Colloquial Expressions Detection Using Machine and Deep Learning Techniques

Shafaat Siddhi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
shafaat.siddhi@g.bracu.ac.bd

Syed Tanvir Shahriar
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
syed.tanvir.shahriar.sizan@g.bracu.ac.bd

***Abstract*—The motivation and goals of the topic are to explore how AI can enhance language learning and teaching by simulating human conversation and providing feedback and guidance.**

Understand the challenges and opportunities of identifying and using colloquial words in natural language processing (NLP), such as variability, sparsity, dynamics, accuracy, and robustness.

Teach AI to use colloquial language appropriately and effectively in different contexts and domains, such as e-commerce, healthcare, education, and entertainment.

***Index Terms*—NLP, Colloquial, Machine Learning, Deep Learning, Slang, Spacy, Roberta, Word Embeddings**

I. Introduction

Conversational AI is a type of AI that can simulate human conversation by using natural language processing (NLP), a field of AI that allows computers to understand and process human language, and foundation models, such as BERT or GPT-3, that power new generative AI capabilities.

Colloquial words are informal words or phrases that are used in everyday speech, often depending on the region, culture, or context of the speaker. They can make the language more natural and engaging, but they can also be confusing or unfamiliar to some listeners or readers.

Some examples of colloquial words in different languages and domains are lol, wanna, or what's up in English, jajaja, vale, or qué tal in Spanish, or pop, biscuit, or swag in American or British English.

Some existing methods and tools for identifying and generating colloquial words in NLP are rule-based methods, which use hand-crafted rules or patterns to match colloquial expressions in text, such as regular expressions or lexicons, statistical methods, which use probabilistic models or features to learn colloquial expressions from data, such as n-grams, word embeddings, or topic models, and neural methods, which use deep neural networks to automatically learn representations and classifiers for colloquial expressions, such as CNNs, RNNs, or transformers.

II. Literature Review

The difficulties of natural language processing (NLP) in dealing with idiomatic expressions (IEs): IEs are sentences whose literal meanings do not adequately convey their figurative or nonliteral meaning. Constantly continuously added to languages, they are an integral aspect of natural language. Nevertheless, they need NLP systems to be able to deal with cultural uniqueness, contextual ambiguity, and semantic non-compositionality, which is a traditional difficulty. Previous research has discussed the effects of IEs on several natural language processing (NLP) applications and offered solutions to these problems. These applications include sentiment analysis, dialog modeling, paraphrase creation, NLP inference, and natural language processing. Idiom processing models and methods: The idiom processing issue has inspired several models and approaches, the two primary of which are idiomatic type classification and idiomatic text classification. While idiom word classification finds out if a particular PIE is used literally or metaphorically in a phrase, idiom type classification attempts to decide whether a collection of MWEs may be used as IEs without taking further context into account. To capture the unique aspects of IEs and differentiate them from literal expressions, most of the current models and approaches depend on linguistic factors including grammatical structure, embeddings of words, and outside sources of data. Data sparsity, poor generalizability, and inadequate semantic portrayal of IEs are common issues with these models and approaches. [1] New developments and potential paths forward: Idiom processing has recently benefited from new developments in natural language processing (NLP), such as transformer-based architectures and pre-trained language models (PTLMs). While some research has shown that PTLMs can handle idiomatic expressions (IEs), the context-dependent and semantic ambiguity present in IEs remains a challenge. Adapters, denoising automatic encoders, comparison enforcing, and contrastive learning are among of the innovative methods suggested in other research as ways to improve IE representation and understanding. [2] Still, a lot of ground has to be covered in terms of future research, including expanding to previously unexplored IEs, bringing in additional external information, and using idiom-aware PTLMs in downstream applications, among other things.

III. Methodology

A. Data Scraping

We scraped Twitter for disaster data as disaster tweets contain more colloquial words and sentences than any other tweets. Roughly 10000 tweets were scraped. we carefully

label each piece of data. also, we used supervised learning to validate the labels of our data points. we divided train and test data into 70% - 30% respectively.

B. Data Prepossessing

- With the help of the pandas python library and regular expression library re we cleaned all the unnecessary usernames, special characters, and web links.
- Moreover, then we proceed to remove stop words from the data. and tokenized the data in three methods. that BERT, Spacy, and Roberta respectively.
- We used three tokenization methods to see if there is any significance in the machine learning model's accuracy with three different tokens.

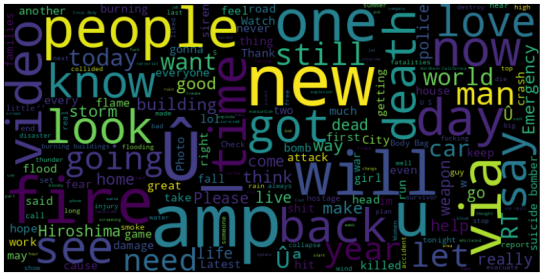

Fig. 1. Word cloud of the data set.

C. Machine Learning

With the processed text we ran 3 machine-learning models that are best fit for nlp related tasks. that is Support Vector Machines, Gaussian Naive Bayes, and Random Forest respectively.

- SVMs, or support vector machines:
  Type: Algorithm for supervised learning. Regression analysis and classification are the two uses for SVM. The basic idea behind SVM is to identify the hyperplane that divides the data into the most distinct classes. The data points that are closest to the decision boundary (hyperplane) are known as "support vectors". SVM is resilient to outliers because it strives to maximize the margin between classes. Kernel Trick: SVM can handle non-linearly separable data by transforming the input space using a kernel trick.

- Random Forest:
  Type: Algorithm for ensemble learning. The main applications of Random Forest are in regression and classification. The idea behind Random Forest is that it creates several decision trees during training and then combines them to produce predictions that are more reliable and accurate. To arrive at the final forecast, all the trees in the forest are trained using a random selection of the data (classification: voting, regression: averaging). Benefits: Random Forest is renowned for its resilience to overfitting, ability to handle huge datasets with high dimensionality, and robustness.

- Naive Bayes:
  Algorithm type: Probabilistic. The primary application of Naive Bayes is in classification jobs. Principle: The foundation of Naive Bayes is the Bayes theorem, which determines a hypothesis's probability using previously determined probabilities. The "naive" assumption of feature independence—that is, the assumption that features are independent given the class—allows Naive Bayes to function successfully in practice for a wide range of real-world scenarios. Text Classification: Naive Bayes is frequently utilized for text classification applications, including sentiment analysis and spam filtering.

.

D. Deep Learning

- Long Short-Term Memory, or LSTM:
  Type: architecture of recurrent neural networks (RNNs). The goal of LSTM is to improve the learning and memory of long-term dependencies in sequential data, such as language sequences, by resolving the vanishing gradient issue that plagues conventional RNNs. Important characteristics: Memory Cells: Long-term information retention is a feature of LSTM memory cells, which enables them to gradually identify dependencies. Gate Mechanisms: Long Short-Term Memory (LSTM) devices can selectively recall or forget information because they employ three gate mechanisms (input, forget, and output) to govern the flow of information. Applications: LSTMs are widely employed in tasks involving sequential data, including speech recognition, machine translation, and language modeling.

- BERT (Bidirectional Encoder Representations from Transformers):
  Architecture of the transformer type. The goal of BERT is to pre-train contextualized word representations by identifying the relationships and bidirectional context between words in a phrase. In several NLP tasks, it has greatly enhanced the state-of-the-art performance. Important characteristics: Bidirectional Context: BERT analyses a word's full context by simultaneously examining its left and right contexts, in contrast to typical models that only analyze words in isolation or a left-to-right or right-to-left context. Attention Mechanism: BERT may focus on different parts of the input sequence with varied levels of attention since it makes use of self-attention mechanisms from the transformer architecture. Applications: BERT embeddings are frequently employed as features or optimized for certain downstream NLP tasks like named entity recognition and question answering.

.

E. Model Evaluation

- Accuracy: The percentage of cases out of all instances that are correctly classified. Precision is a measure of the accuracy of positive predictions that is calculated as

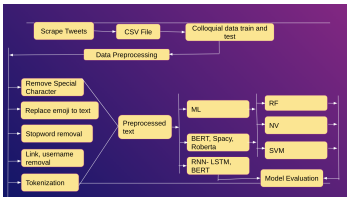

Fig. 2. Dataflow from collection to evaluation.

the ratio of true positives to the total of true positives and false positives. The ratio of true positives to the total of true positives and false negatives is known as recall (sensitivity or true positive rate), and it indicates how well the model can find all pertinent occurrences.

- F1 Score: A balanced metric that takes into account both false positives and false negatives, calculated as the harmonic mean of precision and recall. When there is an imbalance in the distribution of classes, it is very helpful.
- Region Plotting the true positive rate against the false positive rate, the Receiver Operating Characteristic curve (AUC-ROC) is a binary classification statistic.

IV. Results Analysis

we that finding colloquial words is a complex task. we believe further fine-tuning the dataset and the model both with give a satisfactory result.

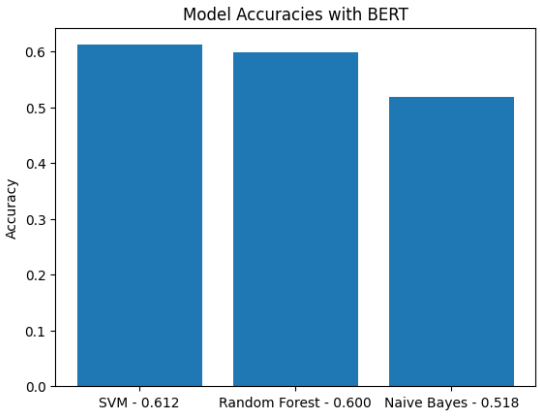

Fig. 3. ML model results with BERT tokenizaiton

TABLE I
Model Evaluation

Model Name	**Accuracy (%)**
SVM	61.7
Random Forest	59.7
Naive Bayes	55.7
LSTM	63.6
BERT	66.2

Among the models **BERT** with deep learning showed the highest accuracy of 66.2% with 3 epochs and batch size

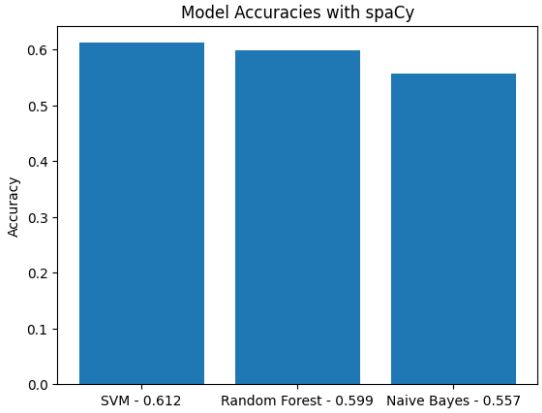

Fig. 4. ML model results with spaCy tokenizaiton

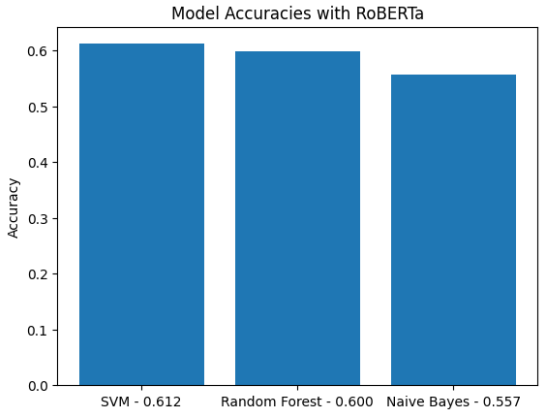

Fig. 5. ML model results with roBERTa tokenizaiton

32. The deep learning model outperformed all other models. However, Machine Learning model were not further.

References

[1] Z. Zeng and S. Bhat, "Getting BART to Ride the Idiomatic Train: Learning to Represent Idiomatic Expressions," Transactions of the Association for Computational Linguistics, vol. 10, pp. 1120–1137, 2022, doi: https://doi.org/10.1162/tacl_a_00510.

[2] [1]Z. Zeng and S. Bhat, "Idiomatic Expression Identification using Semantic Compatibility," Transactions of the Association for Computational Linguistics, vol. 9, pp. 1546–1562, 2021, doi: https://doi.org/10.1162/tacl_a_00442.