# Event classification from the Urdu language text on social media

**Malik Daler Ali Awan** [Corresp., 1], **Nadeem Iqbal kajla** [2], **Amnah Firdous** [3], **Mujtaba Husnain** [4], **Malik Muhammad Saad Missen** [4]

[1] Department of Software Engineering, Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan
[2] Faculty of Computing, Department of Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan
[3] Computer Science and Information Technology, The Govt. Sadiq College and Women University Bahawalpur, Bahawalpur, Punjab, Pakistan
[4] Department of Information Technology, Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan

Corresponding Author: Malik Daler Ali Awan
Email address: daler.ali@iub.edu.pk

Extraction and classification of multiclass events from local languages are challenging tasks because of resource lacking. In this research paper, we presented the event classification for the Urdu language text existing on social media and the news channels. The dataset contains more than 0.1 million (102,962) labeled instances of twelve (12) different types of events. Title, Length, and last-4-words of a sentence are used as features to classify events. The Term Frequency-Inverse Document Frequency (*tf-idf*) showed the best results as a feature vector to evaluate the performance of the six popular machine learning classifiers. Random Forest (RF), Decision Tree, and *k*-Nearest Neighbor out-performed among the other classifiers. Random Forest and K-Nearest Neighbor are the classifiers that out-performed among other classifiers by achieving 98.00% and 99.00% accuracy, respectively.

# Event classification from the Urdu language text on social media

Malik Daler Ali Awan[1], Nadeem Iqbal[2], Amnah Firdous[3], Mujtaba Husnain[4], and Malik Muhammad Saad Missen[5]

[1]The Islamia University Bahawalpur of Bahawalpur, 61300, Pakistan
[2]The Islamia University Bahawalpur of Bahawalpur, 61300, Pakistan
[3]The Government Sadiq College Women University Bahawalpur, 61300, Pakistan
[4]The Islamia University Bahawalpur of Bahawalpur, 61300, Pakistan
[5]The Islamia University Bahawalpur of Bahawalpur, 61300, Pakistan

Corresponding Author:
Daler Ali, The Islamia University of Bahawalpur, Pakistan
Email address: daler.ali@iub.edu.pk

## Abstract

Extraction and classification of multiclass events from local languages are challenging tasks because of resource lacking. In this research paper, we presented the event classification for the Urdu language text existing on social media and the news channels. The dataset contains more than 0.1 million (102,962) labeled instances of twelve (12) different types of events. Title, Length, and last-4-words of a sentence are used as features to classify events. The Term Frequency-Inverse Document Frequency (*tf-idf*) showed the best results as a feature vector to evaluate the performance of the six popular machine learning classifiers. Random Forest (RF), Decision Tree, and *k*-Nearest Neighbor out-performed among the other classifiers. Random Forest and K-Nearest Neighbor are the classifiers that out-performed among other classifiers by achieving 98.00% and 99.00% accuracy, respectively.

## Introduction

In the current digital and innovative era, the text is still the strongest and dominant source of communication instead of pictures, emoji, sounds, and animations [1]. The innovative environment of communication; real-time availability [2] of the Internet and the unrestricted communication mode of social networks have attracted billions of people around the world. Now, people are hooked together via the Internet like a global village. They preferred to share insights about different topics, opinions, views, ideas, and events [3] on social networks in different languages. One of the reasons i.e., Because social media and news channels have created space for local languages [4]. Google input tool[1] provides language transliteration support for more than 88 different languages. Many tools provide the support to use local languages on social media for communication. The google language translator[2] is a platform

[1] https://www.google.com/inputtools/
[2] https://translate.google.com/?hl=en

that facilitates multilingual users of more than 100 languages for conversation. Generally, people prefer to communicate in local languages instead of non-local languages for sake of easiness. A cursive language Urdu is one of the local languages that is being highly adapted for communication. There are more than 300 million [10] Urdu language users all around the world. The Urdu language is a mix-composition of different languages i.e., Arabic, Persian, Turkish, and Hindi [11]. In Pakistan and India, more than 65 million people can speak, understand, and write the Urdu language [12]. It is one of the resource-poor, neglected languages [13] and the national language [14] of Pakistan: the 6$^{th}$ most populous[2] country in the world. Urdu is widely adopted as a second language all over Pakistan [11-14].

In contrast to cursive languages, there exists noteworthy work of information extraction and classification for i.e., English, French, German, and many other non-cursive languages [14-15]. In South Asia other countries [15] i.e., Bangladesh, Iran, and Afghanistan also have a considerable number of Urdu language users. Several tools support the usage of local languages on social media and news channels. Pak Urdu Installer[3]is also one of that software, it supports the Urdu language for textual communication.

Sifting worthy insights from an immense amount of heterogeneous text existing on social media is an interesting and challenging task of Natural Language Processing (NLP). Event extraction and classification is one of those tasks. Event classification insights are helpful to develop various NLP applications i.e., to respond to emergencies, outbreaks, rain, flood, and earthquake [5], etc. People share their intent, appreciation, or criticism [6] i.e., enjoying discount offers by selling brands or criticizing the quality of the product. Earlier awareness of sentimental insights can be helpful to protect from business losses. The implementation of smart- cities possesses a lot of challenges; decision making, event management, communication, and information retrieval. Extracting useful insights from an immense amount of text, dramatically enhance the worth of smart cities [7]. Event information can be used to predict the effects of the event on the community, improve security and rescue the people.

Classification of events can be used to collect relevant information about a specific topic, top-trends, stories, text summarization, and question and answering systems [8-9]. Such information can be used to predict upcoming events, situations, and happening. For example, protesting events reported on social media generally end with conflict among different parties, injuries, death of people, and misuse of resources that cause anarchy. Some proactive measurements can be taken by the state to diffuse the situation and to prevent conflict. Similarly, event classification is crucial to monitor the law-and-order situation of the world.

Extracting and classification of event information from Urdu language text is a unique, interesting, and challenging task. The characteristic features of the Urdu langue that made the event classification tasks more complex and challenging are listed below.

- Cursive nature of the script

---

[2] https://www.worldometers.info/world-population/population-by-country/

[3] http://www.mbilalm.com/download/pak-urdu-installer.php

- Morphologically enriched
- Different structures of grammar
- Right to the left writing style
- No text capitalization

Similarly, the lack of resources i.e., the Part of speech tagger (PoS), words stemmer, datasets, and word annotators are some other factors that made the processing of the Urdu text complex. There exist a few noteworthy works related to the Urdu language text processing (See the literature for more details). All the above-mentioned factors motivated us to explore Urdu language text for our task.

**Concept of Events**

The definition of events varies from domain to domain. In literature, the event is defined in various aspects, such as a verb, adjective, and noun based depending on the environmental situation [16-17]. In our research work event can be defined as "An environmental change that occurs because of some reasons or actions for a specific period." For example, the explosion of the gas container, a collision between vehicles, terrorist attacks, and rainfall, etc. There are several hurdles to process Urdu language text for event classification. Some of them are i.e., determining the boundary of events in a sentence, identifying event triggers, and assigning an appropriate label.

**Event Classification**

"The automated way of assigning predefined labels of events to new instances by using pre-trained classification models is called event classification.". Classification is supervised machine learning; all the classifiers are trained on label instances of the dataset.

**Multiclass Event Classification**

It is the task of automatically assigning the most relevant one class from the given multiple classes. Some serious challenges of multiclassification are sentences overlapping in multiple classes [18-19] and imbalanced instances of classes. These factors generally affect the overall performance of the classification system.

**Lack of Recourse**

The researchers of cursive languages in the past were unexcited and vapid [13] because of lacking resources i.e., dataset, part of speech tagger and word annotators, etc. Therefore, a very low amount of research work exists for cursive language i.e., Arabic, Persian Hindi, and Urdu [20]. But now, from the last few years, cursive languages have attracted researchers. The main reason behind the attraction is that a large amount of cursive language data was being generated rapidly over the internet. Now, some processing tools also have been developed i.e., Part of speech tagger, word stemmer, and annotator that play an important role by making research handier. But these tools are still limited, commercial, and close domain.

Natural language processing is tightly coupled with resources i.e., processing resources, datasets, semantical, syntactical, and contextual information. Textual features i.e., Part of Speech (PoS) and semantic are important for text processing. Central Language of Engineering (CLE)[4] provides limited access to PoS tagger because of the close domain and paid that diverged the researcher to explore Urdu text more easily.

Contextual features [21] i.e., grammatical insight (tense), and sequence of words play important role in text processing. Because of the morphological richness nature of Urdu, a word can be used for a different purpose and convey different meanings depending on the context of contents. Unfortunately, the Urdu language is still lacking such tools that are publicly available for research. Dataset is the core element of research. Dataset for the Urdu language generally exists for name entity extraction with a small number of instances that are

- Enabling Minority Language Engineering (EMILLE) (only 200000 tokens) [22].
- Becker-Riaz corpus (only 50000 tokens) [23]
- International Joint Conference on Natural Language Processing (IJCNLP) workshop corpus (only 58252 tokens)
- Computing Research Laboratory (CRL) annotated corpus (only 55,000 tokens are publicly available data corpora. [24]

There is no specific dataset for events classification for Urdu language text.

**Concept of Our System**

The overall working process of our proposed framework is given in Fig.1.

**Our Contribution**

- In this research article, we claim that we are the first ones who are exploring the Urdu language text to perform multi-class event classification at the sentence level using a machine learning approach,
- A dataset that is larger than state-of-art used in experiments. In our best knowledge classification for twelve 12 different types of events never performed,
- A comprehensive and detailed comparison of six machine learning algorithms is presented to find a more accurate model for event classification for the Urdu language text.

**Our Limitations**

- There is no specific Word2Vec model for Urdu language text,
- There is also no availability of the free (open source) Part of Speech tagger and word stemmer for Urdu language text,
- Also, there exists no publicly available dataset of Urdu language text for sentence classification.

## Related Work

Classification of events from the textual dataset is a very challenging and interesting task of Natural Language Processing (NLP). An intent mining system was developed [6] to facilitate citizens and cooperative authorities using a bag of the token model. The researchers explored the hybrid feature representation for binary classification and multi-label classification. It showed a

[4]http://www.cle.org.pk/

6% to 7% improvement in the top-down feature set processing approach. Intelligence information retrieval plays a vital role in the management of smart cities. Such information helps to enhance security and emergency management capabilities in smart cities [7]. The textual content on social media is explored in different ways to extract event information. Generally, the event has been defined as a verb, noun, and adjective [14]. Event detection is a generic term that is further divided into event extraction and event classification. A combined neural network of the convolutional and recurrent network was designed to extract events from English, Tamil, and Hindi languages. It showed 39.91%, 37.42% and 39.71% F_ Measure [17].

In the past, the researchers were impassive in cursive language, therefore a very limited amount of research work exists in cursive language i.e., Arabic, Persian Hindi, and Urdu [25]. Similarly, in the work of [25], the authors developed a multiple minimal reduct extraction algorithm which is an improved version of the Quick reduct algorithm [26]. The purpose of developing the algorithm is to produce a set of rules that assist in the classification of Urdu sentences. For evaluation purposes, an Arabic-based corpus containing more than 2500 documents was plugged in for classifying them into one of the nine classes. In the experiment, we compared the results of the proposed approach when using multiple and single minimal reducts. The results showed that the proposed approach had achieved an accuracy of 94% when using multiple reducts, which outperformed the single reduct method which achieved an accuracy of 86%. The results of the experiments also showed that the proposed approach outperforms both the K-NN and J48 algorithms regarding classification accuracy using the dataset on hand.

Urdu textual contents were explored [27] for classification using the majority voting algorithm. They categorized Urdu text into seven classes i.e., Health, Business, Entertainment, Science, Culture, Sports, and Wired. They used 21769 news documents for classification and reported 94% precision and recall. Dataset evaluated using these algorithms, Linear SGD, Bernoulli Naïve Bayes, Linear SVM, Naïve Bayes, random forest classifier, and Multinomial Naïve Bayes.

A framework [28] proposed a tweet classification system to rescue people looking for help in a disaster like a flood [29]. The developed system was based on the Markov Model achieve 81% and 87% accuracy for classification and location detection, respectively. The features used in their system are [29]:

- Number of words in a tweet (w)
- Verb in a tweet by (verb)
- Number of verbs in a tweet by (v)
- Position of the query by (Pos)
- Word before query word (before)
- Word after query word (after)

To classify Urdu news headlines [30] by using maximum indexes of vectors. They used stemmed and non-stemmed textual data for experiments. The system was specifically designed for text classification instead of event classification. The proposed system achieved 78.0% for competitors and 86.6% accuracy for the proposed methodology. In comparison, we used sentences of Urdu language for classification and explored the textual features of sentences. We have explored all the textual and numeric features i.e., title, length, last-4-words, and the combinations of these (for more detail see Tab. 1) in detail in this paper that were not reported ever in state-of-art according to our knowledge.

Twitter [31] to detect natural disasters i.e., bush fires, earthquakes and cyclones, and humanitarian crises [32]. To be aware of emergencies situation in natural disasters a framework work designed based on SVM and Naïve Bayes classifiers using word unigram, bi-gram, length, number of #Hash tag, and reply. These features were selected on a sentence basis. SVM and Nave Bayes showed 87.5% and 86.2% accuracy respectively for tweet classification i.e., seeking help, offering for help, and none. A very popular social website (Twitter) textual data was used [33] to extract and classify events for the Arabic language. Implementation and testing of Support Vector Machine (SVM) and Polynomial Network (PN) algorithms showed promising results for tweet classification 89.2% and 92.7%. Stemmer with PN and SVM magnified the classification 93.9% and 91.7% respectively. Social events [34] were extracted assuming that to predict either parties or one of them aware of the event. The research aimed to find the relation between related events. Support Vector Machine (SVM) with kernel method was used on adopted annotated data of Automated Content Extraction (ACE). Structural information derived from the dependency tree and parsing tree are utilized to derive new structures that played important role in event identification and classification. The Tweet classification of the tweets related to the US Air Lines [40] is performed by the sentiment analysis companies that are not related to our work. We tried to classify events at sentence level that is challenging since the Urdu sentence contains very short features as compared to a tweet. It is pertinent to mention that the sentiment classification is different from the event classification. Multiclass event classification is reported [41] comprehensively, deep learning classifiers are used to classify events into different classes.

## Materials & Methods

Event classification for Urdu text is performed using a supervised machine learning approach. A complete overview of the multi-class event classification methodology is given in Fig.1. Textual data classification possesses a lot of challenges i.e., word similarity, poor grammatical structure, misuse of terms, and multilingual words. That is the reason, we decided to adopt a supervised classification approach to classify Urdu sentences into different categories.

### Data Collection

Urdu data were collected from popular social networks (Twitter), famous news channel blogs i.e., Geo News[5], Urdu Point[6] , and BBC Urdu[7]. The data collection consists of the title, the main body, the published date, the location, and the URL of the post. In the phase of data collection, a PHP-based web scraper is used to crawl data from the above-mentioned social websites. A complete post is retrieved from the websites and stored in MariaDB (database). Our dataset consists of 0.1 million (102, 960) label sentences of different types of events. All the different

[5] https://urdu.geo.tv/
[6] https://www.urdupoint.com/daily/
[7] https://www.bbc.com/urdu

types of events used in our research work and their maximum number of instances are shown below in Fig. 2.

There are twelve different types of events that we try to classify in our research work. These events are a factual representation of the state and the situation of the people. In Fig. 2. imbalances number of instances of each event are given. It can be visualized that politics, sports, and Fraud & Corruption have a higher number of instances while Inflation, Sexual Assault, and Terrorist attacks have a lower number of instances. These imbalanced numbers of instances made our classification more interesting and challenging.

Multiclass events classification tasks are comprised of many classes. The different types of events that are used in our research work i.e., sports, Inflation, Murder & Death, Terrorist attacks, Politics, Law and Order, Earthquake, Showbiz, Fraud & Corruption, Weather, Sexual Assault, and Business. All the sentences of the dataset are labeled by the above-mentioned twelve (12) different types of events. Finally, a numeric (integer) value is assigned to each type of event label (See Tab. 2 for more details of the label and its relevant numeric value).

**Preprocessing**

The initial preprocessing steps are performed on the corpus to prepare it for machine learning algorithms. Because textual data cannot directly process by machine learning classifiers. It also contains many irrelevant words. The detail of all the preprocessing steps is given below. These steps were implemented in a PHP-based environment. While the words tokenization is performed using the scikit library [20] in python.

**Post Splitting**

The PHP crawler extracted the body of the post. It comprises many sentences as a paragraph. In the Urdu language script, sentences end with a sign called "- "Hyphen (Khatma-تمہ). It is a standard punctuation mark in the Urdu language to represent the end of the sentence. As mentioned earlier, we are performing event classification at the sentence level. So, we split paragraphs of every post into sentences. Every line in the paragraphs ending at Hyphen is split as a single line.

**Stop Words Elimination**

Generally, those words that occur frequently in text corpus are considered as stop words. These words merely affect the performance of the classifier. Punctuation marks ("!", "@"," #", etc.) and frequent words of the Urdu languages (کا(ka), کے (kay), کی(ki), etc.) are the common examples of stop words. All the stop words [28] that do not play an influential role in event classification for the Urdu language text are eliminated from the corpus. Stop words elimination reduces memory and processing utilization and makes the processing efficient.

**Noise Removal and Sentences Filtering**

Our data were collected from different sources (see section 3). It contains a lot of noisy elements i.e., multilanguage words, links, mathematical characters, and special symbols, etc. To clean the corpus, we removed noise i.e., multilingual sentences, irrelevant links, and special characters. The nature of our problem confined us to define the limit of words per sentence. Because of the multiple types of events, it is probably hard to find a sentence of the same length. We decided to keep the maximum number of sentences in our corpus. All those sentences which are brief and extensive are removed from our corpus. In our dataset lot of sentences varying in length from 5 words to 250 words. We decided to use sentences that consist of 5 words to 150 words to lemmatize our research problem and to reduce the consumption of processing resources.

**Sentence Labeling**

In supervised learning, providing output (Label) detail in the corpus is a core element. Sentence labeling is an exhausting task that requires deep knowledge and an expert's skill of language. All the sentences were manually labeled by observing the title of the post and body of sentences by Urdu language experts (see Tab. 2 for sentence labeling). Three Urdu language experts were engaged in the task of sentence labeling. One of them is Ph.D. (Scholar) while the other two are M.Phil. To our best knowledge, it is the first largest labeled dataset for the multi-class event in the Urdu language.

**Feature Selection**

The performance of prediction or classification models is cohesively related to the selection of appropriate features. In our dataset six (6) features excluding "Date" as a feature are considered valuable to classify Urdu news sentences into different classes. All the proposed features that are used in our research work are listed in Tab.1.

**Why were these features selected?**

**Last- 4-Words of Sentence**

Occurrence, happening, and situations are generic terms that are used to represent events. In general, "verb" represents an event. The grammatical structure of Urdu language is Subject_ Object_ Verb (SOV) [31], which depicts that verb, is laying in the last part of the sentences. For example, the sentence ("احمد نے پودوں کو پانی دیا۔ – Ahmad ney podon ko pani dia"), (Ahmad watered the plants) follows the SOV format. "Pani dia-پانی دیا" is the verbal part of the sentence existing in the last two words of the sentence. It shows the happening or action of the event. Our research problem is to classify sentences into different classes of events. So, that last_4_ words are considered one of the vital features to identify events and non-event sentences. For example, in Tab. 3 in the event column underline/highlighted part of the sentence represents the happening of an event i.e., last_4_words in the sentence. While labeling the sentences we are strictly concerned that only event sentences of different types should be labeled.

**Title of Post**

Every conversation has a central point i.e., title. Textual, pictorial, or multimedia content that is posted on social networks as a blog post, at the paragraph level or sentence level describes the specific event. Although many posts contain irrelevant titles to the body of the message. However, using the title as a feature to classify sentences is crucial because the title is assigned to the contents-based material.

**Length of Sentence**

A sentence is a composition of many words. The length of the sentence is determined by the total number of words or tokens that exist in it. It can be used as a feature to classify sentences because many sentences of the same event have probably the same length.

**Title and Length**

The proposed feature is the combination of the title of the post and the length of the sentence. The title represents the central idea of the post, and the length of the sentence varies from title to title.

**Title and Last-4-words**

The combination of title and last_4_words in Urdu language text is very helpful to classify the sentences. Because last_4_words generally represent the occurrence/happening of some event.

**Length and Last-4-words**

We also consider the combination of length with last_4_words as a valuable feature because the length of a sentence varies from event to event.

**Features Engineering**

Feature Engineering is a way of generating specific features from a given set of features and converting selected features to machine-understandable format. Our dataset is text-based that consists of more than 1 million (102,960 labeled) instances i.e., sports, inflation, death, terrorist attack, and sexual assault, etc. 12 classes.

As mentioned earlier that the Urdu language is one of the resource-poor languages and since there are no pre-trained word embedding models to generate the embedding vectors for Urdu language text, we could not use the facility of Word2Vec embedding technique.

All the textual features are converted to numeric format i.e., (Term Frequency_ Inverse Document Frequency) TF_IDF and Count-Vectorizer. These two features TF_IDF and Count-Vectorizer are used in a parallel fashion. The scikit-learn package is used to transform text data into numerical value [20].

**Count_ Vectorization**

The process of converting words to numerical form is called vectorization. Its working strategy is based on term frequency. It counts the frequency of specific word w and builds the spare matrix-vector using bag-of-words (BOW). The length of the feature vector depends on the size of the bag-of-words i.e., dictionary.

**Term Frequency Inverse Document Frequency**

It is a statistical measure of word w to understand the importance of that word for specific document d in the corpus. The importance of a word is proportionally related to frequency i.e., higher frequency more important. The mathematical formulas related to TF_IDF are given below:

$$\text{Term Frequency (TF)} = \frac{Number\ of\ time\ term\ t\ appears\ in\ document}{Total\ number\ of\ terms\ in\ documents} \quad (1)$$

$$\text{Inverse Document Frequency (IDF)} = Log_e \frac{Total\ number\ \ of\ document}{Total\ number\ of\ documents\ term\ t\ appears} \quad (2)$$

$$\text{TF\_IDF} = TF * IDF \quad (3)$$

## Experimental Setup

Classifiers are the algorithms used to classify data instances into predefined categories. Many classifiers exist that process the textual data using a machine learning approach. In our research work, we selected the six most popular machine learning algorithms i.e., Random Forest (RF)

[10], K-Nearest Neighbor (KNN), Support Vector Machine (SVM, Decision Tree (DT), Naïve Bayes Multinomial (NBM), and Linear Regression (LR).

**Machine Learning Classifiers**

In this section, we presented the detail of six classifiers that were used to classify the Urdu sentences using different proposed features.

**1 Random Forest (RF)**

This model is comprised of several decision trees that act as a building block of RF. Every decision tree is created using the rules i.e., if then else, and the conditional statements, etc. [10]. These rules are then followed by the multiple decision trees to analyze the problem at a discrete level.

**2 *k*-Nearest Neighbor**

It is one of the statistical models that find the similarity among the data points using Euclidean distance [35]. It belongs to the category of lazy classifiers and is widely used for classification and regression tasks.

**3 Support Vector Machine**

It is based on statistical theory [36], to draw a hyperplane among points of the dataset. It is highly recommended for regression and classification i.e., binary classification, multiclass classification, and multilabel classification. It finds the decision boundary to identify different classes and maximize the margin.

**4 Decision Tree**

It is one of the supervised classifiers that work following certain rules. Data points/inputs are split according to the specific condition [37]. It is used for regression and classification using the non-parametric method because it can handle textual and numerical data. Learning from data points is accomplished by approximating the sine curve with the combination of an if-else-like set of rules. The accuracy of a model is related to the deepness and complexity of rules.

**5 Naïve Bayes Multinominal**

It is a computationally efficient classifier for text classification using discrete features. It can also handle the textual data by converting it into numerical [38] format using count vectorizer and term frequency-inverse document frequency (tf- idf).

**6 Linear Regression**

It is a highly recommended classifier for numerical output. It is used to perform prediction by learning linear relationships between independent variables (inputs) and dependent variables (output) [39].

**Training Dataset**

A subpart of the dataset that is used to train the models to learn the relationship among dependent and independent variables is called the training dataset. We divided our data into training and testing using the train_ test_ split function of the scikit library using python. Our training dataset consists of 70% of the dataset that is more than 70,000 labeled sentences of Urdu language text.

**Testing Dataset**

It is also the subpart of the dataset that is usually smaller than size as compared to the training dataset. In our research case, we decided to use 30% of the dataset for testing and validating the performance of classifiers. It comprises more than 30,000 instances/sentences of Urdu langue text.

**Performance Measuring Parameters**

The most common performance measuring parameters [14-18] i.e., precision, recall, and F1_measure are used to evaluate the proposed framework since these parameters are the key indicators while performing the classification in a multiclass environment using an imbalanced dataset.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

## Results

To evaluate our dataset, the Python package scikit-learn is used to perform event classification at the sentence level. We extracted the last-4-words of each sentence and calculated the length of each sentence. To obtain the best classification results we evaluated six machine learning classifiers among others i.e., Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), *k*-Nearest Neighbor, and Naïve Bayes Multinominal (NBM).

We proposed three features i.e., Length, Last-4-words, and Length and Last-4-words to classify sentences into different types of events (see tab. 2). The results were obtained using 'length ' as the feature is shown in Tab. 4. The classifiers i.e., DT, RF, NBM, and LR showed 32% accuracies that is very low. The comparatively second feature that is Last-4-words showed better results for these above-mentioned classifiers. Random Forest showed 52% accuracy that is a considerable result as an initiative for multiclass event classification in the Urdu language text. The detail of results regarding other classifiers can be seen in Tab 5.

We also evaluated these classifiers using another feature that is the combination of both Length and Last-4-grams. It also improved the overall 1% accuracy of the proposed system. The Random Forest showed 53.00% accuracy. The further details of accuracies of other used machine learning models can be seen in Tab.6

The results obtained by using the above features are very low, we deiced to use the title of the post as a feature to improve the performance of the system. We integrated the "Title" of the post with each sentence of the same paragraph that dramatically improves the accuracy of the system. We combined the "Title" of the post with other features i.e., length, and Last-4-words. The detail of the highest accuracies that is obtained by the combination of these features i.e., Last-4-words, length, and title are given in Tab. 7 and Tab. 8. Random forest and *k*-NN showed the highest accuracies. The detail of the confusion matrix related to the proposed system (TP, FP, TN, FN) is also given in Tab. 9 and Tab. 10.

The standard performance measuring parameters i.e., precision, recall, and f1-measure of Random Forest and *k*-NN classifiers using "Title and Last-4words" as features are given in Tab. 11 and Tab. 12 respectively. Similarly other combinations of features i.e., "Title and Length" are used to enhance the accuracy of the system. The Decision Tree and Random Forest showed the highest results as compared to other classifiers for this specific combination of features. A detailed summary of the results related to Decision Tree and Random Forest is given in Tab. 13 and Tab. 14 respectively.

We finally presented the comparison of four classifiers that showed the highest results in fig. 3. The semantics of the script written in the Urdu language is quite different from that of English and Arabic Language which causes the low performance of SVM and k-NN as compared to Random Forest.

## Discussion

Event extraction and classification are tightly coupled with processing resources i.e., Part of speech tagger (PoS), Text annotators, and contextual insights. Usage of local languages being highly preferred over social media. Urdu is one of those languages that have a considerable number of users and a huge bulk of data on social networks. The evaluation reports obtained after analyzing multiple features i.e., Length, Last-4-words, Title, and combination of all these features converged our findings to conclude that length and last-4-words are basic features to classify multiclass events but showed 53% accuracy. To improve the accuracy of the proposed system, we integrated "Title" as the feature with other two features i.e., Length and Last-4-words. The combination of "Title" with "Length and Last-4-words" improved the performance of the proposed system and showed the highest results.

Furthermore, extracting and classification of events from resource-poor language is an interesting and challenging task. There are no standard (benchmark) datasets and word embedding models like Word2Vec or Glove (Exists for the English Language) for Urdu language text.

## Conclusions

A massive amount of Urdu textual data exists on social networks and news websites. Multiclass event classification for Urdu text at the sentence level is a challenging task because of the few numbers of words and limited contextual information. We performed experiments by selecting appropriate features i.e., length, last-4-words, and combination of both length and last-4-words. These are the key features to achieve our expected results. Count_ Vectorizer and TF-IDF feature generating techniques are used to convert text into (numeric) real value for machine learning models. Random Forest classification model showed 52% and 53% accuracy for Last-4-words and combination of length and last-4-words.

The title is the key feature that can dramatically improve the performance of event classification models that works on a sentence level.

## Future Work

- In a comprehensive review of Urdu literature, we found a few numbers of referential works related to Urdu text processing. One of the main issues associated with the Urdu language research is the unavailability of the appropriate corpus like the data set of Urdu sentences representing the event; the close-domain PoS tagger; the lexicons, and the annotator, etc.
- There is a need to develop the supporting tools i.e., the PoS tagger, the annotation tools, the dataset of the Urdu-based languages having information about some information associated with the events, and the lexicons can be created to extend the research areas in the Urdu language.
- In the future, many other types of events and other domains of information like medical events, social, local, and religious events can be classified using the extension of machine learning i.e., deep learning.
- In the future grammatical, contextual, and lexical information can be used to categorize events. Temporal information related to events can be further utilized to classify an event as real and retrospective.
- Classification of events can be performed at the document level and phrase level.
- Deep learning classifiers can be used for a higher number of event classes.

## Acknowledgements

I am very grateful to the Chairman of the Faculty of Computing and Head of Department of Software Engineering of Islamia University Bahawalpur, Pakistan. They encouraged and guided me to make research work more interesting.

## References

[1] Lenhart, A., Ling, R., Campbell, S., & Purcell, K. (2010). Teens and mobile phones: Text messaging explodes as teens embrace it as the centerpiece of their communication strategies with friends. Pew Internet & American Life Project.

[2] Motoyama, M., Meeder, B., Levchenko, K., Voelker, G. M., & Savage, S. (2010). Measuring Online Service Availability Using Twitter. WOSN, 10, 13-13.

[3] Reuter, T., & Cimiano, P. (2012, June). Event-based classification of social media streams. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (pp. 1-8).

[4] Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., & Laredo, J. A. (2013). CrisisTracker: Crowdsourced social media curation for disaster awareness. IBM Journal of Research and Development, 57(5), 4-1.

[5] Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. IEEE Annals of the History of Computing, 27(06), 52-59.

[6] Purohit, H., Dong, G., Shalin, V., Thirunarayan, K., & Sheth, A. (2015, December). Intent classification of short-text on social media. In 2015 ieee international conference on smart city/socialcom/sustaincom (smartcity) (pp. 222-228). IEEE.

[7] Alkhatibl, M., El Barachi, M., & Shaalan, K. (2018, June). Using Arabic Social Media Feeds for Incident and Emergency Management in Smart Cities. In 2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech) (pp. 1-6). IEEE.

[8] Khan, W., Daud, A., Nasir, J. A., & Amjad, T. (2016). Named entity dataset for urdu named entity recognition task. Organization, 48, 282.

[9] Jacobs, I. S. (1963). Fine particles, thin films and exchange anisotropy. Magnetism, 271-350.

[10] Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. ECE591Q Machine Learning Journal Paper, 1-13.

[11] Ghulam, S. M., & Soomro, T. R. (2018, March). Twitter and Urdu. In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-6). IEEE.

[12] Naz, M., & Hussain, S. (2013, December). Binarization and its evaluation for Urdu Nastalique document images. In INMIC (pp. 213-218). IEEE.

[13] Mukund, S., Srihari, R., & Peterson, E. (2010). An Information-Extraction System for Urdu---A Resource-Poor Language. ACM Transactions on Asian Language Information Processing (TALIP), 9(4), 1-43.

[14] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

[15] Riaz, K. (2008, October). Concept search in Urdu. In Proceedings of the 2nd PhD workshop on Information and Knowledge Management (pp. 33-40).

[16] Ramesh, D., & Kumar, S. S. INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY EVENT EXTRACTION FROM NATURAL LANGUAGE TEXT.

[17] Ahmed, K., Ali, M., Khalid, S., & Kamran, M. (2016). Framework for Urdu News Headlines Classification. Journal of Applied Computer Science & Mathematics, (21).

[18] Kong, X., Shi, X., & Yu, P. S. (2011, April). Multi-label collective classification. In Proceedings of the 2011 SIAM International Conference on Data Mining (pp. 618-629). Society for Industrial and Applied Mathematics.

[19] Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, 53, 196-207.

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[21] Vosoughi, S., Zhou, H., & Roy, D. (2016). Enhanced twitter sentiment classification using contextual information. arXiv preprint arXiv:1605.05195.

[22] Baker, P., Hardie, A., McEnery, T., & Jayaram, B. D. (2003, April). Corpus data for South Asian language processing. In Proceedings of the 10th Annual Workshop for South Asian Language Processing, EACL.
[23] Becker, D., & Riaz, K. (2002). A study in urdu corpus construction. In COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization.
[24] Malik, M. K. (2017). Urdu named entity recognition and classification system using artificial neural network. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(1), 1-13.
[25] Alsaedi, N., & Burnap, P. (2015, April). Arabic event detection in social media. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 384-401). Springer, Cham.
[26] Al-Radaideh, Q. A., & Al-Abrat, M. A. (2019). An Arabic text categorization approach using term weighting and multiple reducts. Soft Computing, 23(14), 5849-5863.
[27] Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: a survey. Artificial Intelligence Review, 47(3), 279-311.
[28] Kuila, A., chandra Bussa, S., & Sarkar, S. (2018). A Neural Network based Event Extraction System for Indian Languages. In FIRE (Working Notes) (pp. 291-301).
[29] Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2019). Event classification and location prediction from tweets during disasters. Annals of Operations Research, 283(1), 737-757.
[30] Ali, A. R., & Ijaz, M. (2009, December). Urdu text classification. In Proceedings of the 7th international conference on frontiers of information technology (pp. 1-7).
[31] Agarwal, A., & Rambow, O. (2010, October). Automatic detection and classification of social events. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 1024-1034).
[32] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860).
[33] Hussain, S. (2008). Resources for Urdu language processing. In Proceedings of the 6th workshop on Asian Language Resources.
[34] Usman, M., Shafique, Z., Ayub, S., & Malik, K. (2016). Urdu text classification using majority voting. Int. J. Adv. Comput. Sci. Appl, 7(8), 265-273.
[35] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.
[36] Zhang, Y. (2012, September). Support vector machine classification algorithm and its application. In International Conference on Information Computing and Applications (pp. 179-186). Springer, Berlin, Heidelberg.
[37] Zhong, Y. (2016, August). The analysis of cases based on decision tree. In 2016 7th IEEE international conference on software engineering and service science (ICSESS) (pp. 142-147). IEEE.

[38] Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, 44(1), 48-59.
[39] Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. Information retrieval, 4(1), 5-31.
[40] Rustam, F., Ashraf, I., Mehmood, A., Ullah, S., & Choi, G. S. (2019). Tweets classification on the base of sentiments for US airline companies. Entropy, 21(11), 1078.
[41] Ali, D., Missen, M. M. S., & Husnain, M. (2021). Multiclass Event Classification from Text. Scientific Programming, 2021.

## Table 1 (on next page)

Proposed Features

**Table 1:** Proposed Features

| **Sr. No.** | **Feature _Name** |
|---|---|
| 1 | Length |
| 2 | Last_4_ words |
| 3 | Last_4_words and Length |
| 4 | Title |
| 5 | Title and Length |
| 6 | Title and Last_4-words |

## Table 2(on next page)

Types of events and their labels in the dataset

**Table 2:** Types of events and their labels in the dataset

| Event | Label | Event | Label |
|---|---|---|---|
| Sports | 1 | Earthquake | 7 |
| Inflation | 2 | Showbiz | 8 |
| Murder and Death | 3 | Fraud and Corruption | 9 |
| Terrorist Attack | 4 | Rain/Weather | 10 |
| Politics | 5 | Sexual Assault | 11 |
| Law and Order | 6 | Business | 12 |

# Table 3 (on next page)

Last 4-words representing an event

**Table 3:** Last 4-words representing an event

| Event | | Non_ Event | |
|---|---|---|---|
| Urdu | English | Urdu | English |
| مسئلہ کشمیر کو لے کر پاکستان اور بھارت میں <u>جنگ چھڑ چکی ہے</u>۔ | The battle between Pakistan and India has been started on the conflict of Kashmir. | چند دن پہلے لوگ خوش تھے۔ | Few days ago, people were happy. |

## Table 4(on next page)

Length

**Table 4:** Length

| Algorithms | Accuracy | Feature |
|---|---|---|
| SVM | 17% | Length |
| NBM | 32% | |
| LR | 32% | |
| Decision Tree | 32% | |
| Random Forest | 32% | |
| K-NN | 24% | |

## Table 5(on next page)

Last _4_words accuracy

**Table 5:** Last _4_words accuracy

| Algorithms | Accuracy | Feature |
|---|---|---|
| SVM | 45% | Last _4_words |
| NBMN | 44% | |
| LR | 49% | |
| Decision Tree | 49% | |
| **Random Forest** | **52%** | |
| K-NN | 48% | |

## Table 6 (on next page)

Last _4_words and Length Accuracy

**Table 6:** Last _4_words and Length Accuracy

| Algorithms | Accuracy | Feature |
|---|---|---|
| SVM | 46% | Length and Last _4_words |
| NBMN | 44% | |
| LR | 49% | |
| Decision Tree | 48% | |
| **Random Forest** | **53%** | |
| K-NN | 49% | |

# Table 7(on next page)

Title and Last _4_words accuracy

**Table 7:** Title and Last _4_words accuracy

| Algorithms | Accuracy | Feature |
|---|---|---|
| SVM | 85% | Title and Last _4_words |
| NBMN | 91% | |
| LR | 95% | |
| Decision Tree | 97% | |
| **Random Forest** | **98%** | |
| **K-NN** | **99%** | |

## Table 8(on next page)

Title and Length

**Table 8:** Title and Length

| Algorithms | Accuracy | Feature |
|---|---|---|
| SVM | 87% | Title and Length |
| NBMN | 93% | |
| LR | 98% | |
| **Decision Tree** | **99%** | |
| **Random Forest** | **99%** | |
| K-NN | 94% | |

## Table 9 (on next page)

Random forest TP, FN, FP and TN

**Table 9:** Random forest TP, FN, FP and TN

| | Random Forest | | | | |
|---|---|---|---|---|---|
| Label | Type of Event | TP | FN | FP | TN |
| 1 | Sports | 5646 | 15 | 14 | 25514 |
| 2 | Inflation | 967 | 0.0 | 08 | 30211 |
| 3 | Murder and Death | 2096 | 19 | 22 | 29052 |
| 4 | Terrorist Attack | 865 | 13 | 06 | 30304 |
| 5 | Politics | 9983 | 47 | 86 | 21073 |
| 6 | law and order | 2257 | 36 | 23 | 28872 |
| 7 | Earthquake | 970 | 0.0 | 0.0 | 30219 |
| 8 | Showbiz | 2244 | 15 | 04 | 28929 |
| 9 | Fraud and corruption | 3015 | 29 | 21 | 35924 |
| 10 | Rain/weather | 1031 | 0.0 | 05 | 34888 |
| 11 | Sexual Assault | 889 | 0.0 | 01 | 30300 |
| 12 | Business | 1032 | 20 | 04 | 30134 |

## Table 10 (on next page)

KNN TP, FN, FP and TN

**Table 10:** KNN TP, FN, FP and TN

| | K-Nearest Neighbor | | | | |
|---|---|---|---|---|---|
| Label | Type of Event | TP | FN | FP | TN |
| 1 | Sports | 5638 | 23 | 34 | 25494 |
| 2 | Inflation | 967 | 0.0 | 29 | 30139 |
| 3 | Murder and Death | 2077 | 38 | 32 | 29044 |
| 4 | Terrorist Attack | 858 | 20 | 21 | 30308 |
| 5 | Politics | 9931 | 99 | 98 | 21052 |
| 6 | law and order | 2238 | 55 | 42 | 28854 |
| 7 | Earthquake | 970 | 0.0 | 07 | 30219 |
| 8 | Showbiz | 2242 | 17 | 21 | 28908 |
| 9 | Fraud and corruption | 3023 | 21 | 13 | 28121 |
| 10 | Rain/weather | 1031 | 0.0 | 26 | 10145 |
| 11 | Sexual Assault | 889 | 0.0 | 11 | 30293 |
| 12 | Business | 1001 | 51 | 04 | 30133 |

## Table 11(on next page)

Random Forest performance using the title, and last _4_words

**Table 11:** Random Forest performance using the title, and last _4_words

| Label | Event | Precision | Recall | F1_Measure |
|---|---|---|---|---|
| 1 | Sports | 0.99 | 0.99 | 0.99 |
| 2 | Inflation | 0.99 | 1.00 | 0.99 |
| 3 | Murder and Death | 0.98 | 0.99 | 0.98 |
| 4 | Terrorist Attack | 0.97 | 0.96 | 0.97 |
| 5 | Politics | 0.98 | 0.99 | 0.98 |
| 6 | law and order | 0.98 | 0.96 | 0.97 |
| 7 | Earthquake | 1.00 | 1.00 | 1.00 |
| 8 | Showbiz | 0.99 | 0.98 | 0.99 |
| 9 | Fraud and corruption | 0.99 | 0.98 | 0.98 |
| 10 | Rain/weather | 1.00 | 1.00 | 1.00 |
| 11 | Sexual Assault/Intercourse | 1.00 | 1.00 | 1.00 |
| 12 | Business | 0.98 | 0.95 | 0.97 |
| Overall accuracy | 98.53% | | | |

## Table 12 (on next page)

KNN performance using the title, and last _4_words

**Table 12:** KNN performance using the title, and last _4_words

| Label | Event | Precision | Recall | F1_Measure |
|---|---|---|---|---|
| 1 | Sports | 0.99 | 1.00 | 0.99 |
| 2 | Inflation | 0.97 | 1.00 | 0.99 |
| 3 | Murder and Death | 0.99 | 0.98 | 0.98 |
| 4 | Terrorist Attack | 0.98 | 0.98 | 0.98 |
| 5 | Politics | 0.99 | 0.99 | 0.99 |
| 6 | law and order | 0.98 | 0.98 | 0.98 |
| 7 | Earthquake | 1.00 | 1.00 | 1.00 |
| 8 | Showbiz | 0.99 | 0.99 | 0.99 |
| 9 | Fraud and corruption | 0.99 | 0.99 | 0.99 |
| 10 | Rain/weather | 0.99 | 1.00 | 0.99 |
| 11 | Sexual Assault/Intercourse | 0.99 | 1.00 | 1.00 |
| 12 | Business | 1.00 | 0.95 | 0.97 |
| Overall accuracy | 98.96% | | | |

## Table 13 (on next page)

Decision Tree performance using the ‘Title and Length’

**Table 13:** Decision Tree performance using the 'Title and Length'

| Label | Event | Precision | Recall | F1_Measure |
|---|---|---|---|---|
| 1 | Sports | 1.00 | 1.00 | 1.00 |
| 2 | Inflation | 1.00 | 1.00 | 1.00 |
| 3 | Murder and Death | 0.99 | 0.99 | 0.99 |
| 4 | Terrorist Attack | 0.99 | 0.99 | 0.99 |
| 5 | Politics | 1.00 | 1.00 | 1.00 |
| 6 | law and order | 0.99 | 1.00 | 0.99 |
| 7 | Earthquake | 1.00 | 1.00 | 1.00 |
| 8 | Showbiz | 1.00 | 0.99 | 1.00 |
| 9 | Fraud and corruption | 1.00 | 0.99 | 1.00 |
| 10 | Rain/weather | 1.00 | 1.00 | 1.00 |
| 11 | Sexual Assault/Intercourse | 1.00 | 1.00 | 1.00 |
| 12 | Business | 1.00 | 0.98 | 0.99 |
| Overall accuracy | 99.63% | | | |

## Table 14(on next page)

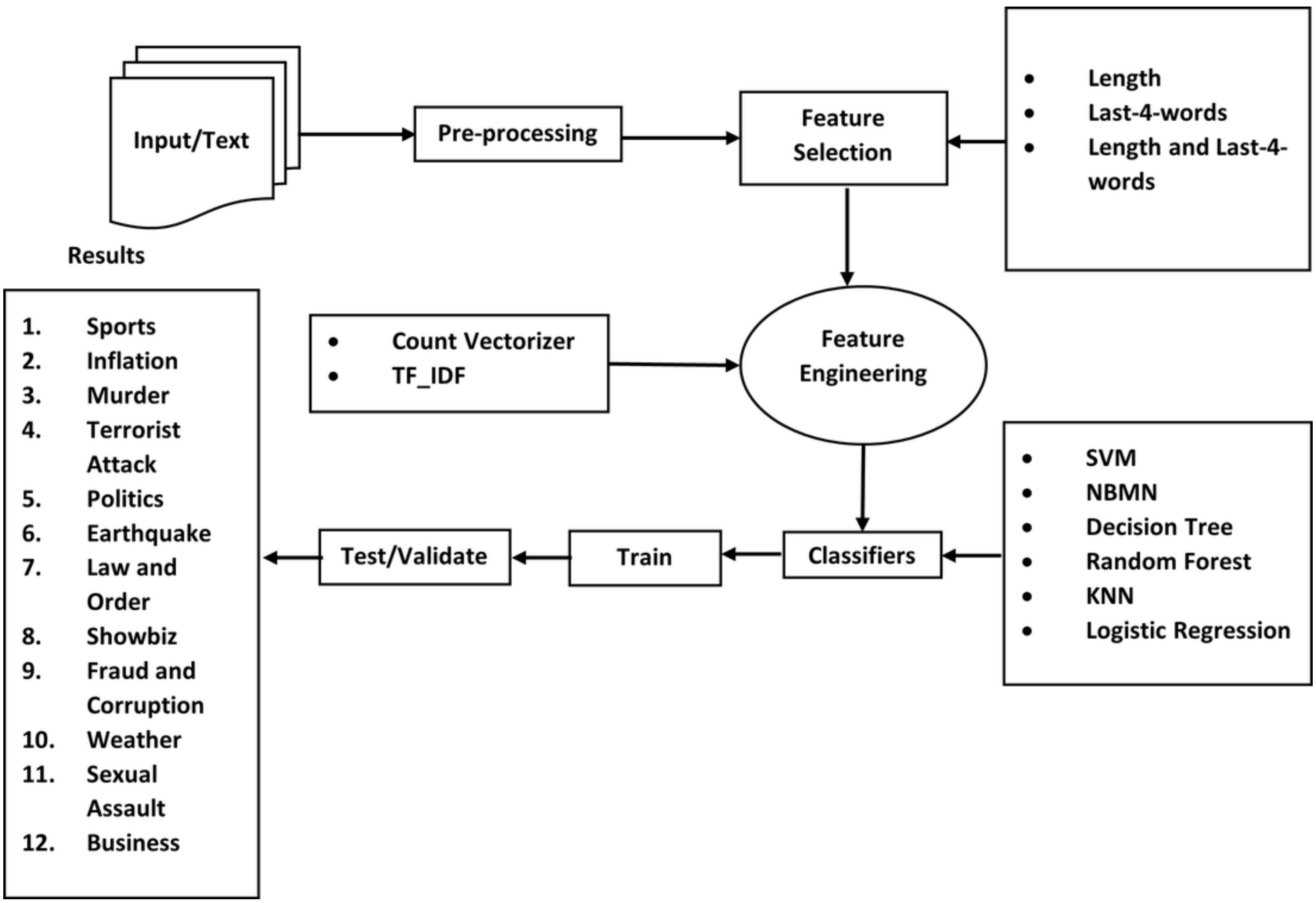
Random Forest performance using the ‘Title and Length’

1 **Table 14:** Random Forest performance using the 'Title and Length'

| Label | Event | Precision | Recall | F1_Measure |
|---|---|---|---|---|
| 1 | Sports | 1.00 | 1.00 | 1.00 |
| 2 | Inflation | 1.00 | 1.00 | 1.00 |
| 3 | Murder and Death | 1.00 | 1.00 | 1.00 |
| 4 | Terrorist Attack | 1.00 | 0.99 | 1.00 |
| 5 | Politics | 1.00 | 1.00 | 1.00 |
| 6 | law and order | 1.00 | 1.00 | 1.00 |
| 7 | Earthquake | 1.00 | 1.00 | 1.00 |
| 8 | Showbiz | 1.00 | 1.00 | 1.00 |
| 9 | Fraud and corruption | 1.00 | 1.00 | 1.00 |
| 10 | Rain/weather | 1.00 | 1.00 | 1.00 |
| 11 | Sexual Assault/Intercourse | 1.00 | 1.00 | 1.00 |
| 12 | Business | 1.00 | 1.00 | 1.00 |
| Overall accuracy | 99.92% | | | |

# Figure 1

Concept Diagram



# Figure 2

Maximum Number of Instances

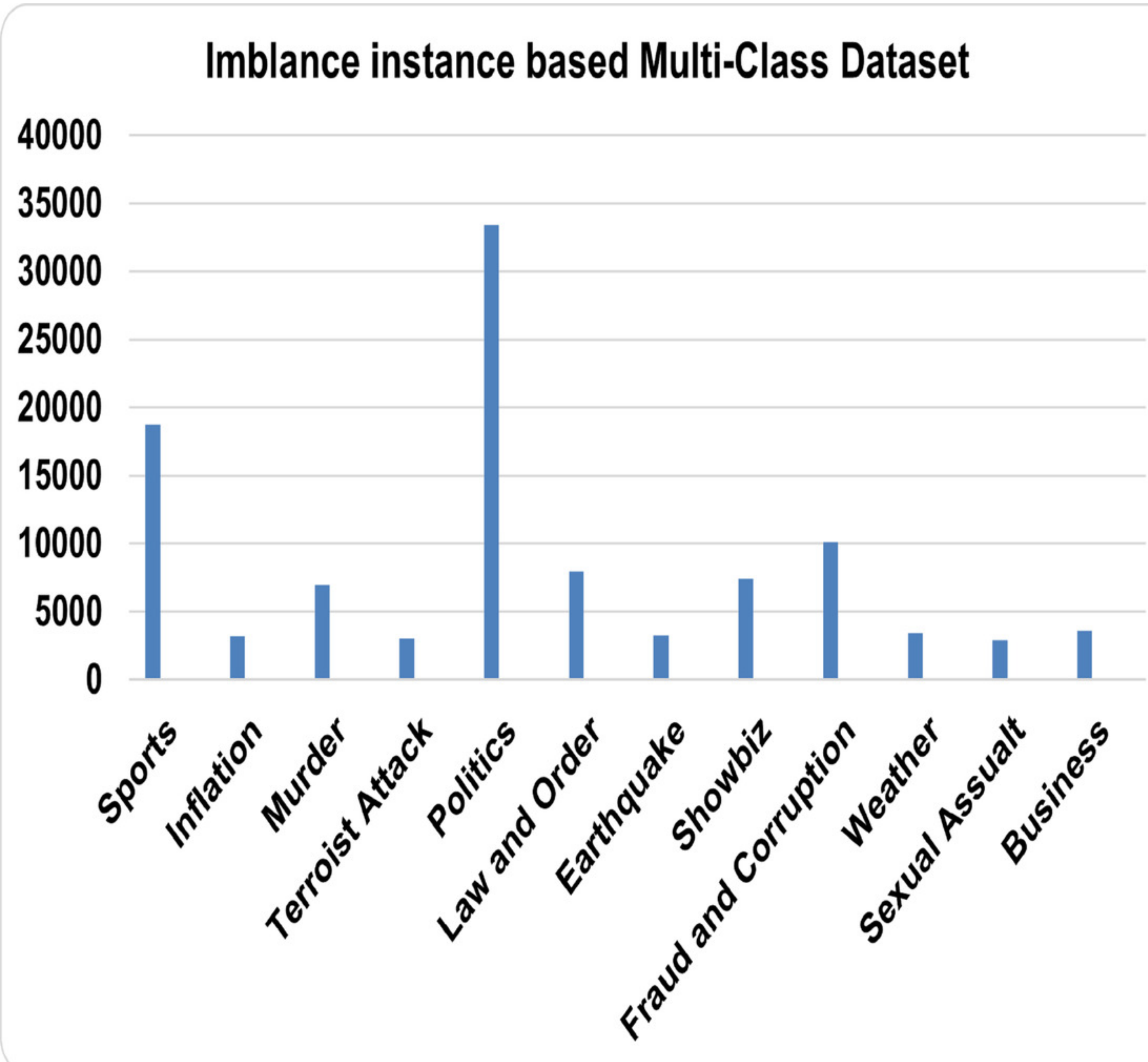



**Figure 2:** Maximum number of instances

# Figure 3

The Best features and the best classi ers