COMPUTER RESEARCH AND MODELING
2024 VOL. 16 NO. 7 P. 1539–1554
DOI: 10.20537/2076-7633-2024-16-7-1539-1554

SPECIAL ISSUE

UDC: 519.6

Computational treatment of natural language text for intent detection

A. S. Adekotujo[1,2,a], T. Enikuomehin[2,b], B. Aribisala[2,c], M. Mazzara[3,d], A. F. Zubair[2,e]

[1]Computer Information and Management Studies Department (CIMSD), The Administrative Staff College of Nigeria (ASCON),
Topo-Badagry, Lagos State, Nigeria
[2]Department of Computer Science, Lagos State University,
Ojo Campus, Lagos State, Nigeria
[3]Innopolis University,
1 Universitetskaya st., Innopolis, 420500, Russia

E-mail: [a] adekotujoakinlolu@gmail.com, [b] toyinenikuomehin@gmail.com, [c] benjamin.aribisala@gmail.com, [d] m.mazzara@innopolis.ru, [e] adamfunsho@hotmail.com

Received 27.10.2024, after completion — 24.11.2024
Accepted for publication 25.11.2024

Intent detection plays a crucial role in task-oriented conversational systems. To understand the user's goal, the system relies on its intent detector to classify the user's utterance, which may be expressed in different forms of natural language, into intent classes. However, lack of data, and the efficacy of intent detection systems has been hindered by the fact that the user's intent text is typically characterized by short, general sentences and colloquial expressions. The process of algorithmically determining user intent from a given statement is known as intent detection. The goal of this study is to develop an intent detection model that will accurately classify and detect user intent. The model calculates the similarity score of the three models used to determine their similarities. The proposed model uses Contextual Semantic Search (CSS) capabilities for semantic search, Latent Dirichlet Allocation (LDA) for topic modeling, the Bidirectional Encoder Representations from Transformers (BERT) semantic matching technique, and the combination of LDA and BERT for text classification and detection. The dataset acquired is from the broad twitter corpus (BTC) and comprises various meta data. To prepare the data for analysis, a pre-processing step was applied. A sample of 1432 instances were selected out of the 5000 available datasets because manual annotation is required and could be time-consuming. To compare the performance of the model with the existing model, the similarity scores, precision, recall, f1 score, and accuracy were computed. The results revealed that LDA-BERT achieved an accuracy of 95.88 % for intent detection, BERT with an accuracy of 93.84 %, and LDA with an accuracy of 92.23 %. This shows that LDA-BERT performs better than other models. It is hoped that the novel model will aid in ensuring information security and social media intelligence. For future work, an unsupervised LDA-BERT without any labeled data can be studied with the model.

Keywords: hate speech, intent classification, Twitter posts, sentiment analysis, opinion mining, intent identification from Twitter posts

Citation: *Computer Research and Modeling*, 2024, vol. 16, no. 7, pp. 1539–1554.

1. Introduction

Users of social media platforms have naturally chosen to communicate their personal information in an open way, according to research trends. This is a result of their attachment and addiction to the unlimited usage of the service. Information can thus be communicated in millions of natural language texts every minute. Now that the information has emerged, it may be explicitly analyzed to determine the text's owner's intent and use that information to make decisions. Users do not need to give their private information in order to establish trust, which is crucial for social media intelligence and information security for many stakeholders [Yashkina et al., 2020].

Recently, computational modeling and analysis of human or natural language have received a lot of attention, thanks to Natural Language Processing (NLP). Furthermore, it has expanded its range of applications to include things like Information Extraction and Machine Translation [Khurana et al., 2023]. The science of NLP encompasses entities, intentions, and utterances.

The field of artificial intelligence and linguistics known as "natural language processing" (NLP) is focused on teaching computers how to comprehend sentences and words written in human languages. In Natural Language Processing (NLP), the result of a behaviour action is intent. A person's intent, in addition to the actions they take to produce an expected desirable result, is very important in negotiations since it allows them to express the outcome, they are interested in. These justifications, in theory, separate conduct from intent. The goal of intent understanding is to identify the action that a user wants a computer system to take or the information that a user would like to get and transmit that through spoken or written words. Intent can also be described as a future-oriented, deliberate activity that can be connected to present-day habits or aspirations, and it can be used to find knowledge that can be used to take action [Adekotujo et al., 2020].

However, there are issues that NLP encounters in intent detection, asides comprehending colloquial language, dealing with ambiguous claims, and distinguishing between similar intents [Al-Garadi et al., 2021; Chang et al., 2020; Collins et al., 2021; Mageto, 2021; Mayrhofer et al., 2019] it also includes Lack of data sources, irregularity of user expression, implicit intent detection, and multiple intent detection [Liu, Li, Lin, 2019].

It is quite challenging to solve the natural language problem of inferring user intent from speech or text in general. However, it can be gradually resolved by reducing and breaking down the purpose scope and domain into different steps. The detection and analysis of intent (ion) is a crucial part of social media modeling [Hutter et al., 2013]. This is mostly because of hidden motives and the risk that they could lead to unrest or improper social behavior if carried out. Recent literary study on determining intent using bi-grams, context, or word order has been scant [Agarwal, Sureka, 2017; Pang, Ruan, 2023; Ramage, Dumais, Liebling, 2010; Sabah, 2023]. With the addition of Contextual Semantic Search (CSS), context analysis requires the application of an upgraded feature model for extraction and a context-based technique that combines word order to effectively detect intent [Khan et al., 2022].

Therefore, it has become essential to understand the intentions of internet users in a variety of commercial sectors, including manufacturing, banking, real estate, tourism, e-commerce, and online marketing. This is crucial for information security and social media intelligence.

In contrast to earlier research, the LDA model is paired with Bert to extract keywords, which is essential for obtaining more precise text information. As a result, this research suggests CSS and LDA-Bert-based ensemble model for public opinion keyword extraction technique.

2. Literature review

The goal of "natural language processing" (NLP), a branch of linguistics and artificial intelligence, is to educate computers to understand words and phrases written in human languages.

The outcome of a behavior action in Natural Language Processing (NLP) is intent. A person's intent, in addition to the actions they take to produce an expected desirable result, is very important in negotiations since it allows them to express the outcome, they are interested in.

The goal of intent understanding is to identify the action that a user wants a computer system to take or the information that a user would like to get and transmit that through spoken or written words. Intent can also be described as a future-oriented, deliberate activity that can be connected to present-day habits or aspirations, and it can be used to find knowledge that can be used to take action [Adekotujo et al., 2020; Chowdhury et al., 2024].

Contextual Semantics for Radicalization Detection on Twitter [Fernandez, Alani, 2018] sought to identify radical content online primarily using radicalization glossaries, i. e., by looking for terms and expressions associated with religion, war, offensive language, etc. The study claimed that such crude methods are highly inaccurate towards content that uses radicalization, hence, an approach for developing a representation of the semantic context of the terms that are linked to radicalized content was proposed.

The study of [Fernandez, Alani, 2018] conducted a study on eight corpora of online hate speech and presented their findings in a paper titled Multilingual Cross-domain Perspectives on Online Hate Speech. The study showcased the NLP techniques used to gather and examine the jihadist, extremist, racist, and sexist contents. The study also concentrated on text categorization, text profiling, keyword and collocation extraction, manual annotation, and qualitative investigation to reveal the significant features.

For the detection of hate speech, dependable solutions are absent [Kiilu et al., 2018], despite the fact that the issues caused by social media content are well acknowledged. The major objective of this project is to create a trustworthy technology for identifying hateful tweets. Using information created by self-identified hateful communities on Twitter, this article provides a method for identifying and categorizing hateful speech. Naive Bayes classifier was used.

According to a study by [Ramadan et al., 2024; Zimmerman, Kruschwitz, Fox, 2018], ensemble methods are being modified for use with neural networks and presented to more accurately classify hate speech and Cross-dialectal Arabic Intent Detection. Neural network approaches are said to be state of the art for text classification challenges. Their approach makes use of an embedding model that is openly available and tested against a Twitter corpus of hate speech. The algorithm employed is Deep Learning Ensembles.

Moreover, using information theory quantifiers (entropy and divergence) to characterize texts, [Almeida et al., 2017] offered a novel method for identifying hate speech. As a unique feature of their method, they capture weighted information of words rather than merely their frequency in documents. [Newman et al., 2009] describes distributed algorithms for the Latent Dirichlet Allocation (LDA) model and the Hierarchical Dirichlet Process (HDP) model, two popular topic models, in their work Distributed Algorithms for Topic Models. The second model directly accounts for scattered data using a hierarchical Bayesian extension of LDA. They came to the conclusion that when compared to their sequential equivalents, LDA and HDP, distributed topic model algorithms Approximate Distributed LDA (ADLDA), Hierarchical Distributed LDA model (HD-LDA), and AD-HDP perform better.

The study [Correa, Sureka, 2013] presents a review of 40 published articles for solutions to detect and analyse online radicalization, examined these methods, and performed trend analysis in their study, Solutions to Detect and Analyse Online Radicalization. For the purpose of organizing the existing literature, they offered a brand-new multi-level taxonomy or framework.

The three primary thematic categories of race, nationality, and religion were used to abstract the problem of hate speech [Njagi et al., 2015]. Their goal is to create a model classifier that uses sentiment analysis and subjectivity detection techniques to identify and evaluate the polarity of sentiment expressions in addition to determining whether a particular sentence is subjective. After that, they created a vocabulary that is utilized to create a classifier for identifying hate speech by employing hate speech-related subjective and semantic criteria.

According to contextual factors related to social, political, cultural, and infrastructure conditions, radical and contentious activism may or may not develop into violent behaviour [Sanfilippo, McGrath, Bell, 2014]. Significant theoretical progress has been made in understanding these contextual factors and the significance of their interrelationships. However, there hasn't been much advancement in the creation of procedures and tools that make use of such theoretical developments to automate the detection of violent intent. They discussed the consequences of using the resulting system to evaluate the emergence of radicalization that results in violence, and proposed a framework that implements such processes and capacities.

Again, a vertical federated learning architecture based on variational quantum circuits was developed by [Yang et al., 2022] to show off a quantum-enhanced pre-trained BERT model's competitive classification of texts ability. Their intent classification tests demonstrate that the BERT-QTC model they propose achieves competitive test results in the Snips and ATIS spoken language datasets. In two text classification datasets, the BERT-QTC specifically improves the performance of the current quantum circuit-based language model by 1.57 % and 1.52 %, respectively.

In addition, the joint models frequently outperform the solo designs since intent classification and slot filling are two essential tasks in natural language understanding (NLU)T. Bidirectional Encoder Representations from Transformers, or BERT (Bidirectional Encoder Representations from Transformers), is one of the promising solutions, according to [Guo et al., 2022]. This method optimizes both the intent classification and slot filling tasks simultaneously.

Z-BERT-A is a Transformer-based, two-stage intent discovery technique that has been optimized with adapters. It was first trained for Natural Language Inference (NLI) and then used for unknown intent classification in a zero-shot context. It was proposed by [Comi et al., 2022]. In two zero-shot circumstances, known intents classification and unseen intent discovery, their tests demonstrate how Z-BERT-A outperforms a wide range of baselines.

The study [Badawi, 2023] asserts that technology has dominated a significant portion of human life. Additionally, people who utilize technology frequently employ words to convey their thoughts and feelings. One of the most active areas of research is sentiment analysis, which identifies human attitudes regarding a specific good, service, or subject. They conducted experiments utilizing the traditional machine learning approach, the most recent deep learning tool in NLP, and Bidirectional Encoder Representations from Transformers on a recently published medical corpus (BERT). The results show that BERT exceeds all machine learning classifiers by scoring (92 %) in accuracy, which is greater than machine learning classifiers, when comparing the results of both machine learning and deep learning.

Additionally, intent classification, detection and slot filling are two crucial tasks in NLU, the combined models typically outperform the solo designs. One of the promising methods is Bidirectional Encoder Representations from Transformers, or BERT [Guo et al., 2022; Huang, Cui, Wang, 2023]. This approach concurrently optimizes intent classification and slot filling duties.

In addition, forecasting the weather and meteorology are essential for anticipating future climatic conditions [Purwandari et al., 2023]. Today, people, particularly those who rely on the weather forecast, use big data to precisely assess information from social media. Three machine learning algorithms in total were looked at: the support vector machine (SVM), multinomial logistic regression (MLR), and multinomial Naive Bayes (MNB), as well as the pretrained bidirectional encoder representations of transformers (BERT), which was adjusted over several layers to guarantee precise classification. The F1-score of 99 % was used to calculate the accuracy of the BERT model, which was higher than any other machine learning technique.

Furthermore, cyber violence is eroding the social media landscape at an increasingly rapid rate [Zhou et al., 2023]. However, the classic self-reporting questionnaire is difficult to use in the contemporary cyber environment due to subjectivity and cost. In their paper, they provided a cyber aggression prediction model based on a state-of-the-art deep learning algorithm. On three fronts, social exclusion, defamatory humor, and guilt induction, they elaborated cyber assault. Then, using the BERT model that has already been trained, they developed the prediction model. Their empirical data demonstrated the BERT model's superior performance and greater prediction compared to conventional machine learning models without additional pretrained data.

Lastly, today's social media landscape is flooded with unfiltered content, which can range from hate speech to cyberbullying and cyberstalking [Mazari, Boudoukhani, Djeffal, 2024]. As a result, locating and eliminating such poisonous language presents a significant problem and is a current research subject. The suggested method builds numerous ensembles learning architectures using the Bidirectional Encoder Representations from Transformers (BERT) model that has already been trained in conjunction with Deep Learning (DL) models. As a result, we show that encoding texts with modern word embedding methods like FastText, GloVe, Bi-LSTM, and Bi-GRU can result in models that, when paired with BERT, can raise the ROC-AUC score to 98.63 %.

3. Proposed methodology

The research workflow entails fetching of tweets from API, the fetched tweets were subjected to contextual semantic search (CSS), and the output collated in a document. The CSS document or file is now pre-processed and stored in a data file. Furthermore, the generated data file is subjected to LDA, BERT, and LDA-BERT models in other to compare similarity ratio of the three models. Lastly, the three models were evaluated and the result enumerated.

Equation (1) can be used to detect intent. An intent detection algorithm produces either True (1) or False (0) as its output. The goal of intent detection is to determine whether a news story or twitter post has intent given the social news engagements it has had from n users. Such that,

$$I(a) = \begin{cases} 1, & \text{if } a \text{ contains intent,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $I(a)$ is the prediction function.

I is a notation for a specific type of formal language called a regular language. The notation "$I\colon \mathcal{E} \to 0, 1$" means that I is a rule or function that maps the empty string ($\mathcal{E}$) to the set of all strings that can be formed by the characters 0 and 1. This is a regular language, which can be recognized by a finite automation or regular expression.

$I(a)$ is a function that takes in a string "a" as input and outputs 1 if the string contains the word "intent" and 0 otherwise. This function can be used to determine if a given input string contains the word "intent".

An individual's purposeful expression or post is influenced by a wide range of variables, including personal interest, religion, and social influence [Mazari, Boudoukhani, Djeffal, 2024]. After performing intent filtering and intent parsing and extraction, we need to mine a specified intent domain in order to make the intent or goal mining problem computationally feasible.

The task intent detection is defined [Kröll, Strohmaier, 2009] as to approximate the unknown function:

$$\begin{gathered} \text{Given that}\colon S_i \in d_i,\ d_i \subset D, \\ f\colon S \times C^I \to \{True,\ False\}, \end{gathered} \tag{2}$$

where $C^I = \left\{c_1^I,\ c_2^I,\ \ldots,\ c_n^I\right\}$ is a set of predefined intent categories from document, where D is a domain that consist of text documents and each reviewed document d_i contains a sequence of sentences $S = \left\{s_1,\ s_2,\ \ldots,\ s_{|S|}\right\}$. It can be understood that in this context, "S" represents a set of sentences within a document, “D” which is a set of documents within a specific domain, and “C_1” is a set of predefined Intent categories. The function “f” takes in an input of a sentence “s_2” from set “S” and an intent category “c^i” from set “C^I” and outputs a Boolean value of either “True” or “False”. It is likely that this function is used to determine if the sentence “S_I” belongs to the intent category “C^I” or not.

Due to the dearth of labeled data in real-world techniques, a machine learning-based strategy with semi-supervised learning is offered.

The pre-processing of the intent dataset will include cleansing the intent tweets of nontextual content and unrelated subjects. After that, stop-word removal, tokenization, stemming, Part of Speech tagging, and lemmatization will be carried out using the relevant lexicons and semantics for the dataset. The classification procedure, as shown in Fig. 1, will next classify the tweets using deep learning models for natural language processing called Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT).

Precision, recall, and accuracy will be used to evaluate the proposed classifier’s performance, and performance will also be compared to that of existing techniques.

4. Data source

Python Application Programming Interface (API) was used to extract 5000 tweets as test data for the research implementation. The unique keywords used for the extraction was also subjected to Contextual Semantic Search (CSS) for a more robust classification of synonyms or semantics.

Connection was made to Twitter Streaming API to download the data using the Python package tweepy. The python code needed to connect to the Twitter Streaming API was then produced in a file called twitter streaming.py, as shown in Figs. 1, 2. The data collecting code for Twitter data collection will require all four personal credentials.

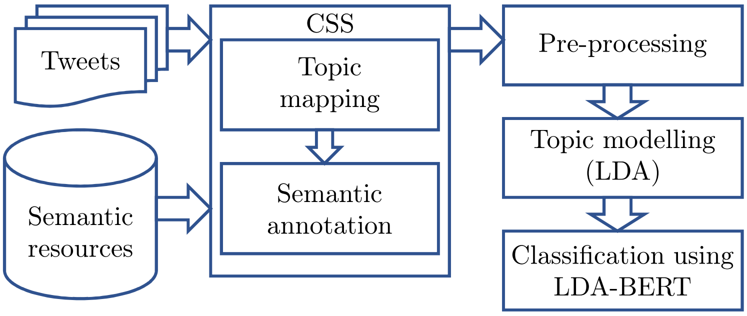


Figure 1. The systematic workflow of the Proposed Model for Intent Classification

5. Features of Twitter dataset

Twitter is a popular social media platform that allows users to post and interact with short messages called “tweets”. A Twitter dataset typically includes information about these tweets and the users who posted them. Here are some common features of Twitter datasets:

1. Tweet text: This is the main content of a tweet. It can contain up to 280 characters and can include links, hashtags, mentions, and emojis.

2. Timestamp: This indicates the date and time when the tweet was posted.

3. User ID: Each Twitter user has a unique ID associated with their account. This allows tweets to be linked to specific users.

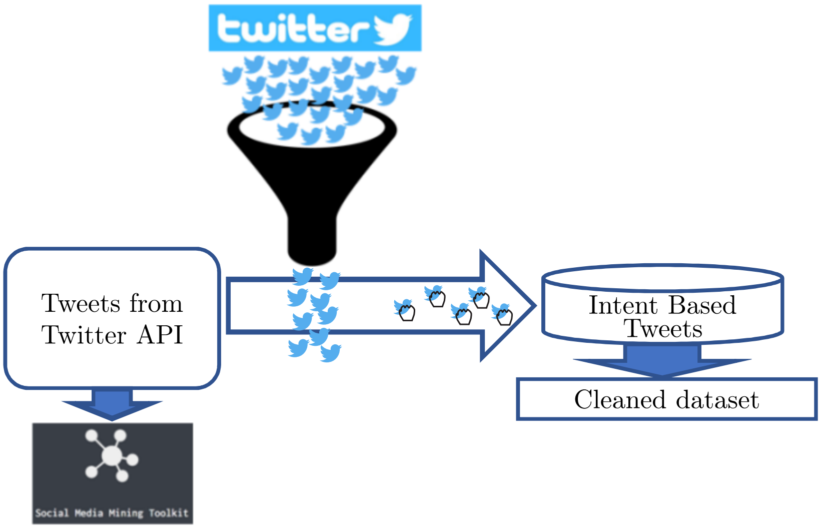


Figure 2. Dataset Flow Scheme

4. User name: This is the public name that a user has chosen to identify themselves on Twitter.
5. Follower count: This is the number of other Twitter users who follow a particular user.
6. Retweet count: This is the number of times a tweet has been retweeted (i. e., shared by other users.
7. Favourite count: This is the number of times a tweet has been favourited (i. e., marked as a favorite by other users).
8. Hashtags: These are keywords or phrases preceded by the "#" symbol that are used to categorize tweets and make them more discoverable.
9. Mentions: These are Twitter usernames preceded by the "@" symbol that are used to reference other users in a tweet.
10. URL links: These are clickable links that can be included in a tweet.
11. Media attachments: Twitter allows users to attach images, videos, and GIFs to their tweets.

6. Feature extraction: word embedding

Word embeddings are a widely used method in Natural Language Processing (NLP) that enables words to be represented as vectors of real numbers in a high-dimensional space. This method has grown to be a crucial component of many NLP models, including named entity recognition, sentiment analysis, and language translation.

Word embeddings generate the semantic and syntactic meaning of words in the context of the text. The key idea is to represent words as dense vectors that are close to each other in this high-dimensional space if they have similar meanings or are used in similar contexts. In contrast, words that have different meanings or are used in different contexts are represented as vectors that are far apart in the space. Word embeddings can be generated in a number of ways. The Word2Vec algorithm is among the most widely used techniques. This approach uses a vast corpus of text to train a neural network, which then learns word embeddings. The network forecasts a word's probability based on its context words, or the other way around. The weights of the neural network's hidden layer, which is usually somewhat lower than the input and output layers, are the word embeddings that are produced.

Examine the following line, for example: "The movie was amazing, I loved every minute of it". A pre-trained word embedding model, such Word2Vec or GloVe, can be used to represent each word in the phrase as a 100-dimensional vector. By factorizing a co-occurrence matrix or training a neural network on a sizable corpus of text, these word embeddings can be generated. This is how the final vectors could appear:

- The: [0.12, 0.35, −0.25, . . .],
- movie: [0.78, 0.44, 0.01, . . .],
- was: [−0.05, 0.21, 0.77, . . .],
- amazing: [0.91, 0.67, −0.12, . . .],
- I: [0.06, 0.22, −0.37, . . .],
- loved: [0.87, 0.54, 0.03, . . .],
- every: [−0.14, 0.31, −0.52, . . .],
- minute: [0.72, 0.41, −0.18, . . .],
- of: [−0.28, 0.09, 0.63, . . .],
- it: [0.13, 0.26, −0.09, . . .].

Attention can now be given to these word embeddings into a single feature vector for the entire sentence, and use this as input to a machine learning algorithm such as a neural network or a support vector machine. The algorithm can learn to predict the sentiment of the sentence based on the relationship between the word embeddings.

In this example, the word "amazing" has a high positive sentiment score in the word embedding space, which would contribute to a positive sentiment label for the entire sentence. Similarly, the word "loved" has a high positive sentiment score and would also contribute to a positive sentiment label. By contrast, if the word "hated" were used instead of "loved", the sentiment label would likely be negative because "hated" has a high negative sentiment score in the word embedding space.

7. Measurement of the result

Accuracy, Precision, Recall, and F1 score are commonly used evaluation metrics in Natural Language Processing (NLP) to measure the performance of a classification model. These metrics are particularly useful in evaluating models that deal with imbalanced datasets, where the number of instances in each class is significantly different.

Accuracy. Accuracy is the simplest evaluation metric that measures the percentage of correctly classified instances out of the total number of instances. It is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where:

- TP: True Positive, the number of instances that are correctly classified as positive.
- TN: True Negative, the number of instances that are correctly classified as negative.
- FP: False Positive, the number of instances that are incorrectly classified as positive.
- FN: False Negative, the number of instances that are incorrectly classified as negative.

For example, suppose you have a sentiment classification model that classifies customer reviews as positive or negative. Out of 1000 reviews, the model correctly classified 900 as positive and 80 as negative, and misclassified 10 as positive and 10 as negative. The accuracy of the model would be:

$$\text{Accuracy} = \frac{900 + 80}{900 + 80 + 10 + 10} = 0.95 \text{ or } 95\,\%.$$

Precision. Precision is a metric that measures the proportion of true positives (TP) among all the instances classified as positive. It measures the model's ability to correctly identify positive instances. It is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

For example, in the above sentiment classification example, the precision of the model for the positive class would be:

$$\text{Precision} = \frac{900}{900 + 10} = 0.989 \text{ or } 98.9\,\%.$$

Recall. Recall is a metric that measures the proportion of true positives (TP) among all the instances that are actually positive. It measures the model's ability to correctly identify all positive instances. It is calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

For example, in the above sentiment classification example, the recall of the model for the positive class would be:

$$\text{Recall} = \frac{900}{900 + 10} = 0.989 \text{ or } 98.9\,\%.$$

F1 Score. F1 score is a harmonic mean of precision and recall, which balances both metrics. It is a good metric to use when the dataset is imbalanced or when both precision and recall are equally important. It is calculated as follows:

$$\text{F1 score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}.$$

For example, in the above sentiment classification example, the F1 score of the model for the positive class would be:

$$\text{F1 score} = \frac{2 \cdot (0.989 \cdot 0.989)}{0.989 + 0.989} = 0.989 \text{ or } 98.9\,\%.$$

Accuracy measures how often the model is correct, precision measures how often the model is correct when it predicts positive, recall measures how often the model is correct when the actual value is positive, and F1 score balances both precision and recall.

8. Experiment setup

The purpose of this experiment is to detect the intent of different tweets, for the purpose of consistency hate is selected as the topic of intent to be detected for this experiment. We intend to classify tweets into two categories, hate speech and non-hate speech, and measure the degree of similarity of each tweet to the topic of hate. The experiment will use pre-trained BERT (Bidirectional Encoder Representations from Transformers) model and Latent Dirichlet Allocation (LDA) which are deep learning models for natural language processing, to classify the tweets.

The experiment will follow the following steps:

a) data collection,

b) data preprocessing,

c) topic modeling,

d) model training,

e) tweets classification.

9. Data collection

A dataset of tweets was collected, containing both hate speech and non-hate speech messages. The experiment was carried out using the web tool colab.research.google.com with reference to the figure 7 flowchart. Following the smooth import of all necessary libraries as shown in the flowchart, all credentials necessary to connect to the Twitter API and to authenticate were then assigned. The tweets collected comprised various meta data such as tweet text, timestamp, user ID, user name, follower count, retweet count, favorite count, hashtags, mentions, URL links, and media attachments features. The total number of datasets collected was about 5000. These reduced to 1432 after pre-processing by removal of incomplete datapoints. This subsampling was implemented to mitigate the potential time-consuming nature of manual annotationDue to a daily constraint of 1 %, only 1432 of the intended 5000 tweets for the test dataset could be collected. The tweepy package and Python are used in the process. The Python script below was used for the data collection procedure.

10. Data pre-processing

The tweet data was pre-processed to prepare it for use by cleaning the data, removing stop words, tokenizing and lemmatization of the data into word-level segments. The @mentions, # symbol, Retweets, hyperlinks, columns, and spaces were removed from the dataset. The polarity and the subjectivity were then computed using a function. Following an equally distributed WordCloud generation, a computation of negative, positive, and neutral analysis was performed.

11. Topic modeling with LDA

In this experiment, as shown in Fig. 3, LDA was used as a step to model the topics of the tweets before feeding the data into the BERT model for classification. This involved fitting an LDA model to the matrix of the pre-processed tweets to uncover the underlying topics.

The topics uncovered by LDA provided valuable insights into the content of the tweets and helped to improve the performance of the BERT model by providing additional context and information about the data. In addition, the topics generated by LDA can be used to gain a better understanding of the distribution of hate speech and non-hate speech tweets in the dataset.

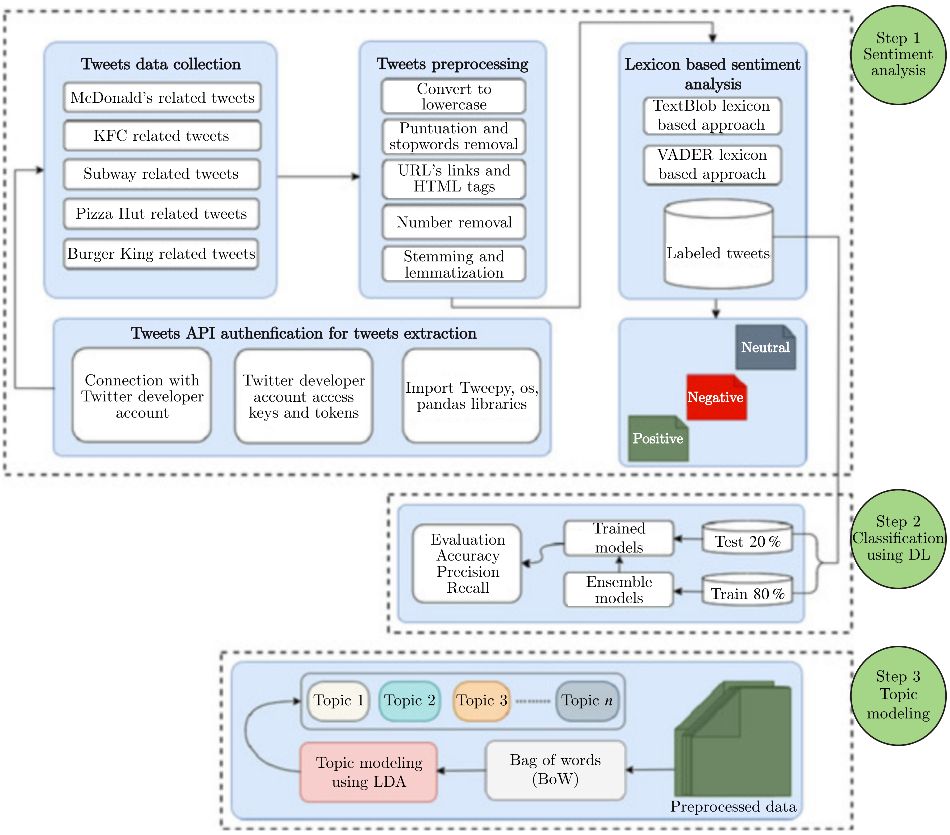

Tweets data collection
McDonald's related tweets
KFC related tweets
Subway related tweets
Pizza Hut related tweets
Burger King related tweets
Tweets preprocessing
Convert to lowercase
Puntuation and stopwords removal
URL's links and HTML tags
Number removal
Stemming and lemmatization
Lexicon based sentiment analysis
TextBlob lexicon based approach
VADER lexicon based approach
Labeled tweets
Step 1 Sentiment analysis
Tweets API authenfication for tweets extraction
Connection with Twitter developer account
Twitter developer account access keys and tokens
Import Tweepy, os, pandas libraries
Neutral
Negative
Positive
Evaluation Accuracy Precision Recall
Trained models
Ensemble models
Test 20 %
Train 80 %
Step 2 Classification using DL
Topic 1
Topic 2
Topic 3
Topic n
Topic modeling using LDA
Bag of words (BoW)
Preprocessed data
Step 3 Topic modeling

Figure 3. The LDA workflow diagram of the adopted methodology for sentiment analysis and topic modeling. Source: [Mujahid et al., 2023]

12. Model training with BERT

The BERT model was fine-tuned on the collected LDA data. This involved adding a classification head to the BERT model and training it to classify the tweets into the two categories of hate speech and non-hate speech.

As shown in Fig. 4, the BERT can solve neural machine translation, question answering, sentiment analysis, text summarization problems because of its ability of understanding languages. However, BERT can be Fine tune to solve other tasks like text classification and compare text by generating similarity score between texts in context.

The python script for training the BERT is as shown in Fig. 5.

13. Classifying new tweets

The trained BERT model was used to classify 1432 new, unseen tweets into the two categories of hate speech and non-hate speech. The model generated output scores indicating the degree of similarity of each tweet to the selected topic (hate).

14. Experiment result

In this experiment, each tweet from the list of new 1432 tweets were classified by similarity to the topic "hate", using BERT model, LDA and Hybrid model (combination of BERT and LDA). The similarity score for each model was recorded respectively (see Table 1).

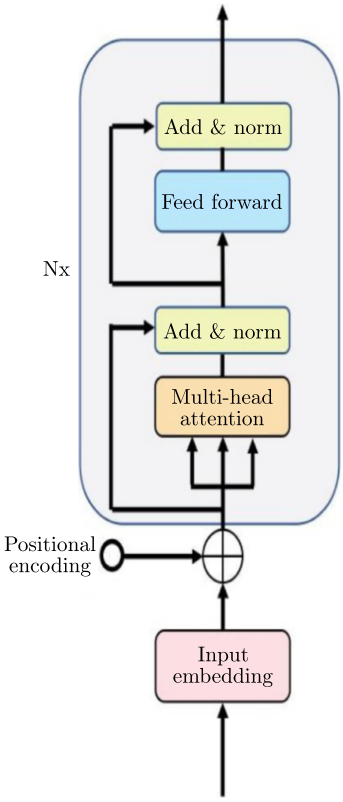
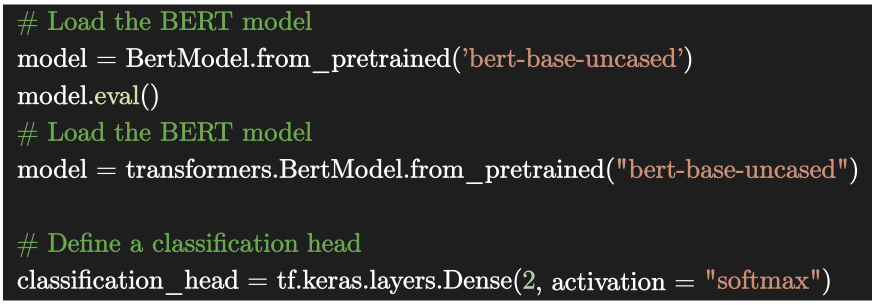


Figure 4. BERT. Source: [Shi et al., 2023]



```
# Load the BERT model
model = BertModel.from_pretrained('bert-base-uncased')
model.eval()
# Load the BERT model
model = transformers.BertModel.from_pretrained("bert-base-uncased")

# Define a classification head
classification_head = tf.keras.layers.Dense(2, activation = "softmax")
```

Figure 5. Part of Python script of BERT for the research

15. Model evaluation

The three models were evaluated using an annotated ground truth of the 1432 tweet dataset to measure accuracy of classifying the tweets. The models were evaluated based on accuracy, precision, recall, and F1 score. The final result is presented in Table 2.

16. Discussion

The implication and justifications of the experiment results are highlighted accordingly. The discussion of the results is presented in two broad sections; the description of the results, implication and justification of the results.

Table 1. Output of research classification

Document text	Hateful label	BERT similarity	LDA	LDA-BERT
The bitch coach said she didnt like my attitude	1	0.754015	0.625836	0.909089
The bitch got some nerve	1	0.749834	0.625836	0.89876
The bitch is free	1	0.749009	0.62557	0.892488
The bitch official though dick harder than a missile you	1	0.736598	0.625151	0.891957
The bitch who just rolled her eyes at me bc she didnt get tipped doesnt listen when u ask for sweet potato fries instead of regular	1	0.732734	0.624851	0.880178
The bitch who shot CJ is going upstate	1	0.728933	0.624755	0.878575
The bitches act like they dont be in the club with they ass all out in too little ass skirts shorts giving em yeast infection n shit	1	0.72106	0.624642	0.865939
The bitches are back	1	0.71801	0.624642	0.862284
The bitches behind me are so annoying	1	0.712104	0.624413	0.857167
The blacks in California are typical niggers	1	0.709797	0.6241	0.853543
The boy who shot albino deer	1	0.704409	0.623845	0.850989

Table 2. Performance ratio using BERT, LDA, and LDA-BERT

Model	Accuracy	Class	Precision	Recall	F1 score
BERT	93.84 %	Hate	90.46 %	93.46 %	93.82 %
		Non-Hate	93.01 %	92.15 %	93.91 %
LDA	92.23 %	Hate	90.12 %	89.63 %	88.43 %
		Non-Hate	91.85 %	91.77 %	90.74 %
BERT + LDA	95.67 %	Hate	93.46 %	94.56 %	93.66 %
		Non-Hate	93.65 %	94.33 %	92.87 %

Result description

The result of our experiment is presented in Table 2. The result shows that LDA-BERT model has better accuracy, precision, recall, and F1 score compare to standalone LDA and BERT models. BERT model also appears to be better than LDA in all the selected metrics.

Implication and justification of results

Despite these limitations, the combination of BERT and LDA has the potential to greatly improve the accuracy of hate intent classification. By considering both the strengths and weaknesses of this approach, researchers and practitioners can effectively utilize this combination to mitigate the impact of hate intent or any form of intent and promote a more inclusive and respectful online environment.

Limitation

One of the major limitations is that only 1432 instances out 5000 were selected from the twitter API, after pre-processing because manual annotation is required and could be time-consuming. Lack of human-labeled data for NLU and other natural language processing (NLP) tasks could also result in poor generalization capability.

Computational resources

The default CPU for Colab is an Intel Xeon CPU with 2 vCPUs (virtual CPUs) and 13 GB of RAM, and the default GPU for Colab is a NVIDIA Tesla K80 with 12 GB of VRAM (Video Random-Access Memory).

17. Conclusion

This research has demonstrated that the deployment of Contextual Semantic Search (CSS) is an efficient technique to boost the effectiveness and efficiency of retrieving tweets by using synonyms of the keywords used during the search from API. The ensemble model, comprising LDA for topic modeling and BERT for intent classification has come out with a better accuracy over and above previous models. The intent detection program has been made available as a free source program on Google Colab. This can be used remotely by any investigator for intent detection. This experiment mainly demonstrated the accuracy of using LDA-BERT for classifying tweets into hate speech and non-hate speech categories compared with LDA and BERT. The similarity ratio of each model was also compared. The results of this experiment can be used to inform the development of automated systems for detecting intent of different categories on social media platforms. The LDA-BERT result has precision, recall, F1 score, and accuracy of 93.5 %, 99.1 %, 96.2 %, and 95.9 %, respectively.

18. Recommendations

The intent detection model could be extended to include more than two models as ensemble model to improve its effectiveness. Inclusion of other models could ensure a more thorough intent classification and detection. LDA is a crucial dimension reduction tool in machine learning since it is a supervised learning method that trains on labeled data. In the study work, an unsupervised LDA-UEL without any labeled data or supervised information can be studied with the model for future work. A corpus of plain text was used to train BERT, an advanced unsupervised, bidirectional language representation. Although BERT pre-training is unsupervised in terms of downstream tasks, it is actually a supervised learning task in its own right. As a result, BERT in the research model can be replaced with an unsupervised model.

References

Adekotujo A. S., Lee J., Enikuomehin A. O., Mazzara M., Aribisala S. B. Bi-lingual intent classification of twitter posts: a roadmap // Proceedings of 6th International Conference in Software Engineering for Defence Applications: SEDA 2018. — 2020.

Agarwal S., Sureka A. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website // arXiv. — 2017. — https://arxiv.org/abs/1701.04931

Al-Garadi M. A., Yang Y.-C., Cai H., Ruan Y., O'Connor K., Graciela G.-H., Perrone J., Sarker A. Text classification models for the automatic detection of nonmedical prescription medication use from social media // BMC Medical Informatics and Decision Making. — 2021. — Vol. 21. — P. 1–13.

Almeida T. G., Souza B. À., Nakamura F. G., Nakamura E. F. Detecting hate, offensive, and regular speech in short comments // Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web.

Badawi S. S. Using multilingual bidirectional encoder representations from transformers on medical corpus for Kurdish text classification // ARO – The scientific journal of Koya university. — 2023. — Vol. 11, No. 1. — P. 10–15.

Chang V., Liu L., Xu Q., Li T., Hsu C.-H. An improved model for sentiment analysis on luxury hotel review // Expert Systems. — 2020. — Vol. 40. — P. e12580.

Chowdhury M. S. A., Chowdhury M., Shanto A., Murad H., Das U. Fired_from_NLP at AraFinNLP 2024: Dual-Phase-BERT-A fine-tuned transformer-based model for multi-dialect intent detection in the financial domain for the arabic language // Proceedings of The Second Arabic Natural Language Processing Conference. — 2024.

Collins B., Hoang D. T., Nguyen N. T., Hwang D. Trends in combating fake news on social media — a survey // Journal of Information and Telecommunication. — 2021. — Vol. 5, No. 2. — P. 247–266.

Comi D., Christofidellis D., Piazza P. F., Manica M. Z-BERT-A: a zero-shot Pipeline for Unknown Intent detection // arXiv. — 2022. — https://arxiv.org/abs/2208.07084

Correa D., Sureka A. Solutions to detect and analyze online radicalization: A survey // arXiv. — 2013. — https://arxiv.org/abs/1301.4916

Fernandez M., Alani H. Contextual semantics for radicalisation detection on Twitter // Semantic Web for Social Good Workshop (SW4SG) at International Semantic Web Conference 2018, 9 Oct 2018, CEUR. — 2018.

Guo Y., Xie Z., Chen X., Chen H., Wang L., Du H., Wei S., Zhao Y., Li Q., Wu G. ESIE-BERT: enriching sub-words information explicitly with BERT for joint intent classification and slot filling // arXiv. — 2022. — https://arxiv.org/abs/2211.14829

Huang J., Cui Y., Wang S. Adaptive local context and syntactic feature modeling for aspect-based sentiment analysis // Applied Sciences. — 2023. — Vol. 13, No. 1. — P. 603.

Hutter K., Hautz J., Dennhardt S., Füller J. The impact of user interactions in social media on brand awareness and purchase intention: the case of MINI on Facebook // Journal of Product & Brand Management. — 2013. — Vol. 22, No. 5/6. — P. 342–351.

Khan M. Q., Shahid A., Uddin M. I., Roman M., Alharbi A., Alosaimi W., Almalki J., Alshahrani S. M. Impact analysis of keyword extraction using contextual word embedding // PeerJ Computer Science. — 2022. — Vol. 8. — P. e967.

Khurana D., Koli A., Khatter K., Singh S. Natural language processing: state of the art, current trends and challenges // Multimedia Tools and Applications. — 2023. — Vol. 82, No. 3. — P. 3713–3744.

Kiilu K., Okeyo G., Rimiru R., Ogada K. Using Naïve Bayes algorithm in detection of hate tweets // International Journal of Scientific and Research Publications. — 2018. — Vol. 8, No. 3. — P. 99–107.

Kröll M., Strohmaier M. Analyzing human intentions in natural language text // Proceedings of the fifth international conference on Knowledge capture. — 2009.

Liu J., Li Y., Lin M. Review of intent detection methods in the human-machine dialogue system // Journal of physics: conference series. — 2019. — Vol. 1267, No. 1. — P. 012059.

Mageto J. Big Data analytics in sustainable supply chain management: a focus on manufacturing supply chains // Sustainability. — 2021. — Vol. 13. — P. 7101.

Mayrhofer M., Matthes J., Einwiller S., Naderer B. User generated content presenting brands on social media increases young adults' purchase intention // International Journal of Advertising. — 2019. — Vol. 39, No. 1. — P. 166–186.

Mazari A. C., Boudoukhani N., Djeffal A. BERT-based ensemble learning for multi-aspect hate speech detection // Cluster Comput. — 2024. — Vol. 27. — P. 325–339.

Mujahid M., Rustam F., Alasim F., Siddique M., Ashraf I. What people think about fast food: opinions analysis and LDA modeling on fast food restaurants using unstructured tweets // PeerJ Comput Sci. — 2023. — Vol. 9. — P. e1193.

Newman D., Asuncion A., Smyth P., Welling M. Distributed algorithms for topic models. Journal of Machine Learning Research. — 2009. — Vol. 10.

Njagi D., Zuping Z., Hanyurwimfura D., Long J. A lexicon-based approach for hate speech detection // International Journal of Multimedia and Ubiquitous Engineering. — 2015. — Vol. 10. — P. 215–230.

Pang H., Ruan Y. Determining influences of information irrelevance, information overload and communication overload on WeChat discontinuance intention: The moderating role of exhaustion // Journal of Retailing and Consumer Services. — 2023. — Vol. 72. — 103289.

Purwandari K., Cenggoro T. W., Sigalingging J. W. C., Pardamean B. Twitter-based classification for integrated source data of weather observations // IAES International Journal of Artificial Intelligence (IJ-AI). — 2023. — Vol. 12, No. 1. — P. 271–283.

Ramadan A., Amr M., Torki M., El-Makky N. M. MA at AraFinNLP2024: BERT-based ensemble for cross-dialectal arabic intent detection // Proceedings of The Second Arabic Natural Language Processing Conference. — 2024.

Ramage D., Dumais S., Liebling D. Characterizing microblogs with topic models // Proceedings of the international AAAI conference on web and social media. — 2010. — P. 130–137.

Sabah N. M. The impact of social media-based collaborative learning environments on students' use outcomes in higher education // International Journal of Human – Computer Interaction. — 2023. — Vol. 39, No. 3. — P. 667–689.

Sanfilippo A., McGrath L., Bell E. Computer modeling of violent intent: A content analysis approach // International Handbook of Threat Assessment / Eds.: J. R. Meloy, J. Hoffman. — New York, NY, US: Oxford University Press, 2014. — P. 224–235.

Shi Z., Luktarhan N., Song Y., Tian G. BFCN: A novel classification method of encrypted traffic based on BERT and CNN // Electronics. — 2023. — Vol. 12, No. 3. — P. 516.

Yang C.-H. H., Qi J., Chen S. Y.-C., Tsao Y., Chen P.-Y. When bert meets quantum temporal convolution learning for text classification in heterogeneous computing // ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — 2022.

Yashkina E., Pinigin A., Lee J., Mazzara M., Adekotujo A. S., Zubair A., Longo L. Expressing trust with temporal frequency of user interaction in online communities // Advanced Information Networking and Applications: Proceedings of the 33rd International Conference on Advanced Information Networking and Applications (AINA-2019). — 2020.

Zhou Z., Yu M., He Y., Peng X. When cyber aggression prediction meets BERT on social media // arXiv. — 2023. — https://arxiv.org/abs/2301.01877

Zimmerman S., Kruschwitz U., Fox C. Improving hate speech detection with deep learning ensembles // Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). — 2018.