output layers. These models are better capable of handling the complex non-linear relationship between input and output data than the traditional single hidden layer neural networks. DNNs are typically feed-forward neural network where data flows from the input to output layers. Each layer of the network is fully connected to previous & next layer and extracts distinct features from the input. The connection between the neuron in the adjacent layers have associated weights which get adjusted during the network training phase. The adjusting of connection weights is done on the basis of error values computed at the output layer. The back-propagation algorithm is employed for this task. As these networks have more hidden layers, therefore, have a large number of parameters and are difficult to train than traditional artificial neural networks.

5.2.3. Convolutional Neural Network (CNN)

CNN (O'Shea & Nash, 2015; Prabhu, 2018) has achieved ground-breaking results in pattern recognition, natural language processing, computer vision and time-series activities by extracting spatially dependent features. Unlike the traditional neural networks in which each neuron is connected to every other neuron in the adjacent layers, CNN works on the concept of receptive field. The neural network implements convolutional layers to map an input to the output. The convolutional layers introduce spatial correlation by connecting a region of input neurons to the output, thereby reducing the number of learning parameters. This layer can be considered as a sliding window filter applied to every part of the input volume. The sliding window filter starting from the top of the input volume calculates the dot product of the input filter with the respective input volume and then slides by one step (stride). In this way, the process is repeated for the entire input volume. The output from the convolutional layer is a feature map generated from the combination of all dot-product values. In convolutional layer, we can define multiple filters, consequently, creating multiple feature maps. Each filter provides a way to learn and extract different feature of the data. In this way, inclusion of multiple filters in the model can lead to extraction of multiple features from the data.

The architectural representation of the CNN model used in the present study is demonstrated in Fig. 4. We have the input volume of size '*L*× *d*', where '*L*' defines length of the input sequence and '*d*' represents dimensions of the word-vector space. In convolution layer, we define '*k*' filters, each of size *n*×*d*, where *n* denotes the number of consecutive words to be considered at a time (n-grams). Sliding each defined filter over the entire input volume while taking dot-product generates a feature map of size (*L-n+1*). Since, we have employed '*k*' filters, the concatenation of these filters will create a matrix of size [*L-n+1, k*]. In the next stage, a pooling layer is applied on each feature map to select the feature with the highest value. The idea behind pooling is downsampling to reduce the computational complexity, avoid over-fitting and return high level & abstract features of the data. The output from the pooling layer is fed as input to the fully connected dense layer. Similar to ANNs, each neuron in the fully connected layer is connected to every other neuron in the adjacent layers. Finally, the last layer is '*Softmax*', which calculates the probability distribution of an input sample over all target classes.

5.2.4. Long Short Term Memory Network (LSTM) and Bi-directional LSTM (Bi-LSTM)

Artificial neural networks have been widely adopted for classification and prediction in nearly all applications domain. These models treat each input sample as independent of the other samples, therefore, are not found useful at handling contextual dependencies present in the data. Recurrent Neural Network (RNN) (Sherstinsky, 2020) serves the desired purpose by utilizing previous state information for handling contextual dependencies. They treat each word of the text sequence as a separate input occurring at a time '*t*' and also uses previous hidden state information to process the current input. Despite the great success in NLP and machine translation tasks, RNN model suffers from some major limitations such as incapable of handling long-term contextual dependencies, vanishing and exploding gradient problems (Sherstinsky, 2020).

Long Short Term Memory Network (LSTM) (Hochreiter & Schmidhuber, 1997), a variant of RNN, introduced by Hochreiter and Schmidhuber overcame the shortcomings of the vanilla RNN. The network employs a gating mechanism to capture long-term contextual dependencies of the text. Fig. 5 shows the architecture of a single LSTM unit with gating mechanism. In the LSTM network, we may have a series of these LSTM units to process the information. The improved memorizing mechanism with the help of gated LSTM model works as follows: The first part of the mechanism is to enable LSTM model to forget the irrelevant information. This is achieved by means of forget gate f_t and is given as:

$$f_t = \sigma(W_f . [x_t,\ h_{t-1}] + b_f) \tag{1}$$

The next part to decide important information to be stored in the current cell state is taken care by input gate i_t. Mathematically, it is given as:

$$i_t = \sigma(W_i . [h_{t-1},\ x_t] + b_i) \tag{2}$$

$$\tilde{S}_t = tanh(W_S . [h_{t-1},\ x_t] + b_S) \tag{3}$$

The *tanh* and *sigmoid* creates a vector of new candidate values and decides which of these information is to be stored respectively. The combination of these two with old cell state determines the new cell state given by:

$$S_t = f_t \odot S_{t-1} \oplus i_t \odot \tilde{S}_t \tag{4}$$

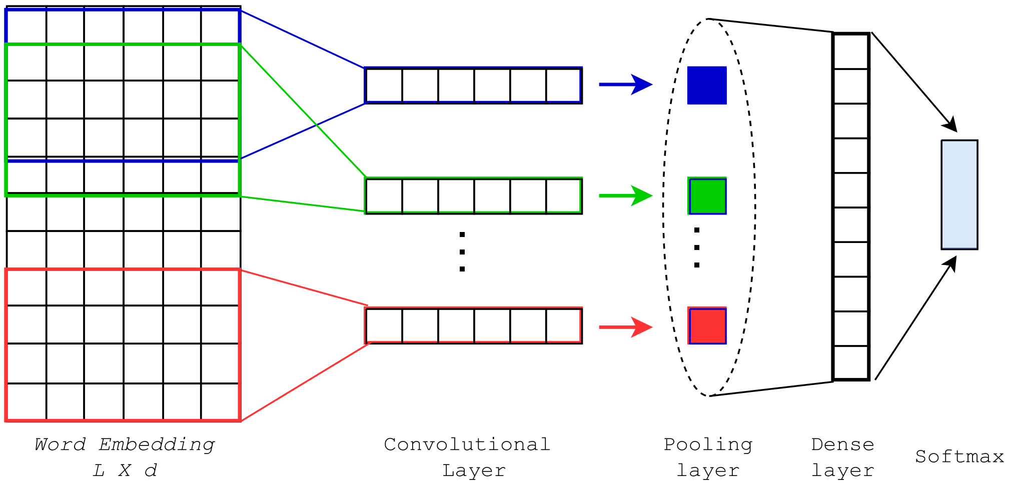


Fig. 4. CNN model architecture.

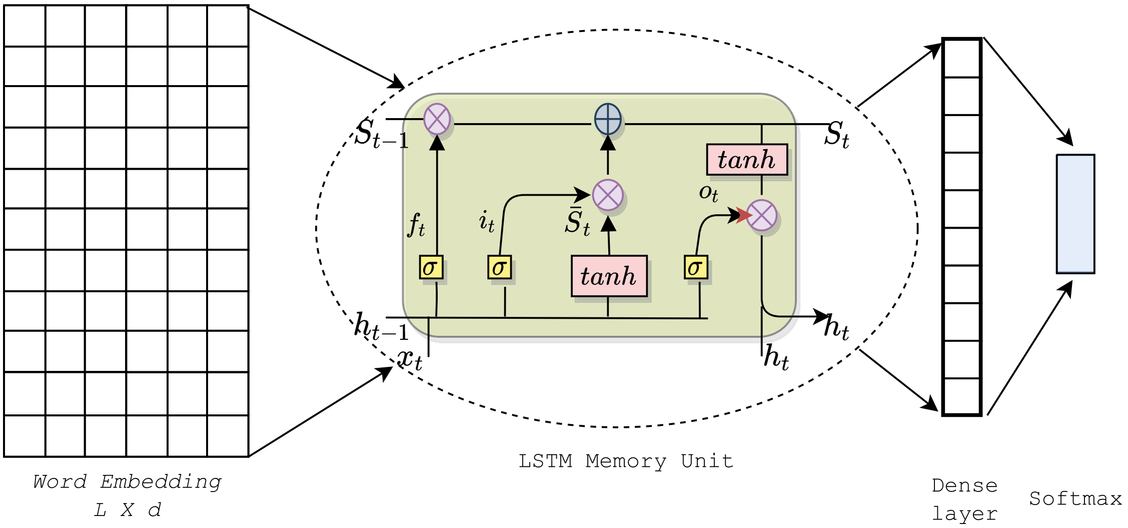


Fig. 5. LSTM model architecture.

The final step to determine the output (o_t) of a memory cell is given as:

$$o_t = \sigma(W_o.\ [h_{t-1},\ x_t] + b_0) \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

where $\odot$ and $\oplus$ denote element-wise multiplication and element-wise summation operation, W_i, W_f, W_S represent weights and b_i, b_f, b_S represent bias values, x_t represents the input at timestamp t and h_t represents the hidden state at timestamp t.

RNN and LSTM models are capable of capturing contextual information in one direction of a language. They consider that future words have no contextual relationship with the word under current consideration. Bi-directional LSTM (Bi-LSTM) model employs two independent recurrent networks to capture contextual information in both backward and forward direction. The network model is fed with input once from start to end and once in the reverse direction to preserve contextual dependencies of both directions.

5.2.5. Bidirectional Encoded Representations for Transformers (BERT)

Text representation plays a vital role in language modelling tasks. Pre-trained contextual representation models like Word2Vec, GloVe generate a unique/single representation for each word in the text. They consider that the word context/meaning remains same across the text corpus. However, this is not the case. In contrast to these existing NLP models which looks at a sequence from either left to right or right to left, BERT support bidirectional training to learn better context. BERT (Devlin et al., 2018) was introduced by Google AI in the year 2018 and is a state-of-the-art model for natural language processing tasks. It makes use of an attention mechanism to identify the context of a word in relation to all other words present in a text sequence. In this way, the model uses both previous and next context to generate a representation of the words present in the corpus.

The architectural representation of the BERT classification model is shown in Fig. 6. BERT model (Devlin et al., 2018) has several encoder layers (12 for the base model and 20 for the large model). These layers also have a large feed-forward neural network and attention heads. Each encoder layer applies self-attention, then passes the results to a feed forwarded neural network and finally to the next encoder. The input to the model is a sequence of words with [*CLS*] token at the beginning. The output from the model is a vector representation of the sequence. For the classification task, BERT model uses the final output of the first token [CLS] as a representation of the input sequence. This output representation, in conjunction with the feed-forward neural network and *Softmax* layer, determines the target class distribution. In the present case, we have employed a pre-trained uncased BERT model after fine-tuning the model parameters on our classification dataset.

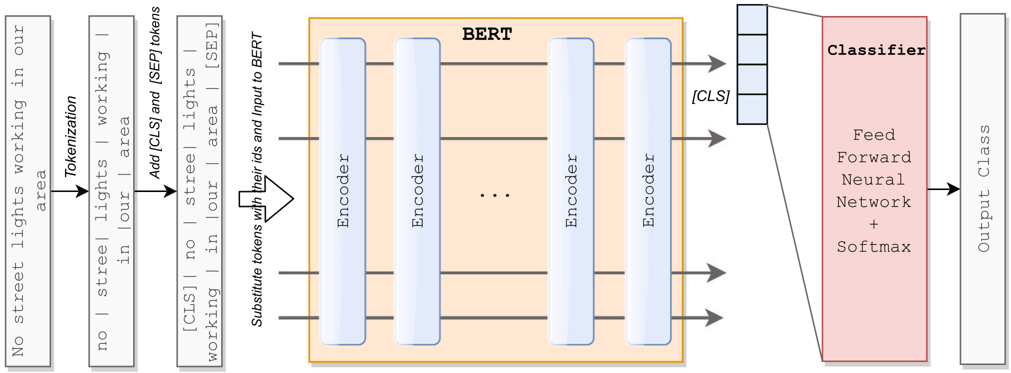

No street lights working in our area
Tokenization
no | street| lights | working | in |our | area
Add [CLS] and [SEP] tokens
[CLS]| no | stree| lights | working | in |our | area | [SEP]
Substitute tokens with their ids and Input to BERT
BERT
Encoder
Encoder
. . .
Encoder
Encoder
[CLS]
Classifier
Feed Forward Neural Network + Softmax
Output Class

Fig. 6. Architectural representation of BERT model.

Table 2
Details of various hyperparameters related to deep neural models.

DNN	CNN	LSTM/Bi-LSTM
num_dense_layers: [4–10],	num_filters = [2–10]	input_size = depends on embedding
num_neurons_layer: [48–200]	size_of_filter = [16–128]	num_lstm_blocks = [2–6]
activation = ‘ReLU’	embedding_matrix size = 300	num_lstm_units = [8–256]
loss = ‘binary_crossentropy’	max_pooling	out_size = 1
	activation = ‘ReLU’	optimizer = ‘adam’
	loss= ‘binary_crossentropy’	loss =‘binary_crossentropy’
	dense_layer_activation = ‘sigmoid’	activation = ‘ReLU’
		dense_layer_activation = ‘sigmoid’

B. Complaints Classification Model

The second major goal of the current study is to help in the identification/classification of electricity-related complaints. For this task, we propose a deep learning-based model to classify the tweets regarding citizens’ electricity-related complaints. The proposed model classifies the input tweets into one of the two below-mentioned target classes depending on the severity of the complaints.

- Class *‘H’ (High-Priority Complaints)*: primarily represent the ‘*community-level complaints*‘ and include tweets which are reported by a group of individuals, and hence should be given more importance. This class also includes the tweets/complaints belonging to the frequent or repetitive issues faced by the citizens. These complaints belongs to the electricity-related issues in public places, streets, offices, etc.
- Class *‘L’ (Low-Priority Complaints)*: represent the ‘*individual level complaints’* and include tweets which are reported by a single person or from a very limited group of people. Primarily, these complaints come from household faults, power-cuts, billing issues etc.

The overall process of providing a deep learning-based solution to the complaints classification task consists of the following two steps:

- *Tweets labelling*: The sentiment classification (explained in Section 5) model assigns input tweets to one of the two target classes (Positive: ‘*P*’ or Negative: ‘*N*’). From the tweets belonging to negative sentiments, we manually identify and assign each tweet (Step 3a: Fig. 2) into one of the following two classes namely, *High-Priority complaint* (‘*H*’) or *Low-Priority complaint* (‘*L*’).
- *Building Complaints Classification Model*: For the purpose of building and evaluating the deep learning based complaints classification models, we initially divide the labelled tweets dataset into three parts: *Train* part, *Validation* part and *Test* part. Similar to the sentiment classification model (explained in Section 5.2), we integrate different word-embeddings (Word2Vec, FastText and GloVe) with deep learning models (*DNN*, *LSTM*, *Bi-LSTM* and *CNN*) for the complaints classification task. Lastly, Google BERT model is also employed to achieve state-of-the-art prediction performance for the complaints classification task.

From the analysis of tweets belonging to electricity-related complaints, we identified the following ten most frequent words to characterize the tweets pertaining to *High-Priority Class ‘H’*: [‘*street lights*’, ‘*fire*’, ‘*daily basis*’, ‘*short circuit*’, ‘*pole*’, ‘*transformer*’, ‘*many days*’, ‘*many complaints*’, ‘*area supplies*’, ‘*from days*’].

6. Experimental results and discussion

6.1. Hyper-parameters selection

Hyper-parameters are the backbone of a learning algorithm. They play an important role in defining the accuracy and reliability of the deep learning models. Determining optimal hyper-parameters value is a difficult and computationally expensive task. The simplest way is to utilize the default value of parameters for building a model. However, this is not an effective method for building an accurate classification model. In the present study, the value of important hyper-parameters such as number of CNN filters, size of filter, number of LSTM units, activation function etc. is determined using the grid-search method (Lee, 2019). The grid-search method works by initially defining a grid of hyper-parameters values and then evaluating the cross-validation accuracy of the model on each combination of specified hyper-parameters. Finally, the model returns a set of parameters value for which the model achieved best accuracy. For the above-specified hyper-parameters, we evaluated the performance of learning models on grids with following range of values: size of filter (2–6), number of filters (50–200), and number of LSTM units (5–100). Based on the cross validation accuracy, the following values are chosen for the hyper-parameters: size of filter

Table 3
Performance comparison of proposed BERT based Sentiment Classification model with other benchmark models on Delhi dataset (*W2V*: Word2Vec, *GV*: GloVe and *fT*: fastText).

Model + Embedding	DNN			LSTM			Bi-LSTM			CNN			BERT
	W2V	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	
Accuracy	72.45	72.68	74.92	76.15	79.16	80.32	75.69	76.98	78.09	78.70	80.09	81.71	**90.32**
Precision	76.22	79.42	76.92	80.06	83.33	84.58	79.43	80.77	83.57	83.87	82.33	84.67	**91.01**
Recall	81.04	78.29	79.27	84.34	81.39	83.64	81.85	81.08	80.96	83.27	84.88	86.24	**95.29**
F-measure	78.55	78.85	78.88	82.14	82.35	84.11	80.62	80.93	82.34	83.57	83.58	85.45	**93.10**

Table 4
Performance comparison of proposed BERT based Sentiment Classification model with other benchmark models on Bangalore dataset (*W2V*: Word2Vec, *GV*: GloVe and *fT*: fastText).

Model + Embedding	DNN			LSTM			Bi-LSTM			CNN			BERT
	W2V	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	
Accuracy	70.80	70.94	71.83	75.45	76.12	78.29	72.05	74.41	76.92	79.58	80.81	83.20	**91**
Precision	72.51	74.50	75.51	78.92	78.07	80.24	81.48	78.51	78.31	80.15	82.57	85.35	**92.59**
Recall	79.25	78.90	79.01	81.27	83.85	85.04	66.66	80.61	83.33	85.64	84.47	87.17	**90.90**
F-measure	75.73	76.63	77.24	80.08	80.85	82.57	73.33	79.33	80.74	82.80	83.50	86.25	**91.74**

Table 5
Performance comparison of proposed BERT based Complaints Classification model with other benchmark models on Delhi dataset (*W2V*: Word2Vec, *GV*: GloVe and *fT*: fastText).

Model + Embedding	DNN			LSTM			Bi-LSTM			CNN			BERT
	W2V	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	
Accuracy	77.77	78.39	82.09	85.80	87.03	89.50	83.03	86.41	87.65	91.97	92.59	94.06	**95.78**
Precision	76.84	79.77	78.88	87.20	82.60	87.20	86.66	88.37	94.59	94.04	94.18	93.97	**98**
Recall	83.90	80.68	87.65	86.20	93.82	92.59	82.65	87.35	81.39	90.80	92.04	95.29	**94.23**
F-measure	80.21	80.22	83.04	86.70	87.86	89.82	84.60	87.35	87.50	92.39	93.10	94.62	**96.07**

Table 6
Performance comparison of proposed BERT based Complaints Classification model with other benchmark models on Bangalore dataset (*W2V*: Word2Vec, *GV*: GloVe and *fT*: fastText).

Model + Embedding	DNN			LSTM			Bi-LSTM			CNN			BERT
	W2V	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	*W2V*	*GV*	*fT*	
Accuracy	74.78	76.17	81.73	80.86	82.60	90.43	79.95	81.73	84.34	87.82	90.43	92.17	**94.64**
Precision	74.19	76.18	76.19	77.61	76.81	92.45	79.12	80.01	78.78	77.61	76.91	92.45	**97.14**
Recall	77.96	79.68	91.07	88.13	92.98	87.50	83.15	84.21	92.85	91.52	92.98	92.85	**94.44**
F-measure	76.03	77.89	82.92	82.53	84.12	89.90	81.08	82.05	85.24	88.52	90.59	92.65	**95.77**

(5), number of filters (128) and number of LSTM units (16). Another important parameter of interest is to determine the number of training epochs. For this, we have implemented the early stopping method which exits when accuracy value gets stabilized i.e., when change in a model accuracy over subsequent iterations is very low. The details of other hyper-parameters related to the deep learning models employed in the current study is listed in Table 2.

6.2. Performance metrics

Performance metrics provide a way to evaluate the reliability and accuracy of a prediction model. The following metrics are used in the current study for the evaluation purpose:

- *Prediction Accuracy* (Bishop, 2006): measures the fraction of samples which are correctly classified by the prediction model and is given as:

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ samples} \tag{7}$$

- *Precision* (Bishop, 2006): measures correctness and defines the fraction of samples which actually turns out to be positive from the totally classified positive samples. It is given as:

$$P = \frac{TP}{TP + FP} \tag{8}$$

- *Recall* (Bishop, 2006): measures completeness and defines the fraction of positive samples correctly classified by the classifiers and is given as:

$$R = \frac{TP}{TP + FN} \tag{9}$$

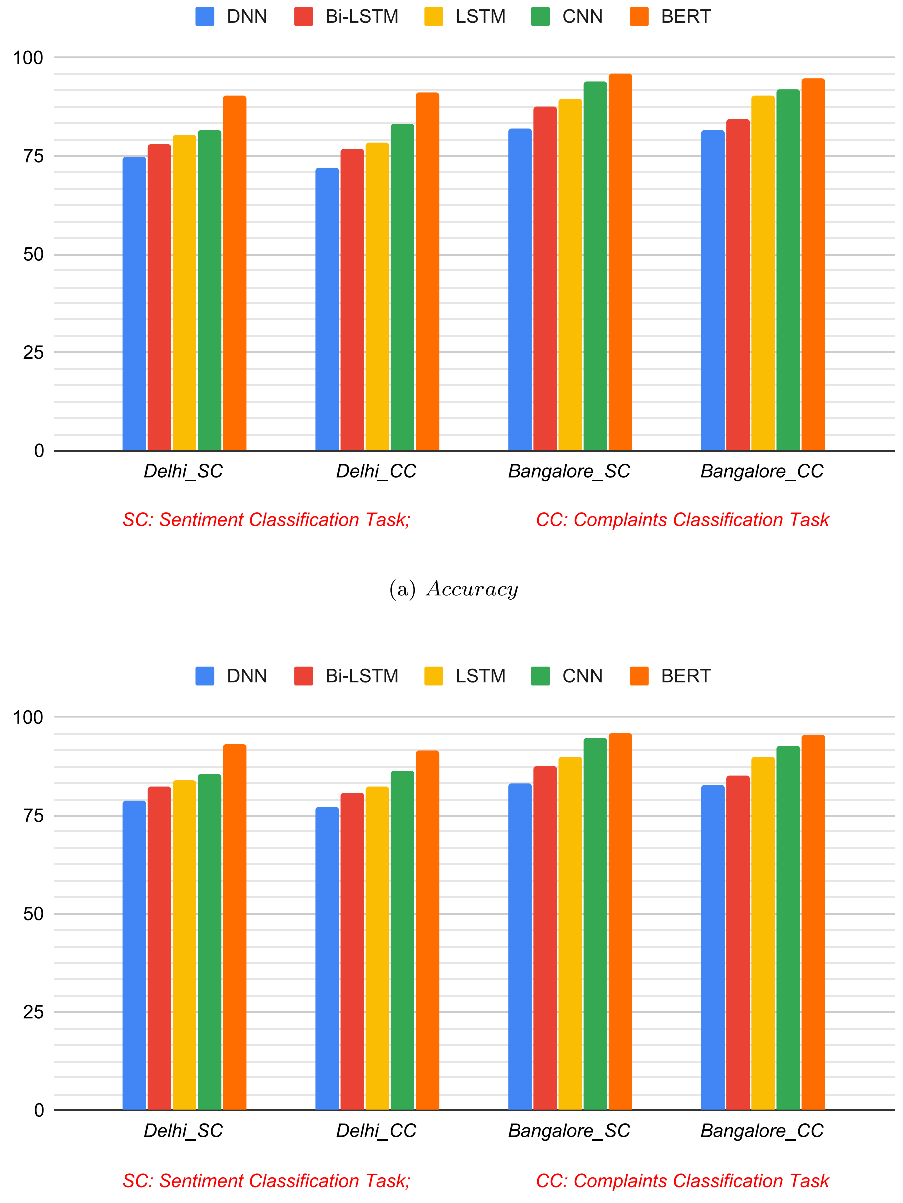
- *F-measure* (Bishop, 2006): combines precision and recall and is given by the harmonic mean of precision and recall.

$$F - measure = 2 \times \frac{P * R}{P + R} \tag{10}$$

Here, FN denotes the samples classified as negative by model and are actually positive, FP denotes the data samples classified positive by the model and are actually negative and *TP* & *TN* denote the samples which are correctly classified as positive and negative by the prediction model respectively.

6.3. Comparison results

To demonstrate the applicability of Proposed BERT based model on Sentiment Classification and Complaints Classification tasks, the proposed classification model is applied on tweets dataset of two Tier-1 cities of India (Delhi and Bangalore). The details of the collected tweet datasets are given in Section 3. Four popular state-of-the-art deep learning classification models (*DNN*, *LSTM*, *Bi-LSTM* and *CNN*) are implemented to compare the prediction results with Proposed BERT based classification model. These deep learning models are integrated with different popular word-embeddings (*Word2Vec*, *FastText* and *GloVe*) to achieve best classification results. Four well-known metrics listed in Section 6.2 are used to evaluate the classification performance of the proposed approach. Furthermore, the present study employs



(a) *Accuracy*

(b) $F - measure$

Fig. 7. Comparison of accuracy and F-score for Sentiment Classification (SC) and Complaints Classification (CC) tasks.

k(5)-fold cross-validation to analyse the prediction performance of the deep learning models. Cross-validation approach helps to define the generalized prediction performance of the learning models. It reduces the chances of over-fitting by giving a less biased model. The step by step working of cross-validation approach is explained as follows: (a) Partition the entire dataset into k(5) equal parts to be used for training and model evaluation. (b) Training the deep learning models on (k-1) parts and then evaluating the performance of the trained model on kth unseen (test) part. (c) Repeating steps (a) and (b) until each fold serves as a test part. (d) The final model accuracy is given by the average of accuracy achieved on the different test parts. In this way, each tweet will get a chance to appear in the test part.

The comparison results of the proposed BERT based classification model with other existing state-of-the-art deep learning models for sentiment classification task are listed in Table 3 (*Delhi*) and Table 4 (*Bangalore*). Likewise, Table 5 (*Delhi*) and Table 6 (*Bangalore*) show the comparison results of proposed BERT based classification model with other existing benchmark deep learning classification models for the Complaints Classification task. Moreover, to provide a better understanding of comparison results, Fig. 7 visualizes the classification performance of the proposed model and other benchmark models in terms of accuracy and F-score measures. From the comparison results listed in Fig. 7 and Tables 3, 5, 4 & 6 for both Sentiment and Complaints classification tasks, the following inferences are drawn:

- In comparison to *Word2Vec* and *GloVe* word-embeddings, the integration of *FastText* word-embedding with all state-of-the-art classification models supports improved classification results.
- Convolution Neural Network (*CNN*) in combination with *FastText* word-embedding provides best accuracy than other existing benchmark deep learning based text classification models.
- Bi-Directional Encoded Representation for Transformer (BERT) model outperforms all other state-of-the-art text classification models by providing the highest classification accuracy and F-score on Sentiment Classification and Complaints Classification tasks.

7. Impact and usefulness of the proposed approach

The current research work provides an in-depth insight into several statistical and analytical factors. The list of important inferences/ insights that can be drawn by employing the proposed approach in the real-time are listed as follows:

- For policymakers, AI Scientists and government organization:
  - Spatially resolved identification of critical areas or hotspots facing high energy-related issues and planning policies to proactively resolve those issues.
  - The proposed approach can be employed to identify areas with the need for more energy resources and extracting real-time opinions of people on resources availability and government policies.
  - Renewable resources planning: Ranking of places based on the need for alternate resources to satisfy the current demand in an area or to resolve the energy availability issues in that area.
- For New Researchers: A deep insight into analytical theory, text mining and most advanced deep learning models.

8. Conclusion

The current research study has exploited Twitter data to propose an accurate deep learning models based approach for energy-related tweets analysis. The approach implements a three-stage architecture for identifying citizens' sentiments and complaints raised by them. Raw Twitter data collected by employing streaming API is very noisy and contains a lot of meta-information. In the first stage, several data preprocessing steps are applied to improve the quality of text data. Manual labelling strategy is utilized to obtain data for building the learning models. The second stage consists of several sub-steps: initially, it involves generating distributed vector representation of the tweets using different embeddings (*Word2Vec, GloVe and FastText*). Subsequently, these word-embeddings are augmented with four deep learning models (*DNN, LSTM, Bi-LSTM* & *CNN*) for the sentiment classification task. Additionally, this stage also involves building the most enhanced BERT model for the sentiment classification task. The third stage includes building the complaint classification models from the negative sentiment class tweets recognized by the sentiment classification model. Similar to the sentiment classification task, BERT and word-embeddings integrated deep learning models are built for the complaints classification task. Finally, the proposed BERT based classification approach and embeddings integrated benchmark deep learning models are applied for sentiment and complaints classification tasks of two tier-1 cities of India (*Delhi* and *Bangalore*). The comparative evaluation is done based on the four popular performance metrics (*Accuracy, Precision, Recall* & *F-measure*). From the comparative evaluation on Delhi and Bangalore datasets, the following conclusions can be stated:

- For both Sentiment classification and Complaints classification tasks, BERT model outperformed other benchmark word-embeddings integrated deep learning models (*DNN, LSTM, Bi-LSTM* and *CNN*).
- The BERT based Complaints classification model could be reliably adopted for identifying and prioritizing the community level complaints.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdar, M., Basiri, M. E., Yin, J., Habibnezhad, M., Chi, G., Nemati, S., & Asadi, S. (2020). Energy choices in alaska: Mining people's perception and attitudes from geotagged tweets. *Renewable and Sustainable Energy Reviews*, *124*, Article 109781.

Agarwal, A., Toshniwal, D., & Bedi, J. (2020). Can twitter help to predict outcome of 2019 indian general election: A deep learning based study. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 38–53). Springer.

Alamoudi, E. S., & Alghamdi, N. S. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, 1–23.

Balogun, A.-L., Marks, D., Sharma, R., Shekhar, H., Balmes, C., Maheng, D., Arshad, A., & Salehi, P. (2020). Assessing the potentials of digitalization as a tool for climate change adaptation and sustainable development in urban centres. *Sustainable Cities and Society*, *53*, Article 101888.

Bedi, J. (2020). Attention based mechanism for load time series forecasting: An-lstm. In *International conference on artificial neural networks* (pp. 838–849). Springer.

Bedi, J. (2021). Transfer learning augmented enhanced memory network models for reference evapotranspiration estimation. *Knowledge-Based Systems*, Article 107717.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blümel, J. H., & Zaki, M. (2021). Comparative analysis of classical and deep learning-based natural language processing for prioritizing customer complaints.

Da Silva, N. F., Hruschka, E. R., & Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, 170–179.

Dabiri, S., & Heaslip, K. (2019). Developing a twitter-based traffic event detection model using deep learning architectures. *Expert Systems with Applications*, *118*, 425–439.

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.

Domeniconi, G., Moro, G., Pasolini, R., & Sartori, C. (2015). A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf. idf. In *International conference on data management technologies and applications* (pp. 39–58). Springer.

Feldman, R. (2013a). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*, 82–89.

Feldman, R. (2013b). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*, 82–89.

Fu, T., Wang, C., & Cheng, N. (2020). Deep-learning-based joint optimization of renewable energy storage and routing in vehicular energy network. *IEEE Internet of Things Journal*, *7*, 6229–6241.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, *69*, 214–224.

Goldberg, Y., & Levy, O. (2014). Word2vec explained: deriving mikolov, et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Huang, Q., Chen, R., Zheng, X., & Dong, Z. (2017). Deep sentiment representation based on cnn and lstm. In *2017 international conference on green informatics (icgi)* (pp. 30–33). IEEE.

Ikoro, V., Sharmina, M., Malik, K., & Batista-Navarro, R. (2018). Analyzing sentiments expressed on twitter by uk energy company consumers. In *2018 fifth international conference on social networks analysis, management and security (snams)* (pp. 95–98). IEEE.

Ilieva, R. T., & McPhearson, T. (2018). Social-media data for urban sustainability. *Nature Sustainability*, *1*, 553–565.

Iniyan, S., Suganthi, L., & Jagadeesan, T. (1998). Renewable energy planning for india in 21st century. *Renewable Energy*, *14*, 453–457.

Jain, A., & Jain, V. (2020). Renewable energy sources for clean environment: Opinion mining. *Asian Journal of Water, Environment and Pollution*, *16*, 9–14.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612. 03651.

Kim, S. Y., Ganesan, K., Dickens, P., & Panda, S. (2021). Public sentiment toward solar energy—opinion mining of twitter using a transformer-based language model. *Sustainability*, *13*, 2673.

Kim, N., & Hong, S. (2021). Automatic classification of citizen requests for transportation using deep learning: Case study from boston city. *Information Processing & Management*, *58*, Article 102410.

Lee, E. (2019). An intro to hyper-parameter optimization using grid search and random search. URL: https://medium.com/@cjl2fv/an-intro-to-hyper-parameter-optimization-using-grid-search-and-r{and}om-search-d.

Li, Y., & Shen, B. (2017). Research on sentiment analysis of microblogging based on lsa and tf-idf. In *2017 3rd ieee international conference on computer and communications (iccc)* (pp. 2584–2588). IEEE.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*, 1–167.

Liu, X., & Hu, W. (2019). Attention and sentiment of chinese public toward green buildings based on sina weibo. *Sustainable Cities and Society*, *44*, 550–558.

Liu, Z., & Na, J.-C. (2018). Aspect-based sentiment analysis of nuclear energy tweets with attentive deep neural network. In *International conference on asian digital libraries* (pp. 99–111). Springer.

Liu, Y., & Zhang, M. (2018). Neural network methods for natural language processing.

Majumdar, A., & Bose, I. (2019). Do tweets create value? a multi-period analysis of twitter use and content of tweets for manufacturing firms. *International Journal of Production Economics*, *216*, 1–11.

Mauri, A., Psyllidis, A., & Bozzon, A. (2018). Social smart meter: Identifying energy consumption behavior in user-generated content. In *Companion proceedings of the the web conference 2018* (pp. 195–198).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mogaji, E., Balakrishnan, J., & Kieu, T. A. (2020). Examining consumer behaviour in the uk energy sector through the sentimental and thematic analysis of tweets. *Journal of Consumer Behaviour*, 1–13.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). Sentiful: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, *2*, 22–36.

Nielsen, F. Å. (2011). *Afinn*: *Richard petersens plads, building*, (p. 321).

of Power, M. (2012). Government of india. URL: https://powermin.nic.in/en/content/energy-efficiency.

Onan, A. (2021). Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. *Computer Applications in Engineering Education*, *29*, 572–589.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Prabhu, . (2018). Understanding of convolutional neural network (cnn) — deep learning. URL: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148.

Raju, K. V., & Sridhar, M. (2020). Based sentiment prediction of rating using natural language processing sentence-level sentiment analysis with bag-of-words approach. In *First international conference on sustainable technologies for computational intelligence* (pp. 807–821). Springer.

Rodrigues, L., Gillott, M., Waldron, J., Cameron, L., Tubelo, R., Shipman, R., Ebbs, N., & Bradshaw-Smith, C. (2020). User engagement in community energy schemes: A case study at the trent basin in nottingham uk. *Sustainable Cities and Society*, *61*, Article 102187.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*, 140–157.

Salur, M. U., & Aydin, I. (2020). A novel hybrid deep learning model for sentiment classification. *IEEE Access*, *8*, 58080–58093.

Seo, S., Kim, C., Kim, H., Mo, K., & Kang, P. (2020). Comparative study of deep learning-based sentiment classification. *IEEE Access*, *8*, 6861–6875.

Sethi, P., Chakrabarti, D., & Bhattacharjee, S. (2020). Globalization, financial development and economic growth: perils on the environmental sustainability of an emerging economy. *Journal of Policy Modeling*.

Shahbaz, M., Mallick, H., Mahalik, M. K., & Sadorsky, P. (2016). The role of globalization on the recent evolution of energy demand in india: Implications for sustainable development. *Energy Economics*, *55*, 52–68.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, *404*, Article 132306.

Shi, T., & Liu, Z. (2014). Linking glove with word2vec. arXiv preprint arXiv:1411.5595.

Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, *471*, 216–232.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*, 267–307.

Tan, L., Yu, K., Shi, N., Yang, C., Wei, W., & Lu, H. (2021). Towards secure and privacy-preserving data sharing for covid-19 medical records: A blockchain-empowered approach. *IEEE Transactions on Network Science and Engineering*.

Toman, M. T., & Jemelkova, B. (2003). Energy and economic development: an assessment of the state of knowledge. *The Energy Journal*, 24.

Wang, M., Cao, D., Li, L., Li, S., & Ji, R. (2014b). Microblog sentiment analysis based on cross-media bag-of-words model. In *Proceedings of international conference on internet multimedia computing and service* (pp. 76–80).

Wang, G., Sun, J., Ma, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, *57*, 77–93.

Xu, J., & Du, Q. (2019). A deep investigation into fasttext. In *2019 ieee 21st international conference on high performance computing and communications; ieee 17th international conference on smart city; ieee 5th international conference on data science and systems (hpcc/smartcity/dss)* (pp. 1714–1719). IEEE.

Zhang, Z., He, Q., Gao, J., & Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C (Emerging Technologies)*, *86*, 580–596.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*, 43–52.