

SCANCPECLENS: A Framework for Automatic Lexicon Generation and Sentiment Analysis of Micro Blogging Data on China Pakistan Economic Corridor

BIBI AMINA^{1b} AND TAYYABA AZIM^{1b}

Center of Excellence in IT, Institute of Management Sciences, Peshawar 25000, Pakistan

Corresponding author: Tayyaba Azim (tayyaba.azim@imsiences.edu.pk)

ABSTRACT With the growing availability of internet and opinion rich resources such as social networks and personal blogs, the task of mining public opinion and exploring facts has become more popular than ever before during the last decade. The latest trend has deeply transformed the way the governments interact with their citizens and offer them various services through continuous public engagement. The proposed framework SCANCPECLENS is an initiative to support performance assessment framework for e-government in Pakistan. The research takes into account the opinion of masses on one of the most crucial and widely discussed development projects, China Pakistan Economic Corridor (CPEC), considered as a game changer due to its promise of bringing economic prosperity to the region. The proposed framework suggests to use machine learning algorithms to automatically discover the public sentiment from microblogs on the matter nationally as well as internationally. We also present an automated way to create sentiment lexicon of positive, negative and neutral words on the subject. To the best of our knowledge, this theme has not been explored for opinion mining before and helps one in effectively assessing public satisfaction over government's policies in the CPEC region. The research is an initiative to discover new avenues of future research and direction for the government, policy making institutions and other stake holders and demonstrates the power of text mining as an effective tool to extract business value from vast amount of social media data.

INDEX TERMS China Pakistan economic corridor (CPEC), k -nearest neighbor, logistic regression, lexicon generation, machine learning, natural language processing, sentiment analysis, support vector machines, text mining.

I. INTRODUCTION

The wide spread use of high speed *internet* and boom in the *smart phone industry* has regenerated the practices of opinion making, promotion, marketing and governance in the last couple of years. The two technological advancements have profoundly changed the way we communicate with one another and make decisions about items or products, ideas and governments' strategies. There are a huge number of social networking and blogging web sites that bear public opinion without going through any reviewing or short-listing procedures as is the tedious custom in print media, for example: Twitter, Facebook, WordPress, Joomla, etc. This

global trend of social media usage has expanded the scale of multimedia web content exponentially, thus presenting new challenges for the governments and decision making bodies worldwide. A substantial number of businesses and decision making agencies have now shifted their attention on mining helpful data from online networking to better comprehend their customers, fill the communication gap between their customers, recognise new sales opportunities and discuss new research areas for offering better services. This paper specifically focuses on microblogging data from Twitter to investigate the public sentiment on China Pakistan Economic Corridor (CPEC),¹ an ongoing mega development venture that has welcomed heated talks, discussions and analyses by

¹The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

¹<http://cpec.gov.pk/>

economic experts and politicians on its suitability nationally as well as globally. The project guarantees economic prosperity in the region by presenting new trade routes with Central Asia, Middle East and Africa. The improved geographical linkages with road, rail and air transportation systems are likely to increase investment and livelihood opportunities, might attract high volumes of trade flow and frequent exchange of people for academic, cultural and regional activities of growth and better understanding.

We here present a machine learning based framework, SCANCPECLENS that analyses public's views on the topic to help create support mechanism for implementing an open and transparent e-government. It is worth noting that the applications of web mining that empower e-democracy by improving political transparency and public participation in decision making via social media have not been explored and practised in Pakistan or CPEC region yet. There are a limited number of case studies discussed in Sec II that demonstrate the applications of social media mining for governance, yet no work has been done on the selected theme in CPEC countries. We explore this avenue of research by setting up a baseline model that is capable of categorising any tweet on the subject of CPEC into positive, negative or neutral class. In addition to the analysis of sentiment polarity, the framework also showcases the formation of a customised sentiment lexicon from Twitter feeds that contains a rich collection of positive, negative and neutral words for the task at hand. Unlike other benchmark dictionaries, this sentiment lexicon offers domain specific vocabulary that provides sufficient coverage of generic as well as specialised content discussed by the masses on social microblogging website Twitter. We believe that with the deployment of such intelligent software, the governments participating in the project would be able to increase the coverage and quality of information and services provided to the general public. In addition to these governments and public sector organisations, journalists and news agencies may also use such a system to identify the impact of governments' policies, identify areas of performance improvement in the project and assess elected government's claims of progress and prosperity against the facts reported by the United Nations (UN) and International Monetary Fund (IMF).

The novel contributions of this work are as follows:

- The SCANCPECLENS framework develops a novel open source repository of microblogs referred to as SCANCPEC (<https://github.com/tabzim/SCANCPEC>), holding a collection of tweets on China Pakistan Economic Corridor (CPEC) theme extracted from July 2013 to August 2017. A subset of these tweets from September 2016 to March 2017 is labelled manually with the help of several human annotators and the prescribed labels are finally accepted using an *inter-annotator agreement* (discussed ahead). To the best of our knowledge, there is no benchmark data set available on the subject of CPEC that offers examination of public sentiment/opinion. The maintained data set

offers the machine learning practitioners, data scientists, government policy makers and news agencies an opportunity to analyse the global population's sentiment on the ongoing mega development project.

- Development of a machine learning based sentiment analysis system for classifying public tweets on CPEC. To the best of our knowledge, there is no such study undertaken so far that deploys the power of machine learning models to intake public opinion on any democratic matter (in particular CPEC) for achieving political transparency, better public service and development in the region.
- Construction of a domain specific sentiment lexicon customised for CPEC developmental project. We have released sentiment specific word embeddings ([https://github.com/tabzim/SCANCPEC/tree/master/Sentiment%20Lexicon%20\(Sep2016-March2017\)](https://github.com/tabzim/SCANCPEC/tree/master/Sentiment%20Lexicon%20(Sep2016-March2017))) learnt from the tweets in corpus. The vocabulary of this lexicon could be adopted off-the shelf in other sentiment analysis, opinion mining and threat detection applications that use tweets for clues about major events or emerging events for public safety.

The remaining paper is organised as follows: Section II gives a brief review of the preliminaries required to understand this work and related research on text mining for benefiting e-government. Section III discusses the development and functionality of the proposed framework needed for microblogging content analysis. This is followed by Section IV that discusses the key results obtained from the conducted experiments. We draw analysis from the obtained quantitative facts and discuss the findings further in Section V. Section VI summarises this case study with directions for the future research and development.

II. PRELIMINARIES AND RELATED WORK

Social multimedia mining requires intelligent techniques to extract useful information from the web where data is available in the form of text, images, videos, meta-data, etc. Sentiment analysis of textual data usually requires natural language processing (NLP) and text analysis techniques to automate the extraction and classification of sentiments. Conventionally, the sentiments are analysed at three different levels [1] based on the granularity of information utilised for feature extraction: Document level, sentence level and aspect level. *Document level* sentiment analysis assumes that the entire document holds opinion about a single topic or entity. This type of sentiment analysis is widely researched, however it is not applicable to forum discussions, blogs, and news articles as such postings generally evaluate and discuss multiple entities. Also it does not describe what aspect of the entity is liked/disliked by the user. *Sentence level* sentiment classification goes farther than document level sentiment classification by assessing the opinion of each sentence in a document. This approach performs a fine-grained analysis yet it still holds the deficiency of aspect responsible for likeness/dislikeness. Another shortcoming of the approach is that it does not work

well for comparative sentences such as ‘Orange juice tastes better than Coke’, as the sentence talks about two entities and have two different opinions for both. Classifying text through both these discussed techniques do not provide opinion on all features or characteristics of the entity being discussed. This gap is filled by research on *aspect level* sentiment analysis [2] that classifies the sentiment with respect to specific characteristics of the entity under discussion. Such systems first identify the occurrence of entity and then their discussed aspects to mine opinion. For example, consider the phrase: “This phone has a good battery life but the voice quality is not good”. This sentence holds multiple opinions about different aspects of the same entity, i.e. phone. Our study focuses on deploying *sentence level* sentiment classification method to understand the sentiments of people on the CPEC development project.

For any type of sentiment classification, feature extraction and selection play a significant role for identifying relevant attributes that can increase the classification accuracy. Some of the most important feature extraction techniques used for natural language processing tasks are organised as follows in order of their popularity and usage: N-gram (unigram, bigram, trigram) [3], bag of words (BoW) [4], term frequency-inverse document frequency (TF-IDF) [5], parts of speech (POS) [6], word2vec [7]. Some of the widely used classifiers deploying these features are: Support vector machines (SVM), logistic regression, naive Bayesian, *k*-nearest neighbour, convolution neural networks, restricted Boltzmann machines, maximum entropy (ME) [8]–[12]. Among these, there is not a single machine learning technique that has outperformed the rest across all application domains. In order to improve feature representation, deep learning models are also adopted by researchers to learn higher level abstract features and improve sentiment classification results [13].

The sentiment classification systems deploy lexicons of sentiment words to identify sentiments in different contextual discussions [14]–[17]. The sentiment lexicons play a key role in determining the semantics of expressions in text. Generally, a sentiment dictionary of positive and negative words is created first and the ratio of the sentiment words determine the polarity of a tweet [18]. The most popular and commonly used general purpose sentiment lexicons are SentiWordNet, WordNet, and ConceptNet. These general purpose sentiment lexicons do not cater domain specific words and therefore incorrectly score the polarity of a word in a different context. To address this issue, one has to develop domain specific sentiment lexicons. The sentiment lexicons can be developed in three different ways: Manual, bootstrapping and corpus oriented. The manual technique needs human annotators for the selection and annotation of words and is therefore time consuming and costly. The sentiment lexicons are not built *manually* anymore, rather a lot of focus has been put on automatic generation of lexicons from any data set through *bootstrapping* [19] and *corpus-oriented* techniques [20]. Our proposed framework deploys corpus

oriented technique to develop domain specific sentiment lexicon on CPEC.

Examining social media content has turned into a very active and dynamic area of research and has deeply transformed the ways of traditional governance and commerce [21], [22]. Some of the relevant works that have deployed the idea of mining from social media in *governance* and *politics* are outlined below: The first official sentiment analysis system [23] was launched by the Singapore government in 2010 to analyse how citizens think about government’s policies. The software, developed by IBM, trawled social media for positive or negative key words to identify current and emerging trends in public sentiment. Ackland [24] used social media mining for exploring the size of the web graph and mapped political parties network on the web. Political web linking and sentiment analysis has also been explored for the US Congress elections [25], [26], Australian federal elections of 2010, where political candidates sentiment was analysed [27] and for the general elections of 2013 in Pakistan [28], where the analytical findings from machine learning techniques were shown to correspond with the results released by the Election Commission of Pakistan. Similar studies utilising the social media content and volume of Twitter data were conducted in Germany and United States to predict the election results [29], [30]. Another interesting case study carried out by Spiliotopoulou *et al.* [31] gathered public insight on the social impact of European Union’s financial crisis by exploiting topic models and stance classification framework. Using online posts, the authors determined citizens’ sentiment polarity for a particular political event, examined whether citizens’ sentiments agree with governmental decisions and predicted outcome of the political decisions made by citizens. Hasbullah *et al.* [32] also presented an automated content analysis tool which focused on helping Malaysian legal firms and official leaders to understand public sentiment for policy making and the future development. They applied semantic role labelling (SRL) techniques along with rule based model to generate new methods for filtering and classifying the comments on Malaysian government’s official pages of Twitter and Facebook. On similar lines, Reddick *et al.* [33] proposed a framework to provide organisational insight through social media text analytics to improve and enhance citizen-centred public service quality. This framework is applied to monitor online interactions of citizen and government on a local government’s Facebook page. Criado *et al.* [34] emphasised on the importance of using social media for the innovation of public sector organisations. The authors have proposed and developed three dimensions of research for using social media in e-government: Tools, goals and topics. The first dimension of *tools* discusses the social media channels governments use explicitly. The second dimension of *goals* refers to the policy and managerial objective behind the use of social media in public sector organisations. The third dimension of *topics* refers to the main areas where social media could be used by government for knowledge

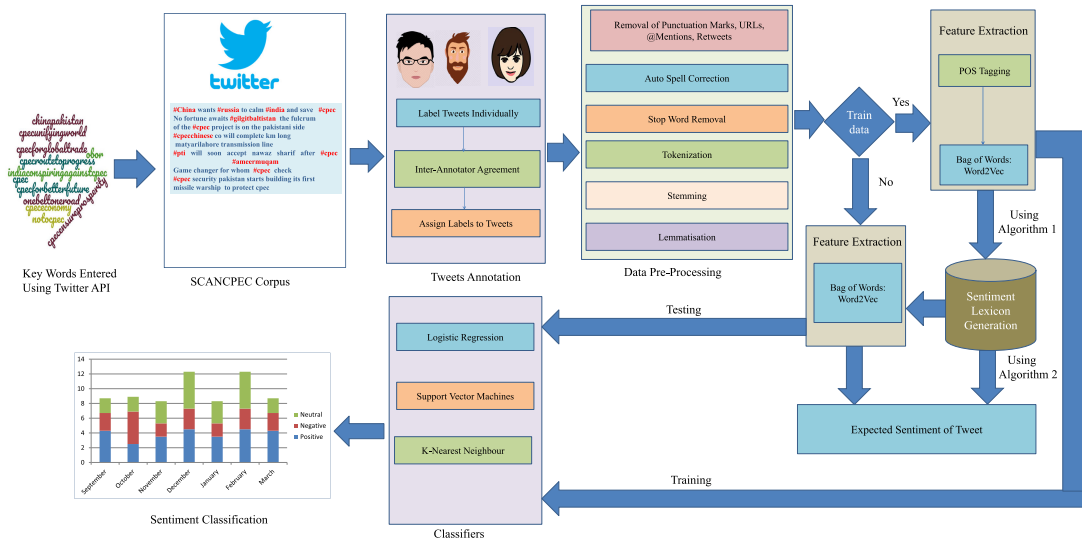


FIGURE 1. Graphical abstract of sentiment classification engine adapted for SCANCPEC corpus.

building. This study shows more promising directions for government’s learning through social media platforms to improve public service quality.

It appears that most of the research in e-governance has shown significance for the political parties and governments than for the general public. Web mining applications that empower democracy by bringing political transparency and participation in decision making process lack in practice, particularly in the developing countries. We aim to fill this gap by developing a sentiment classification system that mines citizens’ interest and viewpoint on CPEC and thus helps improving government’s standard of innovation and development in the region.

III. METHODOLOGY

This section describes the methodology used for developing SCANCPECLENS framework. The proposed sentiment analysis framework consists of the following modules: Domain Corpus Preparation, Corpus Annotation, Data Wrangling, Feature Extraction, Sentiment Lexicon Creation, Sentiment Lexicon Based Annotation and Classification. The graphical abstract of the proposed methodology is given in Figure 1.

A. DOMAIN CORPUS PREPARATION

The first step required for developing a sentiment analysis framework for CPEC is the collection of relevant microblogging data set [35]. We chose to use Twitter platform for sentiment analysis as we realised that it offers a better approximation of public sentiment rather than traditional internet articles and web blogs that offer less data due to their online postings once a day. The public use of Twitter is far more quick and furthermore more broad/common (since the amount of users’ tweets is considerably more

than the individuals who compose web blogs once a day.) We believe that due to Twitter’s widespread usage, we can accomplish an accurate impression of public sentiment by analysing the expressions communicated in the tweets.

In order to gather tweets on the subject of China Pakistan Economic Corridor (CPEC), we have utilised Twitter’s application programming interface (API) destined for the programmers to legitimately access user tweets in JavaScript Object Notation (JSON) format. The open source scraper utilised for the task could be downloaded from here: <https://github.com/taspinar/twitterscraper>, and executed in Python for crawling relevant tweets. The keywords utilised for querying and shortlisting the pool of relevant tweets include popular contextual terms such as #cpec, #chinapakistan, #obor, #onebeltoneroad, etc. The procedure yielded around 222,625 tweets altogether from July 2013 to August 2017. These tweets are saved online in an open source repository named as SCANCPEC Corpus. In order to comply with Twitter’s terms of service, only stripped down version of this raw data with limited fields (time, date, username, tweet text) is shared with the readers at <https://github.com/tabzim/SCANCPEC>. The published data set consists of online discussions demonstrating the expression of individuals on various aspects of the mega developmental project. A list of 25 most frequently occurring keywords in this subset are shown via histogram given in Figure 2.

B. CORPUS ANNOTATION

The tweets gathered in SCANCPEC corpus are unannotated and lend themselves to annotation so that an appropriate ground truth is established for sentiment analysis and opinion mining. Keeping in view the complexity of

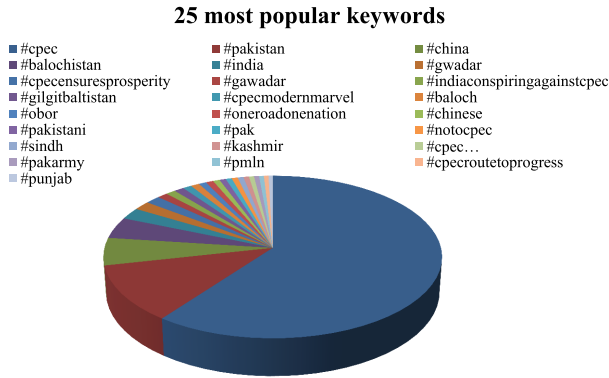


FIGURE 2. List of 25 most frequently occurring keywords in the tweets dating from September 2016-March 2017.

manual annotation task, we’ve manually labelled only a subset of 11,933 tweets from September 2016 to March 2017 for developing this case study. The tweets are manually labelled into positive (*label=1*), neutral (*label=2*) and negative (*label=3*) categories with the help of three human annotators who performed the same annotation task in parallel without any discussion/collaboration during the entire process. This setting ensured that every set of annotations is the work of a single annotator without any influence and any differences in the opinion are recorded too. The label of each tweet is the label agreed by at least two annotators. Any tweet that does not satisfy this confidence measure is excluded from participation in the system.

Manually annotating web data with the help of human experts has been in practice for developing various benchmark data sets [36]–[38], however it is more expensive and takes more time. The annotation task could also be carried out on Amazon Mechanical Turk³ or Crowd Flower,⁴ however due to the payment constraints for the region, we could not avail the service and had to set it up offline using multiple human annotators. The annotation was performed at sentence level and the prescribed labels were accepted using *inter-annotator agreement*. Inter-annotator agreement is a measure of how well two (or more) annotators make the same annotation decision for a certain label in the entire corpus [39]. We measured the inter-annotation agreement of the three annotators using Cohen’s kappa coefficient [40] and found it substantial ($\text{kappa} = 0.701$) for further analysis and improvement in our case study. Generally, a kappa statistic greater than 0.7 shows a good agreement between user annotations based on the equations presented in [41]. We have shared the file measuring the kappa statistic calculation of our pre-processed and annotated 7203 tweets in the open source repository <https://github.com/tabzim/SCANCPEC/blob/master/kappa-cpec3Raters-final.xlsx>. Any conflicting tweets for

TABLE 1. Ground truth distribution of tweets in SCANCPEC corpus. These tweets were manually labelled by multiple human annotators and the annotations were accepted via inter-annotator agreement.

Annotation Schemes	#Positives	#Negatives	#Neutrals	Total
Manual Annotation	2333	2549	2321	7203

which the annotators have disagreed to come up with the same label are not utilised for training/test purpose. Table 1 shows the ground truth distribution of 7203 tweets among the three sentiment classes: Positive, negative and neutral, when the tweets were annotated manually and accepted through inter-annotator agreement.

C. DATA WRANGLING

1) STOP WORDS REMOVAL, TOKENISATION AND STEMMING

Performing natural language processing on textual data from Twitter offers new challenges due to the informal use of language in tweets. In order to remove irrelevant information that does not hold any significance for sentiment classification and data analytics, we have pre-processed the collected microblogs using Natural Language Toolkit (NLTK)² in Python [42]. The preprocessing involves the removal of punctuation marks, uniform resource locators (URLs), at mentions (where people tag other users in tweets) and retweets. All the tweets were converted to lower-case letters and words that were not informative semantically and had dual meanings, such as nail, pool, mine, current, etc. were eradicated. We next checked the tweets for miss-spellings and auto corrected the spelling mistakes in tweets expressed in English language. The open source code utilised for the auto-spell correction task is available here: <https://github.com/phatpiglet/autocorrect>. The code uses a probabilistic approach to determine the correct spelling of used word in a tweet, however some of the Urdu/Hindi words are transliterated as English and hence cannot be corrected. To deal with the issue of transliteration, we eradicated the tweets, 70% of whose content is misspelled. This step removed all the transliterated tweets as well as those not expressed in English Roman script (Urdu, Arabic and Chinese scripts). We have then removed the stop words from the tweets. Stop words are those commonly used short functional words that do not prove very useful when retrieving items by a search query. Examples of such words include: ‘the’, ‘is’, ‘at’, ‘on’, etc. In order to make the annotated data set suitable for language modelling, lexicon analysis and sentiment examination, we have exercised the following steps in sequence next: 1) Tokenisation and 2) Stemming. Tokenisation allows one to break the tweet (string) into word tokens, whereas the stemming technique reduces the inflectional or derived forms of the word by extracting its root/base word.

³<https://www.mturk.com/mturk/welcome>

⁴www.figure-eight.com

²<http://www.nltk.org/>

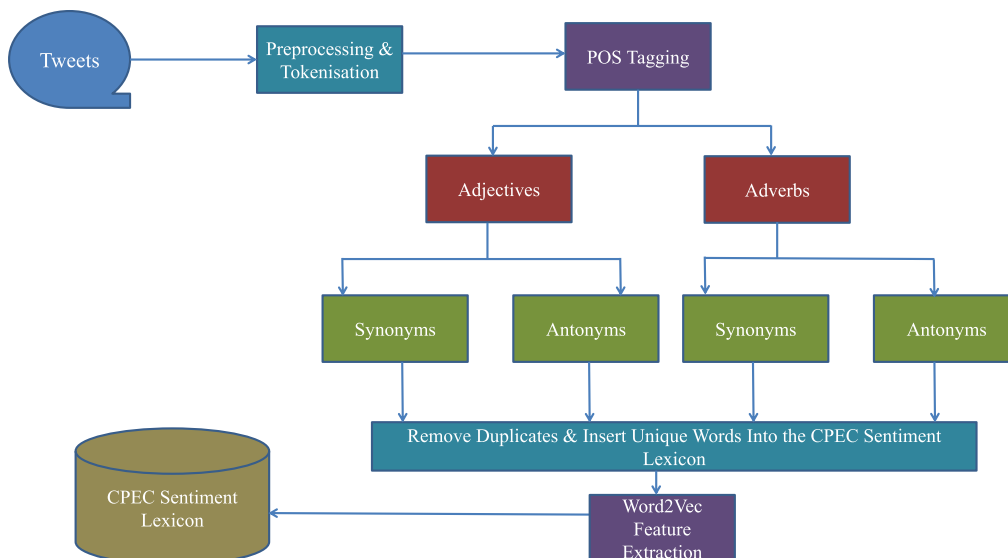


FIGURE 3. Schema depicting feature extraction steps as well as the procedure of automatic lexicon generation for CPEC sentiment corpus, SL3.

D. FEATURE EXTRACTION

We have extracted parts of speech (POS) tags from the set of 7203 annotated and pre-processed tweets. POS tagging processes each word token and attaches a part of speech tag to each word using the NLTK library. In English language, there are nine parts of speech: Noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. We have tagged each word token as an adverb or adjective as they are more likely to convey information related to public sentiments.

The second popular NLP feature that was extracted while building lexicon repositories and comparing the word similarities with lexicons is word2Vec feature. Word2Vec transforms each word token into a dense vector by deploying a shallow two-layered neural network. Word vectors are positioned in the vector space such that the words that share common contexts in the corpus are located in close proximity to one another. Once trained, the neural network is able to reconstruct linguistic contexts of words that can help one in determining the sentiment polarity. Word2Vec is a computationally efficient predictive model for learning word embedding from raw text. This is specifically useful when the text corpus is very large, however this does not reduce the utility of the proposed feature for small or medium sized text corpora and it is still deemed as a better feature in comparison to the traditional natural language processing features such latent semantic analysis (LSA). Word2Vec leverages two different architectures to draw word embedding: (1) Continuous bag of words (CBOW) and (2) skip-grams. The CBOW model tries to predict the current target word (the center word) based on the source context words (surrounding words), whereas skip-gram model tries to predict a whole bunch of context words from a source target word. The literature suggests using

continuous bag of words (CBOW) for small/medium sized text corpora and skip-gram architecture for large text corpora. This is because unlike skip-grams, CBOW smooth over a lot of distributional statistics by averaging over all context words. With small or medium sized data, this regularizing effect of CBOW turns out to be helpful. Keeping in view the size of our annotated data set, i.e. 7203 tweets, we have deployed the continuous bag of words (CBOW) model as shown in Figures 1 and 3. We have used Matlab’s text analytics toolbox (<https://www.mathworks.com/products/text-analytics.html>) for extracting accurate word embeddings. The raw data and extracted features from the tweets have been shared in the repository (<https://github.com/tabzim/SCANCPEC>) for quick development of applications aiming for social media analytics.

E. SENTIMENT LEXICON CREATION

Sentiment lexicon is a database of lexical units (word tokens or phrases) having a specific sentiment orientation. The richer the lexicon vocabulary, the more appropriate would be the result of opinion mining or sentiment analysis of natural language documents. Lexicon based sentiment analysis approaches either utilise the existing natural language *dictionaries* to build sentiment lexicon or utilise some domain specific *corpus* for the purpose. The idea is to develop quality vocabulary repository that can assist in identifying the polarity of public opinion demonstrated in microblogs or documents. We have utilised both the resources, general purpose dictionary as well as the SCANCPEC corpus, in our quest to develop a suitable sentiment lexicon for the task at hand.

We next explain how the sentiment lexicons were created for this work: The first sentiment lexicon was created by

Algorithm 1 Developing SCANCPEC Sentiment Lexicon SL3**Input:** Tweets, Libraries, SL2 Corpus.**Output:** SCANPEC sentiment lexicon SL3.

- 1: **for** All Tweets **do**
- 2: Read the tweet from SCANCPEC corpus.
- 3: Wrangle the tweet's text by removing punctuation marks, auto-correcting spelling mistakes, removing stop-words and changing all the words into lower-case.
- 4: Break the wrangled tweet into word tokens.
- 5: Find parts of speech (POS) tags (adjectives, adverbs) for each word token in the tweet.
- 6: Find synonyms and antonyms of each of these POS tagged word tokens.
- 7: Remove all the duplicate words in retrieved synonyms and antonyms.
- 8: Check the presence of synonyms and antonyms in the SL2 repository and retrieve their class labels.
- 9: Given the class labels of synonyms and antonyms, place the word tokens, synonyms and antonyms into positive, negative and neutral classes.

manually selecting positive, negative and neutral keywords from the set of annotated tweets that belonged to the train set of SCANCPEC corpus. We refer to this sentiment lexicon as SL1. The second sentiment lexicon was created by merging *Opinion Online Lexicon* having a generic list of positive and negative English words with SL1 sentiment lexicon to create a richer set of sentiment vocabulary. We refer to this hybrid sentiment lexicon as SL2. The third sentiment lexicon was created by extending the SL2 sentiment lexicon, following the steps demonstrated in Algorithm 1. This sentiment lexicon is referred to as SCANCPEC Sentiment Lexicon/ SL3. This sentiment lexicon is not just optimised for assessing the sentiment polarity of CPEC related tweets but can also help one take a quick start to mine public opinion on the subject using domain specific dictionary of positive, negative and neutral words. The scheme for creating SCANCPEC Sentiment Lexicon/SL3 is also shown graphically in Figure 3. The words in all the created sentiment lexicons are stored as vectors (word2vec features) computed through Matlab's text analytics toolbox.²

F. SENTIMENT LEXICON BASED CORPUS ANNOTATION

The sentiment lexicons created in Section III-E could also be utilised for automatically generating ground truth for tweets in SCANCPEC corpus or any other corpus from a similar domain. Lexicon based annotation approaches count and weight the sentiment words evaluated and tagged before in a given tweet and assign each message a label accordingly. To accomplish this task, each

²<https://www.mathworks.com/help/textanalytics/ref/fasttextwordembedding.html>

Algorithm 2 Sentiment Lexicon Based Corpus Annotation**Input:** *Tweet***Output:** *TweetSentiment*

- 1: **for** (Every *Tweet*) **do**
- 2: Wrangle the text in *Tweet*.
- 3: Calculate word2Vec feature of each word token in the preprocessed *Tweet*.
- 4: Find feature similarity of the word tokens with the features stored in sentiment lexicon and assign them sentiment labels accordingly.
- 5: Count all the positive, negative and neutral words in the *Tweet*.
- 6: Calculate *Target* using formula:
 $Target = \text{Sum of Positive Words} - \text{Sum of Negative Words}$.
- 7: **if** ($Target < 0$) **then**
- 8: *TweetSentiment* = 'Negative';
- 9: **if** ($Target > 0$) **then**
- 10: *TweetSentiment* = 'Positive';
- 11: **if** ($Target = 0$) **then**
- 12: *TweetSentiment* = 'Neutral';

preprocessed tweet as well as the saved lexicons are converted into vectors first with the help of *word2vec* model (<https://code.google.com/p/word2vec/>). The word2vec feature of each word token is compared with word2vec features of sentiment lexicon and labels are assigned to each word token accordingly. Given these labels, each tweet is assessed for the total number of positive N_{pos} and negative words N_{neg} . We then calculate the difference between the two counts, i.e. $N_{pos} - N_{neg}$. If this difference is greater than zero, the tweet is labelled as positive, if it turns out to be less than zero, the tweet is classified as negative and if the number of positive and negative words are equal in a tweet, i.e. the difference is zero, the tweet is annotated as neutral. The steps of the proposed annotation algorithm are described in Algorithm 2. We have used each of the three sentiment lexicons (SL1, SL2 and SL3) to generate annotations (ground truth) for classification. The quality of the annotations generated by each of the sentiment lexicons will be discussed in experiments section ahead.

G. CLASSIFICATION IN THE SCANCPELENS FRAMEWORK

The SCANCPELENS framework allows one to assess the polarity of opinion revealed in microblogs in a supervised way. For developing this framework, we select 7203 annotated and preprocessed tweets from the SCANCPEC corpus and divide them into train and test sets via 10-fold cross validation. The k -fold validation scheme splits the data set into 80:20 ratio, where 80% of the examples are used for model training and the remaining 20% are reserved for model testing. Thus, from a collection of 7203 tweets, 5762 tweets are allocated to the train set and 1441 tweets are assigned to the test set respectively. Before confining to this train-test

ratio, we also checked 70:30 and 60:40 train-test splits, however the validation results did not favour our choice of data division. A split of 80:20 allows the model to learn better from more training data and hence is more prepared to analyse sentiments of unseen microblogs. The train set is further used to achieve three objectives: (1) Derive a validation set required to optimise model parameters, (2) Develop sentiment lexicons as suggested in Algorithm 1, (3) Use sentiment lexicons developed from the train set to annotate the SCANCPEC corpus using Algorithm 2.

Using Matlab's text analytics toolbox, we extract word2vec features of both train and test tweets. The three state of the art classifiers utilised for sentiment analysis are: 1) k -nearest neighbor (k -NN) [43]–[46], 2) logistic regression [47] and 3) support vector machines (SVM) [3], [44], [46], [48], and [49]. Although simple, k -nearest neighbor (k -NN), logistic regression and support vector machines (SVM,) are still very popular classifiers used in the industry and academia for benchmarking the performance of different models and features. Once we have these baseline results, we can always seek for more fancy methods to improve the accuracy and robustness of the classification framework. The three classifiers are passed these feature vectors from the train set along with their actual labels for training purpose. The actual annotations are the ones returned by Algorithm 2 using each sentiment lexicon (SL1, SL2 and SL3) separately. Once these classifiers are trained with optimal parameters, word2vec features from the test examples are passed to the classifiers to find out the labels of the test data. These predicted labels are compared with the ground truth, i.e actual labels returned by sentiment lexicon based annotation to measure the classification performance of the proposed classifiers. The graphical abstract of the proposed supervised classification framework is shown in Figure 1.

H. PERFORMANCE EVALUATION METRICS UTILISED

The performance of the proposed sentiment classification framework is assessed with the help of following standard evaluation metrics: a) Accuracy, b) Precision and c) Recall.

Accuracy measures how good a model is, by estimating the proportion of all the predictions that are correct, i.e.

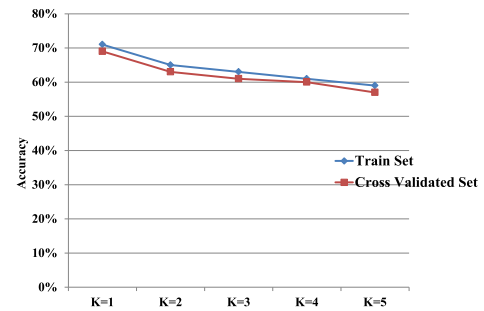
$$\text{Accuracy} = \text{Correct Predictions} / \text{All Predictions.} \quad (1)$$

When the data is not equi-balanced between classes, precision and recall metrics are used in addition to accuracy to estimate the efficiency of the classifiers.

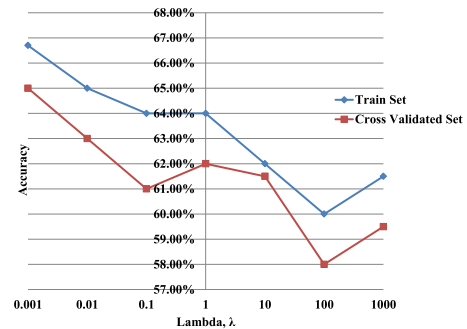
Precision calculates proportion of all the positive predictions that are correct. It measures the number of positive predictions that were actually positive observations.

Precision
 = Positives Predicted Correctly / All Positive Predictions. (2)

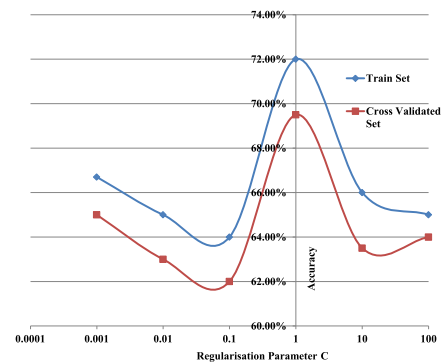
Recall is the true positive rate and measures the number of actual positive observations predicted correctly.



(a) Seeking the best value of k in k -nearest neighbour classifier.



(b) Searching for optimal regularisation parameter, λ in logistic regression. The most optimal performance was observed at $\lambda=0.001$.



(c) Searching for optimal regularisation parameter C in support vector machines classifier using linear kernel. The model was found to perform well at $C=1$.

FIGURE 4. Tuning the regularisation parameters of classifiers used for determining the sentiment polarity of masses on CPEC theme.

Recall

$$= \text{Positives Predicted Correctly} / \text{All Positive Observations.} \quad (3)$$

I. COMPUTATIONAL REQUIREMENTS OF THE PROPOSED FRAMEWORK

The tools used for developing the framework are Python 3.5 (Anaconda framework), NLTK library for text preprocessing, Scikit library for utilising the machine learning models and Matlab 2018 for the simulation of results. The source code of

TABLE 2. Accuracy of classifiers using word2Vec features and three different sentiment lexicons used for annotation.

Classifiers	Sentiment Lexicon (SL1)	Hybrid Sentiment Lexicon (SL2)	CPEC Sentiment Lexicon (SL3)
k -NN ($k=1$)	73.35 ± 0.35	71 ± 0.34	81 ± 0.28
Regularized Logistic Regression ($C=0.001$)	67.2 ± 0.25	78.8 ± 0.34	86.18 ± 0.28
SVM Linear Kernel ($C=1$)	67.5 ± 0.32	76.4 ± 0.05	86.8 ± 0.24
SVM RBF Kernel ($C=10, \gamma=0.01$)	70.2 ± 0.35	79 ± 0.04	86.6 ± 0.34

TABLE 3. Precision of classifiers using word2Vec features and three different sentiment lexicons for ground truth acquisition.

Classifiers	Sentiment Lexicon (SL1)	Hybrid Sentiment Lexicon (SL2)	CPEC Sentiment Lexicon (SL3)
k -NN ($k=1$)	73 ± 0.06	80 ± 0.03	72 ± 0.306
Regularized Logistic Regression ($C=0.001$)	66 ± 0.32	81 ± 0.07	87 ± 0.45
SVM Linear Kernel ($C=1$)	68 ± 0.02	76 ± 0.33	88 ± 0.45
SVM RBF Kernel ($C=10, \gamma=0.01$)	70 ± 0.34	81 ± 0.05	88 ± 0.06

TABLE 4. Recall rate of classifiers using word2Vec features and three different sentiment lexicons used for annotation.

Classifiers	Sentiment Lexicon (SL1)	Hybrid Sentiment Lexicon (SL2)	CPEC Sentiment Lexicon (SL3)
k -NN ($k=1$)	73 ± 0.43	67 ± 0.02	73 ± 0.06
Regularized Logistic Regression ($C=0.001$)	67 ± 0.34	78 ± 0.05	84 ± 0.06
SVM Linear Kernel ($C=1$)	67 ± 0.45	74 ± 0.23	84 ± 0.04
SVM RBF Kernel ($C=10, \gamma=0.01$)	70 ± 0.05	78 ± 0.42	84 ± 0.05

the proposed system was run on Intel Core i5 with 2.40 GHz processor and 4 GB RAM.

IV. EXPERIMENTAL RESULTS

This section describes how the classifiers deployed in SCANCPELENS (k -NN, SVM and logistic regression) were regularised to achieve optimal results on the test set. For regularising the parameters used in the deployed classifiers, a cross validation set is obtained first by splitting the train data set using 80:20 rule again. Seeking optimal parameter configuration regularises the models so that overfitting is avoided on the test data. For k -nearest neighbour, we have assessed the performance of the classifier with different values of k , i.e. 1, 2, 3, 4 and 5 using Euclidean distance metric and found out the best results on the validation set at $k = 1$. Figure 4(a) shows that the accuracy of k -NN begins to drop as the value of k is increased. Figures 4(a), 4(b) and 4(c) demonstrate the classification performance of k -NN, logistic regression and SVM classifiers on train and cross validation sets when the regularisation parameters are tweaked using the grid search method. Once optimal model parameters for each classifier are figured out by observing their performance on the validation set, the entire train set of 5762 tweets is utilised for model training and its performance is assessed on the test set containing 1441 tweets.

Tables 2, 3 and 4 outline the classification accuracy, precision and recall rates achieved by each of the machine learning classifiers on SCANCPEC data set. The predicted results are compared with the actual ground truth retrieved from sentiment lexicons (SL1, SL2 and SL3) using Algorithm 2. The classification results reveal that none of the classifiers outperforms the rest on all three sentiment lexicons using word2Vec features. On SL1 sentiment lexicon, k -NN shows the best classification performance, SVM with RBF kernel

demonstrates the best results using hybrid sentiment lexicon SL2, whereas SVM with linear kernel outperforms the rest using CPEC sentiment lexicon SL3. On the whole, the best classification results on the problem are yielded by SVM with linear kernel. Among all the sentiment lexicons, the vocabulary provided by CPEC sentiment lexicon (SL3) is more domain specific and therefore yields better ground truth labels required for training and testing the classifiers. We maintain that this vocabulary could be utilised off-the-shelf in other sentiment analysis, opinion mining and threat detection applications that use tweets for clues about major events or emerging events for public safety.

V. DISCUSSION AND ANALYSIS

A. ANALYSIS ON USER PARTICIPATION

We have also explored the geographical location of Twitter population to analyse the spatial patterns of public sentiment on CPEC worldwide. The geographical meta data is gathered with the help of Twitter API provided by Twitter officially (<https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.html>). Figure 5 reveals that the highest number of tweets came from Pakistan as compared to the other neighbouring countries, in specific China which is the biggest partner countries in the project. We believe that one of the main reasons for getting low participation of Chinese population in the microblogging data set is the censorship of Twitter in China. The ban of the social website in country has hampered the goal of gaining access to the Chinese population living in China and know their views in English. Other micro blogging websites such as Weibo, have taken the place of Twitter instead which allow users to express their views in the native language. Since this research was confined to extract opinion from text in English only, we could not include Chinese alternative to

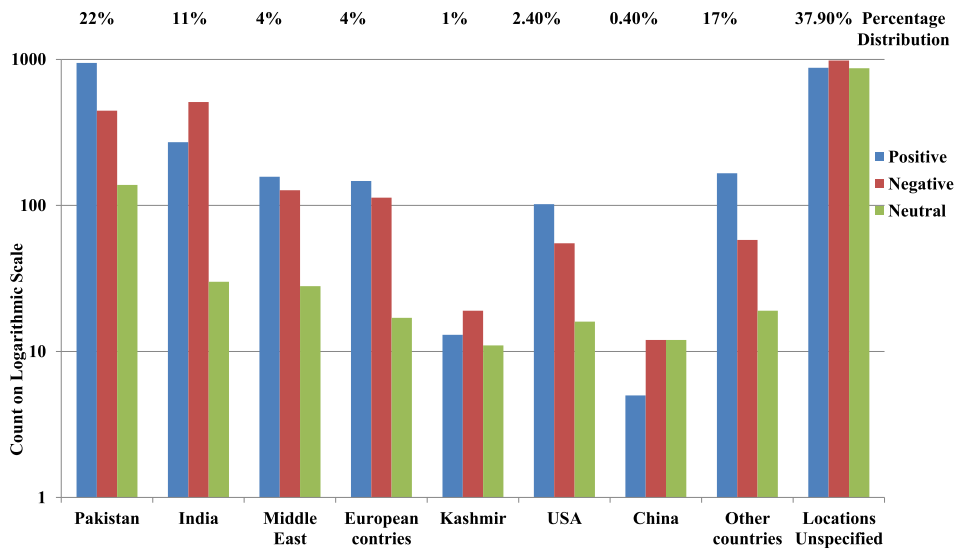


FIGURE 5. Region-wise sentiment statistics on CPEC, the percentage at top of each region is revealing the participation on twitter particularly on CPEC from that region we have 22% participation from Pakistan followed by 11% India, 4% Middle east and other regions likewise, 37.9% participation is from population whose location information cannot be tracked down from the meta data.

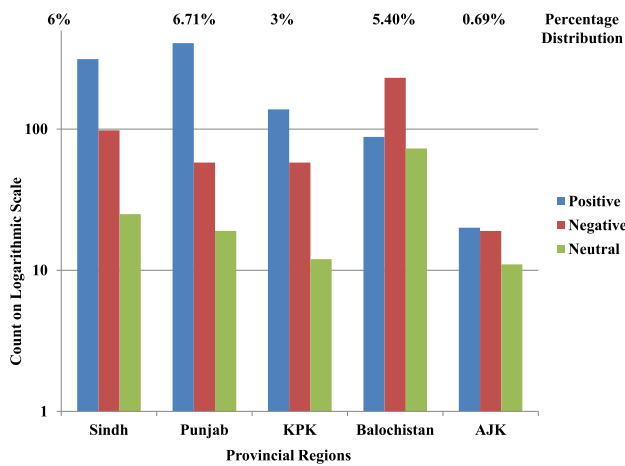


FIGURE 6. Estimation of tweets sentiment polarity at provincial level in Pakistan and Azad Jamun Kashmir (AJK).

Twitter, i.e. Weibo. However, we note that the Chinese population living outside China were able to contribute to this study. The graph reveals that the highest number of positive tweets came from Pakistan and the highest number of negative tweets came from India.

We further analyse the geo-spatial patterns of microblogs received from within Pakistan and compare its sentiment polarity with the number of development projects³ planned by the government for the entire region. A comparison of Figure 6 and Figure 7 reveals that the polarity of sentiments is not positively correlated to the count of development projects aimed by the government. We observe that the highest number of negative tweets came from the provincial region of

³<http://cpec.gov.pk/>

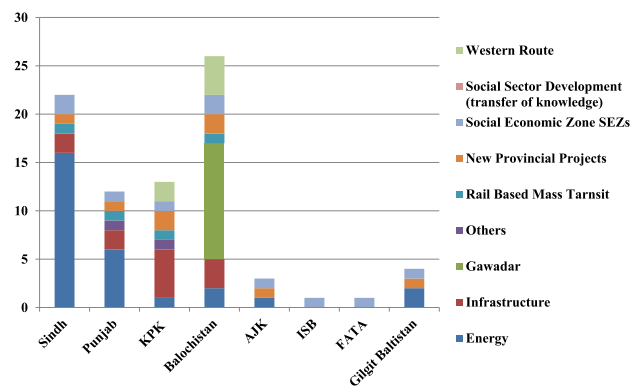


FIGURE 7. Different development projects planned in CPEC for each province of Pakistan and Azad Jamun Kashmir (AJK).

Balochistan, despite the government’s plan to execute largest number of development projects for the region. There could be several possible reasons for this disparity, some of which are highlighted below: 1) *Digital Divide*: Not everyone can afford the cost of computers and internet, thus online opinions are often diverted in favour of certain social groups who are literate and are willing to express their opinions online. Balochistan is ranked as the poorest province of Pakistan according to UNDP statistics, thus increasing the probability of having reduced public representation on Twitter, 2) *Democratic Experience*: The public on Twitter may not sound positive due to their past disadvantageous experience with the ruling governments who have not fulfilled their earlier promises of development, 3) *Disproportional Investment in the Country*: The CPEC investment benefits the Eastern part of the country more which includes the areas of Eastern

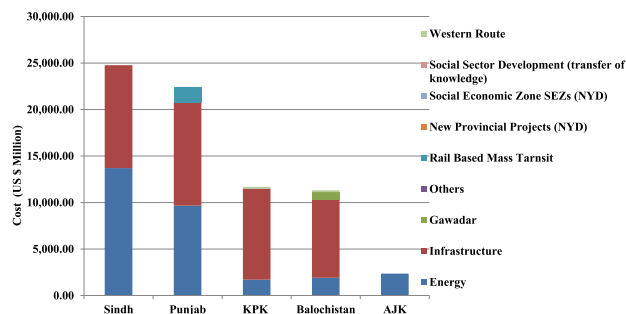


FIGURE 8. Distribution of investment in various development projects in CPEC.

Punjab and Southern Sindh, where the per-capita income and literacy rates are higher, standards of education and health are better and most of Pakistan's industry are located. See Figure 8 for illustration. This is also one of the reasons why the largest number of positive tweets came from these two provinces, whereas the neglecting region, i.e. Balochistan registered its protest through negative tweets. The sentiment polarity results and analysis shown here reflects public's faith on government's policies and highlights the need to take public into confidence for running mega development projects like CPEC. It is the government's responsibility to maximise public outreach and enhance their satisfaction through their planned projects and policies.

VI. CONCLUSION AND FUTURE WORK

This research deploys machine learning algorithms on microblogs to automatically discover public sentiment on the subject of China Pakistan Economic Corridor (CPEC) nationally as well as internationally. To the best of our knowledge, there is no such study undertaken so far that deploys the power of machine learning models to intake public opinion and sentiment on any democratic matter (in particular CPEC) for achieving political transparency and development in Pakistan. In order to pursue this research, we have constructed a domain specific repository of public tweets namely SCANCPEC, learned a domain specific sentiment lexicon and proposed an automatic framework for sentiment lexicon generation and sentiment classification. The open source annotated corpus and sentiment lexicon can be adopted off the shelf by any sentiment analysis application for further exploration and development.

In future, we aim to bring the opinion of disconnected groups into consideration by modelling data from other forums such as online newspapers (with a printing press), blogs and other social platforms like Facebook, Instagram and Wordpress. In order to analyse the aspects responsible for positive and negative project feedback on social media websites, we also aim to integrate topic models within the framework. Currently, the CPEC sentiment lexicon repository serves classification of explicit opinions with a focus on sentiment polarity rather than subjectivity detection or opinion detection where challenges such as sarcasm, anaphora resolution and double negation are more prevalent. Since sentiment

analysis frameworks are highly domain dependent and rely on rich lexical vocabulary to detect sentiments, we need to scale up the developed CPEC sentiment lexicon with more contextual words to tackle the mentioned challenges in future. We assert that such sentiment analysis systems should be deployed by the governments to devise policies for sensitive national issues and high budgeted projects like CPEC. It is difficult to maintain social harmony and justice if public's interest and opinion is ignored by the governments. Incorporation of such opinionated systems in e-government can break the barriers and bring government's service to all citizens with transparency. To the best of our knowledge, the mentioned subject (CPEC) despite its significance and high budget has not been researched for opinion mining before and will help one discover new avenues of future research and direction for the government, policy making institutions and other stake holders.

ACKNOWLEDGMENT

The authors would like to thank all the annotators who participated in labelling the tweets in corpus. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

- [1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
- [2] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2016.
- [3] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016.
- [4] G. Salton and J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1986.
- [5] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion mining and sentiment polarity on Twitter and correlation between events and sentiment," in *Proc. IEEE 2nd Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar./Apr. 2016, pp. 52–57.
- [6] R. Ghosh, K. Ravi, and V. Ravi, "A novel deep learning architecture for sentiment classification," in *Proc. 3rd Int. Conf. Recent Adv. Inf. Technol. (RAIT)*, Mar. 2016, pp. 511–516.
- [7] F. Enriquez, J. A. Troyano, and T. López-Solaz, "An approach to the use of word embeddings in an opinion classification task," *Expert Syst. Appl.*, vol. 66, pp. 1–6, Dec. 2016.
- [8] S. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," 2013, *arXiv:1308.6242*. [Online]. Available: <https://arxiv.org/abs/1308.6242>
- [9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, vol. 10, 2002, pp. 79–86.
- [10] S. Tan and J. Zhang, "An empirical study of sentiment analysis for Chinese documents," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2622–2629, 2008.
- [11] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for Twitter sentiment classification," in *Proc. SemEval COLING*, 2014, pp. 208–212.
- [12] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford Univ., Stanford, CA, USA, Project Rep. CS224N, 2009, p. 12, vol. 1.
- [13] X. Li, J. Cao, and Z. Pan, "Market impact analysis via deep learned architectures," *Neural Comput. Appl.*, pp. 1–12, Mar. 2018.
- [14] Z. Hailong, G. Wenyang, and J. Bo, "Machine learning and lexicon based methods for sentiment classification: A survey," in *Proc. IEEE Conf. Web Inf. Syst. Appl.*, Sep. 2014, pp. 262–265.
- [15] G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu, and C. Chen, "DASA: Dissatisfaction-oriented advertising based on sentiment analysis," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6182–6191, 2010.

- [16] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. 2nd Int. Conf. Knowl. Capture*, 2003, pp. 70–77.
- [17] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 168–177.
- [18] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 122–129.
- [19] Y. Choi, Y. Kim, and S.-H. Myaeng, "Domain-specific sentiment analysis using contextual feature generation," in *Proc. 1st Int. CIKM Workshop Topic-Sentiment Anal. Mass Opinion*, 2009, pp. 37–44.
- [20] M. Z. Asghar, S. Ahmad, M. Qasim, S. R. Zahra, and F. M. Kundi, "Senti-health: Creating health-related sentiment lexicon using hybrid approach," *SpringerPlus*, vol. 5, no. 1, p. 1139, 2016.
- [21] L. Tang and Z. Ni, "Emerging opinion leaders in crowd unfollow crisis: A case study of mobile brands in Twitter," *Pattern Anal. Appl.*, vol. 19, no. 3, pp. 731–743, 2016.
- [22] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *Int. J. Inf. Manage.*, vol. 33, no. 3, pp. 464–472, 2013.
- [23] R. Hicks, "Singapore to mine citizen sentiment online," Government Anal., Singapore Government, Singapore, Tech. Rep., 2010. [Online]. Available: <http://www.unpan.org/PublicAdministrationNews/tabid/115/mctl/ArticleView/ModuleID/1467/articleId/23532/Default.aspx>
- [24] R. Ackland, "Mapping the U.S. political blogosphere: Are conservative bloggers more prominent?" in *Proc. Conf. BlogTalk Downunder*, Sydney, NSW, Australia, 2005, pp. 178–185.
- [25] K. A. Foot and S. M. Schneider, "Online action in campaign 2000: An exploratory analysis of the U.S. Political Web sphere," *J. Broadcast. Electron Media*, vol. 46, no. 2, pp. 222–244, 2002.
- [26] K. Foot, S. M. Schneider, M. Dougherty, M. Xenos, and E. Larsen, "Analyzing linking practices: Candidate sites in the 2002 US electoral Web sphere," *J. Comput.-Mediated Commun.*, vol. 8, no. 4, p. JCMC84, 2003.
- [27] X. Zhou, X. Tao, J. Yong, and Z. Yang, "Sentiment analysis on tweets for social events," in *Proc. IEEE 17th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, Jun. 2013, pp. 557–562.
- [28] M. A. Razzaq, A. Qamar, and H. S. M. Bilal, "Prediction and analysis of Pakistan election 2013 based on sentiment analysis," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 700–703.
- [29] J. Gulati and C. B. Williams, "Communicating with constituents in 140 characters or less: Twitter and the diffusion of technology innovation in the United States Congress," presented at the Annu. Meeting Midwest Political Sci. Assoc., Chicago, IL, USA, 2010.
- [30] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpel, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 178–185.
- [31] L. Spiliotopoulou, D. Damopoulos, Y. Charalabidis, M. Maragoudakis, and S. Gritzalis, "Europe in the shadow of financial crisis: Policy making via stance classification," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 2835–2844.
- [32] S. Hasbullah, D. Maynard, R. Z. W. Chik, F. Mohd, and M. Noor, "Automated content analysis: A sentiment analysis on Malaysian government social media," in *Proc. 10th Int. Conf. Ubiquitous Inf. Manage. Commun.*, 2016, Art. no. 30.
- [33] P. C. G. Reddick, A. T. Chatfield, and A. Ojo, "A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use," *Government Inf. Quart.*, vol. 34, no. 1, pp. 110–125, 2017.
- [34] J. I. Criado, R. Sandoval-Almazan, and J. R. Gil-Garcia, "Government innovation through social media," *Government Inf. Quart.*, vol. 30, no. 4, pp. 319–326, 2013.
- [35] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. LREC*, vol. 10, 2010, pp. 1320–1326.
- [36] N. Alsaedi, P. Burnap, and O. Rana, "Can we predict a riot? Disruptive event detection using Twitter," *ACM Trans. Internet Technol.*, vol. 17, no. 2, p. 18, 2017.
- [37] H. Saif, M. Fernandez, Y. He, and H. Alani, "Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-gold," in *Proc. AI*IA Conf.*, Turin, Italy, 2013.
- [38] S. Narr, M. Hulphenhaus, and S. Albayrak, "Language-independent Twitter sentiment analysis," *Knowl. Discovery Mach. Learn.*, pp. 12–14, Sep. 2012.
- [39] R. Artstein, "Inter-annotator agreement," in *Handbook of Linguistic Annotation*. Dordrecht, The Netherlands: Springer, Jun. 2017, pp. 297–313. [Online]. Available: http://link.springer.com/10.1007/978-94-024-0881-2_11
- [40] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Med.*, vol. 22, no. 3, pp. 276–282, 2012.
- [41] J. L. Fleiss, B. Levin, and M. Paik, "The measurement of interrater agreement," in *Statistical Methods for Rates and Proportions*. Hoboken, NJ, USA: Wiley, 2003.
- [42] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text With the Natural Language Toolkit*. Newton, MA, USA, O'Reilly Media, 2009.
- [43] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of roman-urdu opinions using Naïve Bayesian, decision tree and KNN classification techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, 2016.
- [44] M. R. Huq, A. Ali, and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 19–25, 2017.
- [45] T. S. Raghavendra and K. G. Mohan, "Web mining and minimization framework design on sentimental analysis for social tweets using machine learning," *Procedia Comput. Sci.*, vol. 152, pp. 230–235, Jan. 2019.
- [46] S. S. Istia and H. D. Purnomo, "Sentiment analysis of law enforcement performance using support vector machine and K-nearest neighbor," in *Proc. 3rd Int. Conf. Inf. Technol., Inf. Syst. Elect. Eng.*, Nov. 2018, pp. 84–89.
- [47] H. Hamdan, P. Bellot, and F. Bechet, "LSISLIF: CRF and logistic regression for opinion target extraction and sentiment polarity analysis," in *Proc. 9th Int. Workshop Semantic Eval.*, 2015, pp. 753–758.
- [48] F. Luo, C. Li, and Z. Cao, "Affective-feature-based sentiment analysis using SVM classifier," in *Proc. IEEE 20th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2016, pp. 276–281.
- [49] P. Chikersal, S. Poria, and E. Cambria, "SeNTU: Sentiment analysis of tweets by combining a rule-based classifier with supervised learning," in *Proc. 9th Int. Workshop Semantic Eval.*, 2015, pp. 647–651.



BIBI AMINA was born in Peshawar, Pakistan, in 1991. She received the M.Sc. degree in computer science from Shaheed Benazir Bhutto Women University, Peshawar, in 2014, where she has been a Visiting Lecturer with the Computer Science Department, since 2016. She is currently pursuing the M.S. degree in computer science with the Institute of Management Sciences, Peshawar. Her research interests include data mining, social media mining, social media analytics, topic models, and machine learning for industry.



TAYYABA AZIM was born in Rawalpindi, Punjab, Pakistan, in 1986. She received the master's degree in computer science from the National University of Computer and Emerging Sciences, Islamabad, in 2009, and the Ph.D. degree in computer science from the University of Southampton, U.K., in 2014. She was a Data Scientist with the Horizon Research Institute, University of Nottingham, and a Research Associate with Cortexica Vision Systems based at Imperial College London, U.K. Since 2015, she has been an Assistant Professor on Tenure track with the Institute of Management Sciences. She is the author of a book and has several conference and journal publications in the area of computer vision and machine learning. Her research interests include deep learning, topic models, kernel methods, and real time systems.

She was a recipient of Startup Research Grant, National Grassroots ICT Research Initiative Fund, IGNITE Research and Development Fund, and Overseas Ph.D. Scholarship, and has received the Best Paper Award at ICPRAM, in 2017.

...