Turkish Journal of Computer and Mathematics Education Vol.12 No.10 (2021), 3039-3054

Research Article

Banking news-events representation and classification with a novel hybrid model using DistilBERT and rule-based features

Varun Dogra[a], Sahil Verma[b,*], Aman Singh[a], Kavita[b], M N Talib[c], Mamoona Humayun[d]

[a] Lovely Professional University, Phagwara, 144411, India
[b] Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab, 140413, India
[c] Papua New Guinea University of Technology, Lae, PNG
[d] Department of Information Systems, College of Computer and Information Sciences, Jouf University, Al-Jouf, Saudi Arabia

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract

This paper discusses a novel hybrid approach to text classification that integrates a machine learning algorithm along with DistilBERT, a pre-trained deep learning framework for natural language processing, offers a base model fine-tuned on Indian Banking News-Events with a rule-based method that is used by filtering false positives and dealing with false negatives to enhance the results given by the previous classifier. The major benefit is that by incorporating unique rules for certain chaotic or overlapping categories that have not been effectively trained, the system can be quickly fine-tuned. This research also compares the effectiveness of state-of-art deep contextual language representation DistilBERT used in our proposed hybrid model with the most preferred context independent language representation TFIDF, on supervised learning of the classification of multiclass Banking News-Events. Both representations are fed into the machine learning classifiers, Logistic Regression, Linear SVC, Decision Tree, and Random Forest. The findings indicate that DistilBERT can better transfer generic domain knowledge to other domains as compared to the baseline TFIDF and results into higher accuracy with Random Forest. The result is further fed into the Rules based method as per our proposed model. And, our proposed hybrid model resulted into improved accuracy.

Keywords: Multiclass text classification; TFIDF; DistilBERT; Machine or Deep Learning models; Rule-based method

1. Introduction

Hundreds of events are published every day in the form of articles on web news portals or company's website and the need to automatically organize and understand is becoming crucial. With the advancement of machine learning algorithms and its problem solving power, Natural Language Processing has been explored in many ways. One of its common applications is Event Extraction and Classification, the method of collecting data about recurring events contained in documents, automatically extracting information about what happened and when. However, transformation of data is mind-numbing due to unstructured nature of original data.

The extraction of events from unstructured data such as news reports could be useful in various ways for IE systems. The events can be useful in applications of monitoring systems mentioned by [1], the authors [2] assessed of risk analysis in conjunction with events, they [3] presented impact on German stock market of US news events. Financial markets are very subtle to news events. Economic actions for example mergers and attainments, share splits, bonus declarations, etc. play a vital role in broker everyday judgements, where brokers can be people or computers. The authors investigated the role of acquiring firm's news on stock prices [4]. In addition to being capable to understand news more rapidly, computers are able to handle higher dimensions of breaking news, have access to more information than people do, and make better up-to-date decisions.

Automatic monitoring of financial news is concerned with continuous observing of information sources related to banking and non-banking actions and timely exposure of contrary events. The system aims to provide up-to-date knowledge of banking news-events so that people can take preventive actions in-concern to the nature of event. Sometimes events are directly mentioned in the news articles; however certain news articles do not publish content directly mentioning the name of event. It becomes quite important to design technique which can understand the published news-articles, extract event, and label it. The researchers have provided FIDS system to classify financial news automatically in to different domains on the basis of content and extract information from it [5].

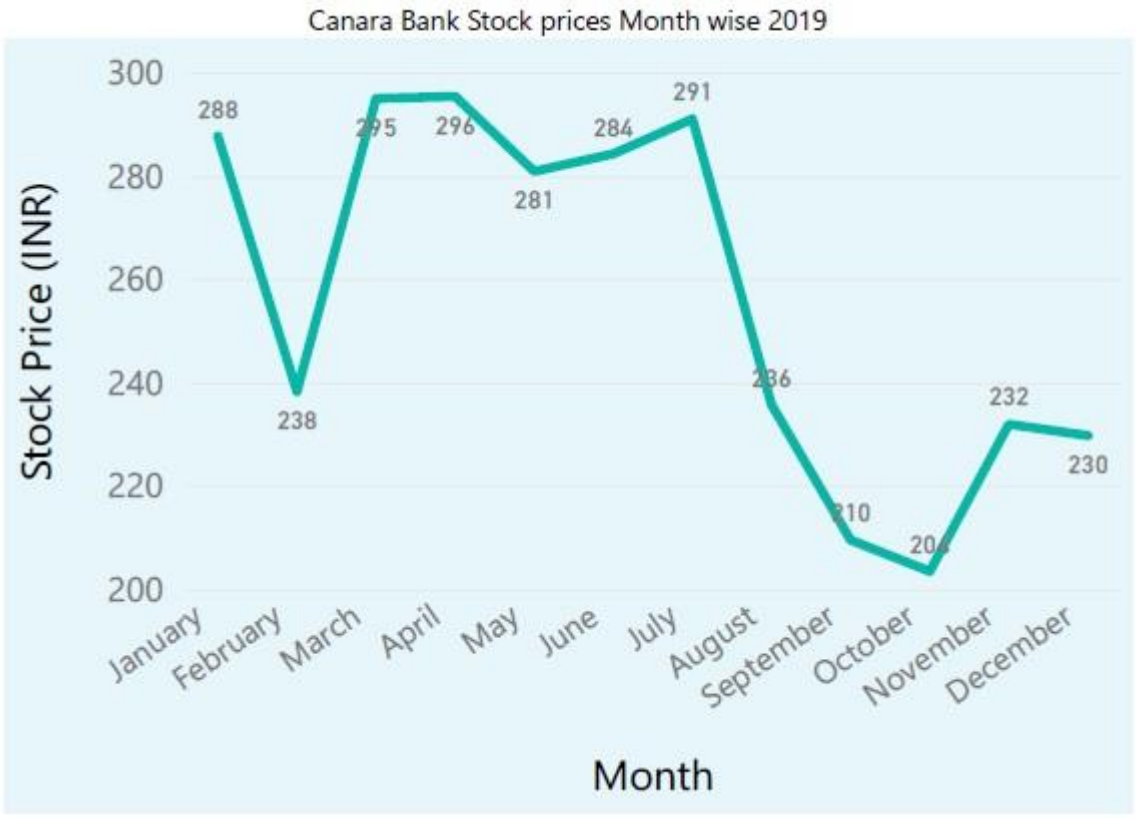

Fig 1. Movement of Stock price of Canara Bank (NSE) for month wise 2019

Considering the favourable prospective for event extraction applications and the assumption that the tasks of real-time extraction and event arrangement can be adequately addressed, it is worth investigating that text mining practices are suitable for this purpose. The extraction of information on financial activities, which are anomalies captured in terminology referring to common (complex) money and risk terms – such as mergers and acquisitions, stock splits, dividend announcements, etc., is of paramount importance. In addition, this can be related to participants, times and locations. Financial markets are highly sensitive to news [6]. For example as mentioned in Fig. 1, the stock price of Canara Bank in NSE touched a low value INR 200 and high value INR 300 in 2019, which can be attributed in large part to the impact of financial events related to Canara Bank that took place during the same fiscal time. The stock prices are subject to rapid price swings, frequently combined with major financial events.

In May 2019, Canara Bank presented Q4 result with loss narrowed of fiscal 2018-19. The stock has moved to high 300 INR from 280 INR in a period of a month. During same period, government of India has announced the plan of capital infusion in PSU banks. It had risen up the sentiments of investors. However, the PSU banks had taken a hard sharp knock after the Union budget was being presented in same period. Also the news was started addressing of PSU banks on non-performing loans which was reached at peak in FY20. The stock of Canara Bank had slipped sharply to INR 210 in September 2019. In start of December 09, 2019 there was news of Canara Bank to dilute 30% of stake in Can Fin Homes with premium. Also the government of India addresses the issue of economic slowdown in December 2019. The expert had started giving buy call on Canara Bank in mid-December 2019 and stock has re-gained to 230 INR levels on December 23, 2019. This example shows that when used in financial applications, events contained in news messages can play an important role. This can be used in applications such as prediction of stock price change and algorithmic trading. Algorithmic trading is the use of computer programs to join commercial orders with algorithms that decide on factors such as timing, size, and order quantity.

An interrelated application to information extraction is the **text classification**, in which a document is assigned to one of several classes based on its content. This problem has been studied well before and various techniques have been explored in the literature, comprising approaches of machine learning studied by [7] Naïve Bayes, [8] provided on the basis on local feature selection, [9] presented feature relation based selection for classification, [10] presented Deep Learning based technique. Extraction of information is viewed in some context as a special form of classification of text, where the number of categories is reduced to two. On the other hand, the text comprehension approach, which attempts to summarize a document by extracting only the relevant information, tackles a more challenging task. It differs in a major aspect from categorizing text and filtering content. A more thorough analysis of the document is required in order to locate the target information, which will inevitably involve the use of natural language processing techniques. It undoubtedly makes the task harder.

The understanding of descriptions of events depends heavily on context. A design needs to take into account the mentions of events in accordance with the sense of discourse to make correct calculations. For example, the following headline from financialexpress.com:

> "FE Online | February 16 2018 4:43 PM HDFC Securities is bullish on the shares of Max Financial Services after the company reported strong **Q3 results**. Last week the company reported a 10.3 percent jump in its **Q3 net profit** at Rs. 88.8 crore against Rs. 80.5 crore in the same period fiscal."

The news headline mentions about the event of quarterly Q3 results of company. It describes about entities Max Financial Services which reports strong results and HDFC securities view point on this stock and it also describes the net profit of the company which consists event related information in context of first line.

In this paper, we design a technique of event extraction and classification into appropriate pre-defined seven event-labels from financial news of Indian banking domain. The task is divided in to two phases. In phase-1, the script is written using python libraries to scrap the news articles from various online news portals, stored in excel files and followed by pre-processing tasks. Furthermore, the classification technique was used to categorize banking news from overall financial news. In phase-2, the events were extracted from news articles of banking domain using hybrid model where a rule based approach is connected with machine learning algorithm to produce reasonable accuracy. We conduct extensive experiments on banking news collected and pre-processed from various news portal, and presents that our technique outperforms the state-of-art techniques for event/text classification.

In this paper, we cover the phase-2; baking event extraction and classification of the overall objective i.e. Financial News collection, event extraction and classification. The following section 2 discusses the recent literature in the field Rule based, Machine Learning or state-of-art Deep Learning and hybrid models for keyword or feature extraction for event classification.

2. Related work

Automatic classification of text is the job of automatically assigning to a given text written in a natural language one or more predefined labels (or topics) according to their similarities with respect to a previously classified corpus used as a sample collection. It is possible to perform text classification in two separate ways: manual or automatic classification. In the former, the content of the document is interpreted by a human annotator and classified accordingly. Typically, this approach can have quality performance, but it's time-consuming and costly. The above applies machine learning, natural language processing and other AI-based methods to identify text automatically in a quicker, more cost-effective and more detailed way.

There are two typical methods for text categorization [11]. The first method, the most common approach to the development of automated document classifiers in the 1980s, consisted of manually constructing an expert model capable of taking decisions by means of knowledge engineering techniques. Usually, such an expert scheme will consist of a series of manually formulated logical rules of some type, one per class. For instance: if (expression) then (class/label). This strategy has the added advantage of being easily understandable to humans. However, apart from the inherent complexity of modelling a text category with a list of logical operations on words, some expert knowledge about the domain, as well as detailed knowledge about the specifics of the rule set as a whole, is needed. The biggest drawback in this situation is that the design of the rule set when working with several hundreds or even thousands of classes or labels is an enormous job that, in most real-world situations, places this technique out of reach [12].

In the other hand, since the late 1990's, the machine learning approach has been the prevalent one. In this case, a set of pre-classified (labelled) texts for each class is given to the system, which is used as a training set, and a classifier is automatically generated from them. The benefit is that domain awareness is only required for of current text in the training set to be assigned a label, which requires a far lower overhead than setting the rules. The following sub-sections discuss the literature of rule-based and data-driven approaches for text classification.

2.1. Rule-based approaches

Rule based approaches are used to write either lexico-syntactic or lexico-semantic patterns using linguistic rules. To identify possible relationships between entities, they are closely based on linguistic pre-processing and morphological analysis. Events are discovered by searching for possible combinations of entities using syntactic dependencies or semantic associations. Therefore, named entities are extracted using rules, and relationships are identified using another method, or vice versa. The author proposes set of linguistic resources require for knowledge-based event extraction system [13].

In the other terms, Rule-based methods define text by using a series of handcrafted linguistic guidelines into ordered categories. These guidelines advise the framework to use text elements that are semantically significant in

order to define the relevant groups on the basis of their content. Each rule consists of a pattern and a class that is expected. The CONSTRUE method is a frequently cited example of this strategy [14]. Several conditional statements using if-else were used in CONSTRUE to identify documents in the Reuters dataset's particular class. A recent research deals with the rule-based approach which uses the embedding technique for transforming a document to vector files [15] and to classify documents using rules.

These systems need deep knowledge of the domain to begin with. They are often time intensive, as it can be very difficult to create rules for a complex structure which typically needs a lot of research and checking. Rule-based systems are often difficult to maintain and do not scale well, provided that the effects of the pre-existing rules may be changed by introducing new rules.

2.2. Data driven approaches

The alternative, a data-driven approach to text classification has been the dominant since 1990's. This is based on methods of machine learning. Machine learning text classification, rather than relying on manually designed rules, learns to make classifications based on previous observations. Machine learning algorithms can learn the various correlations between pieces of text by using pre-labelled examples as training data, and that a certain output (i.e. tags) is required for a particular input (i.e., text). The pre-determined class or group into which any given text can fall is a tag.

Text representation: feature extraction

Feature extraction is the first step in training a machine learning NLP classifier: a procedure is used to turn each text into a vector-shaped numerical representation. Bag of words [16], where a vector in a predefined dictionary of words reflects the frequency of a word, is one of the most commonly used methods. The machine learning algorithm is then provided with training data consisting of pairs of feature sets as vectors and tags to construct a model of classification. The machine learning model will start to make precise predictions until it is equipped with enough training samples. To convert unseen text into feature sets that can be fed into the classification model to get tag predictions, the same feature extractor is used. In fact, on complicated NLP classification tasks, text classification with machine learning is typically far more effective than human-crafted rule systems. Often, it is easier to manage classifiers for machine learning and you can still tag new instances to learn new tasks.

Feature construction and weighting (context-independent)

Two empirical findings concerning text are focused on the machine learning classifiers:

- The more frequently a word appears in a text, the more important it is to the document's topic.
- The more times the term appears in the compilation of documents, the more badly it discriminates between documents.

Supervised NLP methods usually take a term and transform it to a symbolic ID, which is then transformed using a one-hot representation into a feature vector [17]. The feature vector consists of n dimensions in which the quantity of unique terms in the whole dictionary is defined by n. More frequency indicates that the token / feature has more meaning. But some researchers claim that the use of Bag-Of-Words will induce bias and dominate outcomes in the model with terms of very high frequency [18]. The authors have provided a method for assigning weights to token frequencies in the form of a matrix, called TF-IDF (Term frequency-inverse text frequency), to resolve the limitation of BOW. It believes that in another context, very high word frequency will not be able to offer a lot of data gain; uncommon words add more to the model [19]. However, a vast number of significant terms may also consist of a small set of documents, contributing to the issue of scalability or high-dimensionality.

Contextualized feature embeddings

In the context of statistical language processing, the main option for modelling complex text classification or other natural language tasks has arisen. This culminated in the motivation to learn distributed representations of low-dimensional terms of space. The authors suggest representation in order to counteract the curse of dimensionality by studying a distributed representation for words that allows each training sentence to tell the model of an infinite number of semantically surrounding terms [20].

The word embedding approach encourages words to be interpreted in a manner that captures their meanings, semantic associations and the ways in which they are used. The great increasing popularity of word embedding may have been due to the continuous bag-of-words (CBOW) and skip-gram paradigm to efficiently produce high-quality distributed vector representations. Ideally, successful word embedding would represent terms in such a way that two separate words with identical semantic meanings would have similar representations of vector space. It is also important to maintain other linguistic connections between different words. The prediction based continuous bag-of-words model explicitly learns word representation. To predict the word in the middle, the distributed representations of context (or surrounding words) are combined in the CBOW model [21]. The paradigm of the skip-gram that further reshaped Word embedding, its design functions in reverse of the continuous model of bag-of-words. The model forecasts words from the target word for each context. It iterates on the terms in the given corpus of each sentence and uses the present word to predict its neighbours (its context), so the model

is called "Skip-Gram" [21].

The author's approach to creating a word embedding glove (Global Vector, combines count-based and predict-based methods) model for distributed word representation was suggested. Unsupervised learning algorithm, where the model is trained on general word-word co-occurrence statistics that indicate how much it occurs in a corpus and the result obtains the vector representation of words with linear substructures of the word vector space [22]. For texts, like word2vec, we have come up with embedding, which turns a word into an n-dimensional vector. In essence, we will go into an approach to generating sequence embedding that takes a sequence into a Euclidean space, to map the terms into a Euclidean space [23]. We can do standard Machine Learning and Deep Learning on sequence datasets with these embedding approaches. Usually, sequential data requires substantial embedding creation before it is fed into algorithms for data mining. Sequence embedding, where the goal is to transform a sequence into a fixed-length embedding, is one of the sequential data feature learning issues [24].

Researchers have repeatedly shown the significance of pre-training a neural network model with transfer learning on an existing problem in the field of computer vision, as well as doing fine tuning using the learned neural network as the basis for a new intent-specific model. ELMo is such a conceptual embedding that brings the words that surround it into consideration. It models features of word use, such as morphology, and how it is used in multiple contexts [23]. The authors have shown that these representations can be effectively extended to existing frameworks and substantially reinforce state-of-the-art NLP challenges such as answering questions, textual interaction and emotional perception [25].

Another approach to dealing with long dependencies in texts is a transformer [26]. The Transformer is based on an attention mechanism through the encoder and decoder. The Transformer enables the use of this tool to store information in its word vector about the basic meaning of a given term. BERT [27] utilizes Transformer, an attention process that in a document learns contextual relations between words (or sub-words). The authors have argued that traditional technologies restrict the capacity of pre-trained representations, especially for fine tuning approaches. The key limitation is that current language models exist unidirectional, which reduces the number of constructs that can be used during pre-training. BERT is designed to pre-train deep bidirectional representations from unlabelled text documents by mutual conditioning of both left and right contexts in both layers [28]. It remains difficult to run these broad models on the edge and/or within constrained algorithm training or inference budgets. The author proposes a methodology called DistilBERT to pre-train a smaller general purpose language representation model, which can then be fine-tuned on a wide variety of tasks such as its larger counterparts with good performances [29]. While most previous research explored the use of distillation for constructing task-specific models, during the pre-training process, the authors exploit information distillation and demonstrate that the size of a BERT model can be reduced by 40%, while keeping 97% of its language comprehension capability and being 60% quicker. The authors incorporate a triple loss incorporating language processing, distillation and cosine-distance losses in order to exploit the inductive biases gained from larger models during pre-training. In our research, we have used the capabilities of DistilBERT for feature extraction and embedding for computation of event-classification using Machine learning classifier. The following section covers the state-of-art machine or deep learning classifiers for text classification.

Machine learning models

In a number of forms, the inductive architecture of text classifiers has been tackled. Here we will discuss the most common strategies of Text Classification, but we will also briefly discuss the presence of alternate methods that are less common. In the machine learning models, the classifier is trained on the basis of the features chosen from the text documents. The efficiency of learning models is determined by the selection of suitable features in the feature space. When it is trained on a large volume of data, the Naïve Bayes results in the best classification model [30]. The reduction of features, however, remains a concern. Naïve Bayes is then used as a pre-step for SVM, which translates text documents into vectors before the process of classification begins [30]. This resulted in the whole method being strengthened thus wasting very sufficient classification time by reducing to low dimensional space. In other terms, there is a possibility of overfitting, because textual data is always high-dimensional. Support Vector Machines (SVMs) have built-in overfitting resistance, which is one of the reasons why text categorization has been shown to work well [31].

The KNN, a non-linear classifier, was stated in the research, where the algorithm classifies the document by going through all training documents that are identical to that document. In general, the nearest K neighbours in the training set are first defined for each unseen example. After that, the maximum a posteriori concept is used to evaluate the label set for the unseen case, based on statistical knowledge obtained from the label sets of these neighbouring instances, i.e. the number of neighbouring instances belonging to each possible class [32]. Traditional algorithms have some shortcomings that draw researchers to boost efficiency by a) reducing computational overheads by generating low-dimensional space b) accelerating the computational ability to locate nearest neighbours in KNN or locating decision limits in SVM c) enhancing efficiency by not losing accuracy [33].

The authors choose the Random Forest algorithm in one of the works, as it can achieve high output metrics in text classification [34]. The authors demonstrate from the observations that the average values of Naïve Bayes, Random Forest, and Support Vector Machine classification accuracy are identical and the difference is not highly significant. With the exception of Logistic Regression, the efficiency of the classification methods evaluated is more reliable and the overall classification accuracy values are less distributed [35]. It was discussed in the book

that all linear classification models derive their inspiration from the parent issue of linear regression, which traditionally precedes the formulations of classification [36]. In the past study, from systematic experimentation, the authors have found that the process of feature extraction and selection has a major impact on the precision of the model of machine learning [37]. The influence of using various pre-processing techniques, such as the use of N-grams, removal of stop-words, and stemming, is also studied [38]. The results show that the choice of method of feature extraction has a significant effect on the resulting classifications.

Deep learning models

To learn hierarchical representations of data, deep learning methods utilize numerous processing layers and have achieved state-of-the-art results in many fields. In computer vision applications and pattern recognition, deep learning architectures and algorithms have already made remarkable progress. Recent NLP research is now increasingly focused on the use of new deep learning approaches, following this pattern. Machine learning techniques solving NLP issues have been focused on shallow models trained on very high dimensional and sparse features for decades (e.g., logistic regression and SVM). Neural networks based on dense vector representations have provided superior results for different NLP tasks in the last few years. The progress of word embedding [39] and deep learning [40][25] techniques is triggering this trend. In recent deep learning research for NLP, supervised learning is the most common practice. However, we have unlabelled data in many real-world situations that involve sophisticated, semi-supervised or unsupervised approaches.

Among the straightforward deep neural networks for text representation are feed-forward networks. Yet, on many text classification benchmarks, they have reached high precision. A text is treated by these models as a bag of words. Using an embedding model such as word2vec [41]or Glove [22], they learn a vector representation for each word, use the vector sum or average of the embeddings as the text representation, and transfer it through one or more feed-forward layers and then define the representation of the final layer using a classifier such as Naïve Bayes, logistic regression, or SVM.

RNN-based models interpret text as a word sequences, and are intended for text classification to capture word dependencies and text structures. Vanilla RNN models don't always work perfectly, indeed, feed-forward neural networks frequently underperform. Long Short-Term Memory (LSTM) is by far the most common architecture among many variants of RNNs, built to catch long-term dependencies better. The authors apply the Tree-LSTM to tree-structured network topologies, a generalization of LSTMs [42]. And it outperforms all other variants of RNN and LSTM on text classification.

Initially, CNNs were built for computer vision activities, but later found their way into different NLP applications. CNNs comprise of three layer types: (1) convolution layers in which a sliding kernel is added to a text section to retrieve local features; (2) nonlinear layers where (local) obtained features are applied to a nonlinear activation function; and (3) pooling layers where local features are aggregated to form global features (via the max-pooling or mean-pooling operation). In order to use the 1D structure (namely, word order) of text data for accurate prediction, the author studies CNN on text categorization [43].

Attention has recently become an increasingly general method and useful technique in the development of NLP deep learning models [26]. In a brief, focus can be represented in language models as a vector of weights of value. We measure how strongly it is associated with other terms in order to determine a term in a sentence, using the attention vector, and take the sum of their values weighted by the attention vector as the approximation of the target. Self-attention is a special mechanism of attention that causes the similarity of terms in the same sentence to be understood [23]. Transformers also use self-attention, which will be described in next paragraph.

Transformers [26] measure a “attention score” in parallel with each word in a sentence or text to model the effect each word has on another. BERT is also a type of text representation that is a combination of a number of state-of-the-art deep learning algorithms, such as LSTM and Transformers bidirectional encoders. BERT was created in 2018 by Google researchers and has been shown to be state-of-the-art for a number of tasks in natural language processing, such as text classification, text summary, text generation, etc.

To claim that BERT has substantially contributed the NLP paradigm is not a misconception. To obtain state-of-the-art results on different NLP tasks, consider using a single model that is trained on a large unlabelled dataset with little fine-tuning [28]. Another method was provided, called DistilBERT [29], to pre-train a smaller general-purpose language representation model. On a wide variety of functions, such as its larger counterparts, this model can be fine-tuned with favourable performance. While most previous work studied the use of distillation for the creation of task-specific models, during the pre-training process, the authors use information distillation and demonstrate that the size of a BERT mode can be decreased. In recent years, the emergence of Natural Language Processing (NLP) Transfer Learning approaches with large-scale pre-trained language models has become a central method in many NLP activities.

2.3. Hybrid approach

There are disadvantages to both knowledge-driven approaches and data-driven approaches. Formerly one need a lot of high-quality expert manual work and is complicated when porting an existing set of extraction rules to the novel domain. Later one requires large corpus annotation, comparable amount of manual work, and their findings are more difficult to interpret. A research [12] explores a hybrid text categorization method that incorporates a

machine learning algorithm that provides a simple model trained with a labelled corpus, with a rule-based expert system that is used for filtering false positives and working with false negatives to boost the outcomes given by the previous classifier. The authors also explore the hybrid approach of framing rules clubbed with machine learning technique [44]. However, there are still a limited number of contributions and it does not cover all issues.

Until now, hybrid methods have not been used to extract information from banking news articles. In this paper, we address the challenge of applying them to the classification of events from news articles, particularly banking news. The proposed model is discussed in detail in section 3.

3. **System architecture: proposed model**

Event extraction and classification refers to the task of identifying and deriving structured information about events in any text and assign the correct label. It has attracted many academics and industry in terms of the completeness of extracted information and its relevance to many real-life issues. In addition, an increasing number of methods have been proposed to improve the performance of event extraction systems and improve the accuracy of derived information from rule-based approaches to machine learning techniques. It is mentioned that the machine learning method frequently makes mistakes in extracting the de-verbal entities denoted events [44]. This discovery encourages researchers in tandem with machine learning to employ multiple techniques. In this work, we propose a hybrid approach that combines these two approaches in search of better performance by taking advantage of each. The architecture of the proposed hybrid system for event extraction from news articles and classification, particularly banking news-events, is in Fig. 2.

Our work includes an experimental evaluation of Indian banking news for event extraction and classification using hybrid approach by finding event extent and event triggers using rule based approach in conjunction with machine learning algorithm. Data collection was based on a python script for monitoring and scrapping unstructured online news; a set of news documents were chosen to classify documents suitable for the banking domain, such as documents detailing banking events. We are interested in dealing with seven events occur in Indian Banking system time to time mentioned in the Section 4.

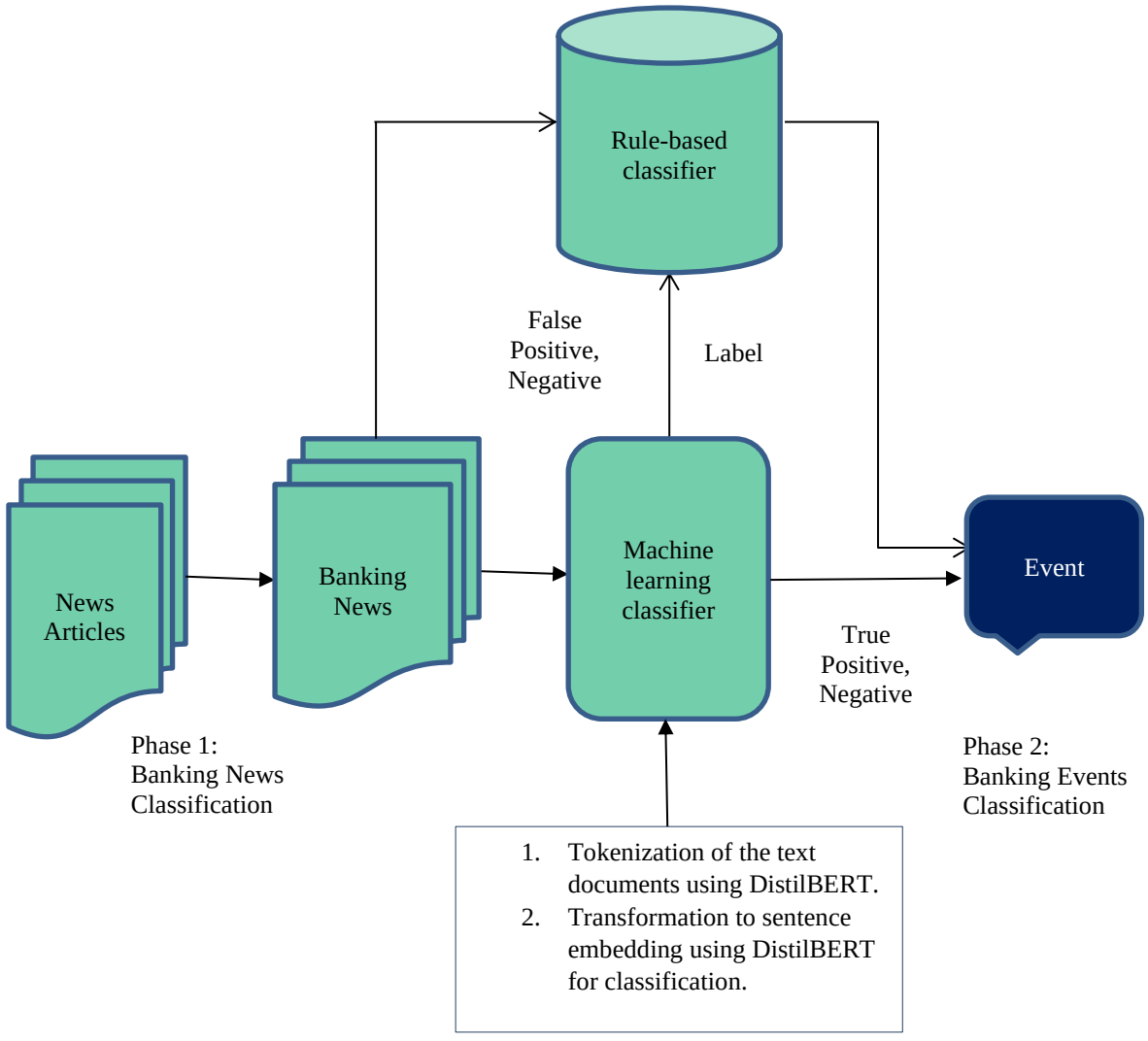

Fig. 2. Steps for News-Events representation and classification using hybrid approach of Rule based and machine learning model based on DistilBERT

Our method for events classification involves three steps: (1) Representation of news articles using DistilBert; (2) Fine Tuning DistilBert along-with machine learning models for classification; (3) Validating the result of step 2 with Rules for false positives and false negatives.

3.1. DistilBERT for banking news-events representation

Embedding is one of the exciting implementations of both unsupervised learning and transfer learning since the large unlabelled corpus promotes embedding. The need for the transfer learning was all high. There were some advantages of transformer-based model. The benefits are twofold: (1) Instead of generating an input sequence token by token, these models take the whole sequence as input in one go, which is a major change over RNN-based models, so the GPUs will now speed up the model. To pre-train these models, we don't need labelled info. This means that to train a transformer-based model, we just have to provide a large amount of unlabelled text data. For NLP task, News-Events classification, we used this transfer learning model.

Training a BERT model on a small dataset from scratch will result in overfitting. So as a starting point, it is easier to use a pre-trained BERT model trained on a large dataset. On our comparatively smaller dataset, we can then further train the algorithm and this process is known as model fine-tuning. However, transfer learning through large-scale pre-trained systems grows prominently in Natural Language Processing, it remains difficult to run these large models under limited computational training or inference budgets. The authors suggest a method to pre-train DistilBERT, a simplified general-purpose language representation model that can then be fine-tuned on a wide variety of tasks such as its larger counterparts with good results [29]. There are some questions regarding the trend toward bigger models in terms of parameters count as shown in Fig. 3. First is the environmental cost of increasingly growing the computing requirements of these models. Second, though running these on-device models in real-time has the potential to allow new and exciting language processing applications, these models' increasing computing and memory requirements can impede broad adoption [45].

The findings demonstrated it on a variety of downstream tasks, a 40 percent smaller Transformer pre-trained by distillation through the guidance of a larger Transformer (based on BERT) language model can attain comparable efficiency, while being 60 percent faster at inference time.

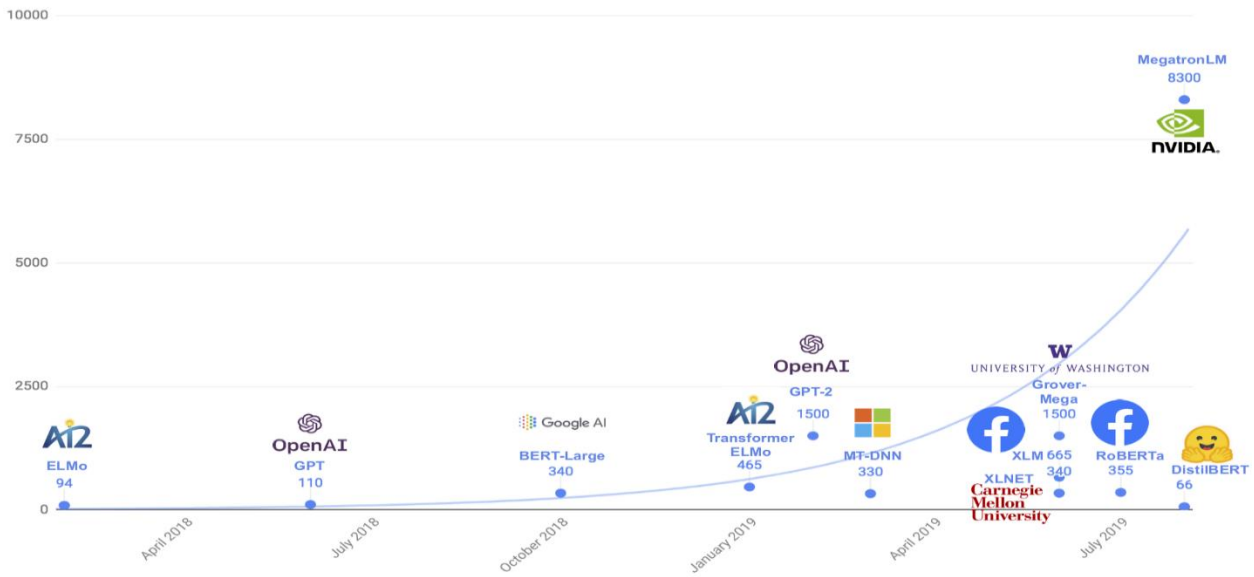

Fig. 3. Several pre-trained models with parameters count [29]

3.2. Machine learning classifiers

In order to compare the classification efficiency in detail, different classifiers are used. Amongst the most important and commonly used Machine Learning Algorithms is logistic regression. Logistic regression is generally one of the first few subject areas that individuals choose when learning predictive modelling. Logistic regression is not an algorithm for regression, but a model for probabilistic classification. The authors apply the model to a variety of problems with document classification and demonstrate that it yields at least as accurate compact predictive models as those developed using SVM coupled with the selection of features [46]. Linear SVC is another applicable machine learning algorithm for the NLP problem. A Linear SVC tries to match the data we have, returning a “best fit” subspace that divides or classifies our data. The text classification application was developed using Support Vector Classifier and achieves high accuracy [47]. The Decision Tree classifier uses a hierarchical

decomposition of training data to separate the data based on the state of data points. The inclusion or absence of words is usually a requirement for separation. The attribute feature is associated with each non-leaf node, and each leaf node is associated with a positive or negative classification value [48]. The decision tree is commonly used in research, and it is easy to comprehend the rules created by the decision tree. The rules created are however, biased against certain features. There are still too many small branches in the tree in a cost-sensitive scenario and the overfitting issue is severe. In addition, because the decision tree has only one root, the rules created are typically essential locally, but ineffective globally. The author creates a decision forest, which consists of many strong decision trees, to discover more efficient rules. In developing a predictive algorithm, it works better than the traditional decision tree methods on imbalanced data [49]. Random Forest is an effective and interpretable algorithm for machine learning; it achieves specific predictions for different forms of datasets since it employs random sampling and ensemble strategies. Among most machine learning algorithms, the interpretability of the model and the prediction accuracy of Random Forest are very rare. In comparison, in training results, Random Forest is less prone to outliers then there's no need to trim the trees because the bootstrapping and ensemble method makes Random Forest effectively address 0verfitting's problems. Random Forest thus has all the benefits of decision trees and, leading to the use of bagging on samples, random subsets of variables and voting systems, it produces better outcomes much of the time [50]. The performance of each classifier is discussed in the Section 5.

3.3. Fine tuning with rules

A rule-based method is usually composed of a series of if then rules represented in such a way that in the field of artificial intelligence there are different approaches to information processing. The most prominent one though, may be in the format of if then rules defined as: IF condition (set of rules) THEN result (pattern set). Identifying evet's extent is the first step in annotating an Event mention in the text. The extent of an Event summary is always the whole word describing the Incident. While it is not an assessed function to classify the extent of the case, it will be critical in the process of annotation. In particular, it defines whether or not Values and Entities in the text can be used in neighbouring Events as **arguments**. Only Entities and Values are permissible arguments within the extent of an event. A word or phrase describing its occurrence clearly. Event Trigger can be nouns, verbs and sometimes adjectives like "scam" or "bankrupt". It can be considered that rule-based systems are information that can be interpreted in different contexts as a set of rules deciding what to do or what to infer. Event attributes are always entities and values that are not properly included in the context of an event as mentioned in Table 1 for each class.

Table 1: List of Events labels and Attributes

Event Label	**Attributes (Supporting terms)**
Results	quarter, quarterly, q1, q2, q3, q4, q o q, yearly, year-on-year, y o y, profit, net profit, loss, net loss, estimates, npa, fiscal, flat, report, reported, provision, provisioning.
RBI_Policies	issue, issued, interest rate, consumer price index, cpi, consumer price index, gdp, gross domestic product, circular, slump, slumped, economists, economy, complaint, complaints.
Merger_Or_Acquisition	amalgamation, amalgamated, subsidiary, merger, merged, merging, integration, disinvest, disinvestment, acquisition, acquire, reduce.
RatingsAgencies_Experts_View	rating, ratings, outlook, maintained, revised, bb-, bb+, cc-, ccc, ccc+, bbb+, bbb-, a+, b+, cc+,a1+, faa, stable, asset, quality.
Fraud	scam, fraud, probe, unfold, mis-governance, scandal, inquiry, case, register, registered, booked, loan default.
Global	trade war, fed, global, globally, asian, dollar, trade data, import, export, world bank, barrel, brent crude oil, geo-political, opec.
Governmental	union budget, govt borrowing, indian government bond, gst, election, ltcg, tax, state, center, minister, lok sabha, nda, parliament.

4. **Experimental set up**

4.1. Data sets

In our experiment, we gathered data by scrapping news from public news sources such as Bloomberg, Financial Express, Money Control, and Times of India using python-written code. As a result, we have been collected more than 10000 instances of financial news articles from the year 2017 to 2020. The news articles belong to various events. We are interested to classify the banking and other most related news-articles in to 7 events as mentioned in Table 2. The data has been collected by running python script several times a day. In phase-1 of the model, the banking and other most related news were separated and classified from financial news, so that in phase-2 (in this research), news-articles related to our area of interest could be entertained. The news data are then pre-processed such that the machine learning or deep learning algorithms may learn from the training dataset and adapt them in an acceptable way to the testing dataset. Therefore, these are pre-processed for the machine learning models to be explored from the training sample and implemented in an appropriate format to the test data set.

Table 2: List of Events labels and news samples for classification

Event Label	**News Samples**	**Date and Time**
Results	Private lender HDFC Bank NSE -1.84 % on Saturday reported an 18.17 per cent year-on-year (YoY) growth in net profit at Rs 4,601.44 crore for the quarter ended June.	Jul 21, 2018, 05:34 PM
RBI_Policies	Reserve Bank of India cuts interest rates to the lowest in over 2 decades as the central bank expects the GDP to contract this financial year.	MAY 22, 2020, 12:45 PM
Merger_Or_Acquisition	Oriental Bank of Commerce (OBC) and United Bank of India will be merged into Punjab National Bank (PNB). After the merger, these together will form the second-largest public sector bank in the country, after State Bank of India (SBI).	Apr 01, 2020, 08:18 AM
RatingsAgencies_Experts_View	Fitch Ratings upgraded private lender IDBI Bank's Viability Rating (VR) by one-notch from "ccc" to "ccc+". The upgrade in VR is due mainly to the improved core capitalization and the high loan-loss coverage.	December 04, 2020, 12:53 P.M
Fraud	A plunge in the shares of PSU banks kept the headline indices Sensex and Nifty with minimal rise in the morning deals on Monday. The shares of the fraud-hit PNB slumped over 4% in today's trade.	February 20, 2018, 10:35 AM
Global	Trump signed two proclamations that levied a 25 percent tariff on steel and a 10 percent tariff on aluminium imported from all countries except neighbouring Canada and Mexico.	March 09, 2018, 07:21 PM
Governmental	LTCG tax has blunted the edge of equities as compared to last year; however, from a growth perspective, equities should be able to perform better than other asset classes, which ideally should ensure the inflows into equity continue.	March 15, 2018, 11:01 AM

4.2. Data transformation and classification

The next step, data transformation, is to turn the data into suitable forms suited to the classification process, and by creating a vector collection of features, DistilBert is fine-tuned on the news articles together with machine learning classifier. In this stage, the feature vector is normalized and scaled to prevent an unbalanced dataset using up-sampling technique. Also, the classes (events) on a set of training data are defined in this step, and then the

machine learning classifier uses these classes to perform the data classification. The 4 most prominent machine learning classifiers has been tested for classification task, those were Logistic Regression, Linear SVC, Decision Tree and Random Forest. The machine learning algorithms were implemented on data for classification using Python's libraries scikit-learn and imblearn for dealing imbalanced dataset of Python. The transformer library of Python was used to implement DistilBERT. The training of the model was done on google-colab.

We conducted evaluation in terms of Precision and Recall by means of micro-averaging. Precision, the probability that in that class, what has been labelled as being is actually. Recall the probability that an object is currently identified in a class as being in that class. We compared the performance of above mentioned classifiers with respect to seven Banking News-Events mentioned in Table 2. The Random Forest classifier has resulted into higher accuracy amongst all machine learning algorithms.

4.3. Applying rules

This base model was further fine-tuned by our Rule-based section that is based on simple or complex logical expression of Natural Language processing. Our model supports some rule description in this section that can impose such conditions on the words that appear in the news-articles or not. However a simpler rule language that makes the development of rules easier has been defined for our implementation. There can be one or many rules connected with each event-label. Each rule is checked against the input news-article to accept or reject a category, depending on whether or not the input news-article fits the criteria set out in the rule. Rejecting a label eliminates false positives returned by the classifier for machine learning, thereby improving accuracy. Any false negative is overcome by adding a new label, so recall improves. The benefit of our model is that by fine-tuning the classifier by writing specific rules for each label, the accuracy of our model has surpassed to the base model by 1.0%. Writing rules for each label did not require much expert workload as compared to the expert's efforts required to train the classifier on training set with set of documents for each label with defined rules. For each category, the following constituents are considered for validating categories:

- Collection of attributes for i^{th} event $E_i = \langle e_{i1}, e_{i2}, \dots, e_{in} \rangle$
  The news-events must consist at-least one of these attributes;

  if $\langle e_{i1}$ ***or*** e_{i2} ***or*** $\dots$ ***or*** $e_{in} \rangle$ **then**
  label the category (i)
  else
  pass to the next article
- Collection of non-supporting attributes for i^{th} event $NE_i = \langle ne_{i1}, ne_{i2}, \dots, ne_{in} \rangle$
  The news-events must not consist any of these attributes;
  if $\langle ne_{i1}$ ***or*** ne_{i2} ***or*** $\dots$ ***or*** $ne_{in} \rangle$ **then**
  pass to the next article
  else
  label the category (i)
- The decision of labeling the news-article will depend on the count of supporting attributes resulted for each news-article in favor of a particular category. The total number of attributes for each category is kept same.

5. Results and discussion

We have carried out several experiments on our pre-processed news-articles utilizing state-of-art context-dependent representation and sentence embedding technique DistilBert, the embedding are further passed to conventional machine learning algorithms. The key purpose of these experiments is to determine the right classifier that gives the best performance. Every classifier's output for classification is calculated using the metrics Precision, Recall, and F_1-score. The outcomes of the best classifier 'Random Forest' is further fine-tuned using Rules based expert system.

5.1. Results of Banking News-Events classification with DistilBERT

Table 3 lists the result of DistilBERT fine-tuned with different machine learning classifiers on banking news-events in terms of precision, recall and F-1 score. Table 4 lists the result of accuracy of the models.

Table 3. Results of the machine learning classifiers fine-tuned with DistilBERT

Classifier	Logistic Regression			Random Forest			Decision Tree			Linear SVC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Fraud	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Global	1.00	0.88	0.93	1.00	1.00	1.00	0.97	1.00	0.98	1.00	1.00	1.00
Governmental	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98	1.00	1.00	1.00
Merger_Or_Acquisition	0.96	1.00	0.98	0.97	1.00	0.98	1.00	1.00	1.00	0.97	1.00	0.98
RBI_Policies	0.94	1.00	0.97	1.00	1.00	1.00	0.91	1.00	0.95	1.00	1.00	1.00
RatingsAgencies_Experts_View	1.00	1.00	1.00	1.00	0.96	0.98	1.00	0.82	0.90	1.00	0.89	0.94
Results	1.00	1.00	1.00	0.96	1.00	0.98	1.00	1.00	1.00	0.93	1.00	0.96

Table 4: Accuracy of the classifiers with DistilBERT

Classifier	Accuracy
Logistic Regression	0.98
Random Forest	0.99
Decision Tree	0.97
Linear SVC	0.98

From the different classifiers Decision Tree, Linear SVC, Logistic Regression, and Random Forest, the Random Forest performed best with accuracy 99% as shown in Table 4. The Random Forest achieved the F_1-score 1.00, 1.00, 1.00, 0.98, 1.00, 0.98, and 0.98 for classes Fraud, Global, Governmental, Merger_Or_Acquisition, RBI_Policies, RatingsAgencies_Experts_View and Results respectively. The comparison of all the mentioned classifiers for 7-different classes on Banking News Events is shown in Table 3 in terms of Precision, Recall and F-1 score. And the accuracy of the classifiers varies from 97% to 99%.

5.2. Results of banking news-events classification with TFIDF

Table 5 lists the result of TFIDF language representation with different machine learning classifiers on banking news-events in terms of precision, recall and F-1 score. Table 6 lists the result of accuracy of the models.

Table 5. Results of the machine learning classifiers with TFIDF

Classifier	Logistic Regression			Random Forest			Decision Tree			Linear SVC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Fraud	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Global	1.00	0.87	0.93	1.00	0.89	0.94	1.00	0.76	0.87	1.00	0.89	0.94
Governmental	1.00	1.00	1.00	1.00	1.00	1.00	0.85	1.00	0.92	1.00	1.00	1.00
Merger_Or_Acquisition	0.96	1.00	0.98	0.97	1.00	0.98	0.97	1.00	0.99	0.97	1.00	0.98
RBI_Policies	0.94	1.00	0.97	1.00	1.00	1.00	0.98	0.93	0.96	1.00	1.00	1.00
RatingsAgencies_Experts_View	1.00	0.96	0.98	1.00	0.96	0.98	1.00	1.00	1.00	1.00	0.89	0.94

Results	0.93	1.00	0.96	0.96	1.00	0.98	0.91	1.00	0.95	0.93	1.00	0.96

Table 6. Accuracy of the classifiers with TFIDF

Classifier	Accuracy
Logistic Regression	0.97
Random Forest	0.98
Decision Tree	0.95
Linear SVC	0.97

From the different classifiers Decision Tree, Linear SVC, Logistic Regression, and Random Forest, the Random Forest performed best with accuracy 98% as shown in Table 6. The Random Forest achieved the F_1-score 1.00, 0.94, 1.00, 0.98, 1.00, 0.98, and 0.98 for classes Fraud, Global, Governmental, Merger_Or_Acquisition, RBI_Policies, RatingsAgencies_Experts_View and Results respectively. The comparison of all the mentioned classifiers for 7-different classes on Banking News Events is shown in Table 5 in terms of Precision, Recall and F-1 score. And the accuracy of the classifiers varies from 95% to 98%.

5.3. Results of banking news-events classification with hybrid approach

Table 7 lists the result of proposed hybrid model and DistilBERT fine-tuned with Random Forest classifier on banking news-events in terms of precision, recall and F-1 score. Table 8 lists the result of accuracy of the both models.

Table 7. Results of the DistilBERT fine-tuned with Random Forest classifier and proposed Hybrid model

Approach	(Hybrid) DistilBERT+Random Forest+Rules			DistilBERT+RandomForest		
	P	R	F_1	P	R	F_1
Fraud	1.00	1.00	1.00	1.00	1.00	1.00
Global	1.00	1.00	1.00	1.00	1.00	1.00
Governmental	1.00	1.00	1.00	1.00	1.00	1.00
Merger_Or_Acquisition	0.98	1.00	0.99	0.97	1.00	0.98
RBI_Policies	1.00	1.00	1.00	1.00	1.00	1.00
RatingsAgencies_Experts_View	1.00	0.98	0.99	1.00	0.96	0.98
Results	1.00	1.00	1.00	0.96	1.00	0.98

Table 8. Accuracy of the DistilBERT fine-tuned with Random Forest classifier and proposed Hybrid model

Classifier	Accuracy
(Hybrid) DistilBERT+RandomForest+Rules	1.00

DistilBERT+RandomForest	0.99

From the both approaches DistilBERT fine-tuned with Random Forest classifier and proposed Hybrid model, the Hybrid model performed best with accuracy 100% as shown in Table 8. The Hybrid model achieved the F_1-score 1.00, 1.00, 1.00, 0.99, 1.00, 0.99, and 1.00 for classes Fraud, Global, Governmental, Merger_Or_Acquisition, RBI_Policies, RatingsAgencies_Experts_View and Results respectively. The comparison of both approaches for 7-different classes on Banking News Events is shown in Table 7 in terms of Precision, Recall and F-1 score.

6. **Conclusion and future direction**

This paper aims to classify the banking news amongst seven events mentioned in Table 2. This multi-class classification helps to get the news on the commonly occurred events in the Indian Banking system. The development of a system for classifying banking news and other related news is a major and untested problem. We are interested in seeking news-events from Indian banks, the Indian government, and the global. We take a structured approach to choose and group the banking news articles into 7 classes. The news articles are gathered from numerous online news sources and labelled to derive the banking-events and other related news-events to achieve the paper's goal. To automate the classification process, we propose a novel hybrid technique which is used to classify the banking-news articles into 7 events (Fraud, Global, Governmental, Merger_Or_Acquisition, RBI_Policies, RatingsAgencies_Experts_View and Results). Since our data set faces the class imbalance issue, we used SMOTE method to align the data set between classes, and the classifier's output is evaluated using balanced data set. We used precision, recall, F-1, and accuracy parameters to evaluate our model. It is evident from results that the proposed hybrid model with the DistilBERT language representation model fine-tuned with Random Forest classifier and Rules achieved the highest accuracy of 100%. Based on our results, our trained classification model can be used to automatically classify the other domain-specific news into the domain-specific events by modifying the rules for those events. As the DistilBERT is pre-trained deep learning model, the model can be further fine-tuned on domain-specific texts. The small volume of data was labelled for that purpose manually with the help of the domain expert. In our future research, we may plan to apply the technique with a larger volume of data for more banking news-events.

References

[1] J. Park, S. Ha, and F. K. Chang, "Monitoring impact events using a system-identification method," AIAA J., vol. 47, no. 9, pp. 2011–2021, 2009, doi: 10.2514/1.34895.

[2] T. Flohrer, H. Krag, and H. Klinkrad, "ESA's process for the identification and assessment of high-risk conjunction events," Adv. Sp. Res., vol. 44, no. 3, pp. 355–363, 2009, doi: 10.1016/j.asr.2009.04.012.

[3] T. Dimpfl, "The impact of US news on the German stock market-An event study analysis," Q. Rev. Econ. Financ., vol. 51, no. 4, pp. 389–398, 2011, doi: 10.1016/j.qref.2011.07.005.

[4] M. Erickson and S. W. Wang, "Earnings management by acquiring firms in stock for stock mergers," J. Account. Econ., vol. 27, no. 2, pp. 149–176, 1999, doi: 10.1016/S0165-4101(99)00008-7.

[5] W. Lam and K. S. Ho, "FIDS: An intelligent financial web news articles digest system," IEEE Trans. Syst. Man, Cybern. Part ASystems Humans., vol. 31, no. 6, pp. 753–762, 2001, doi: 10.1109/3468.983433.

[6] F. Hogenboom, Automated Detection of Financial Events in News Text. 2014.

[7] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," Expert Syst. Appl., vol. 36, no. 3 PART 1, pp. 5432–5435, 2009, doi: 10.1016/j.eswa.2008.06.054.

[8] N. Armanfard, J. P. Reilly, and M. Komeili, "Local Feature Selection for Data Classification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 6, pp. 1217–1227, 2016, doi: 10.1109/TPAMI.2015.2478471.

[9] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 447–462, 2011, doi: 10.1109/TKDE.2010.110.

[10] N. Majumder, I. Politécnico, N. Soujanya Poria, A. Gelbukh, I. P. Nacional, and E. Cambria, "AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS Deep Learning-Based Document Modeling for Personality Detection from Text," IEEE Intell. Syst. (Vol. 32 , Issue 2 , Mar.-Apr. 2017), vol. 32, no. 2, 2017.

[11] F. Sebastiani, "Machine Learning in Automated Text Categorization," vol. 34, no. 1, pp. 1–47, 2002.

[12] J. Villena-Román, S. Collada-Pérez, S. Lana-Serrano, and J. C. González-Cristóbal, "Hybrid approach combining machine learning and a rule-based expert system for text categorization," Proc. 24th Int. Florida Artif. Intell. Res. Soc. FLAIRS - 24, pp. 323–328, 2011.

[13] V. Solovyev and V. Ivanov, "Knowledge-Driven Event Extraction in Russian: Corpus-Based Linguistic Resources," Comput. Intell. Neurosci., vol. 2016, 2016, doi: 10.1155/2016/4183760.

[14] L. Hirsch, M. Saeedi, and R. Hirsch, "Evolving rules for document classification," Lect. Notes Comput.

Sci., vol. 3447, pp. 85–95, 2005, doi: 10.1007/978-3-540-31989-4_8.

[15] A. M. Aubaid and A. Mishra, "A rule-based approach to embedding techniques for text document classification," Appl. Sci., vol. 10, no. 11, 2020, doi: 10.3390/app10114009.

[16] B. S. Harish, D. S. Guru, and S. Manjunath, "Representation and classification of text documents: A brief review," IJCA, Spec. Issue Recent Trends Image Process. Pattern Recognit., no. 2, pp. 110–119, 2010, [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Representation+and+classification+of+text+documents:+A+brief+review#0.

[17] Y. Xiong et al., "Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity," BMC Med. Inform. Decis. Mak., vol. 20, no. Suppl 1, pp. 1–7, 2020, doi: 10.1186/s12911-020-1045-z.

[18] Sivic and Zisserman, "Video Google: a text retrieval approach to object matching in videos," Proc. Ninth IEEE Int. Conf. Comput. Vis., pp. 1470–1477 vol.2, 2003, doi: 10.1109/ICCV.2003.1238663.

[19] D. Isa, L. H. Lee, V. P. Kallimani, and R. Rajkumar, "Text document preprocessing with the bayes formula for classification using the support vector machine," IEEE Trans. Knowl. Data Eng., vol. 20, no. 9, pp. 1264–1272, 2008, doi: 10.1109/TKDE.2008.76.

[20] J. D. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, "Distributed Representations of Words and Phrases and their Compositionality," Adv. Neural Inf. Process. Syst. 26, pp. 1389–1399, 2013, doi: 10.18653/v1/d16-1146.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," vol. 5, pp. 135–146, 2016, [Online]. Available: http://arxiv.org/abs/1607.04606.

[22] C. D. M. Jeffrey Pennington, Richard Socher, "GloVe: Global Vectors for Word Representation," Assoc. Comput. Linguist., pp. 1532–1543, 2014.

[23] M. E. Peters et al., "Deep contextualized word representations," NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 2227–2237, 2018, doi: 10.18653/v1/n18-1202.

[24] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.

[25] S. Minaee, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1, pp. 1–42, 2020.

[26] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[27] M. C. Kenton, L. Kristina, and J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 1953.

[28] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11856 LNAI, no. 2, pp. 194–206, 2019, doi: 10.1007/978-3-030-32381-3_16.

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," pp. 2–6, 2019, [Online]. Available: http://arxiv.org/abs/1910.01108.

[30] G. Forman, "Feature Selection for Text Classification," pp. 257–276, 2010, doi: 10.1201/9781584888796.pt4.

[31] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features," Eur. Conf. Mach. Learn. Springer, pp. 137–142, 1998.

[32] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognit., vol. 40, no. 7, pp. 2038–2048, 2007, doi: 10.1016/j.patcog.2006.12.019.

[33] S. Manne and S. S. Fatima, "A Novel Approach for Text Categorization of Unorganized data based with Information Extraction," Int. J. Comput. Sci. Eng., vol. 3, no. 7, pp. 2846–2854, 2011.

[34] E. R. H. Estevam R.HruschkaJr., Nádia F.F.da Silvaa, "Tweet sentiment analysis with classifier ensembles," Decis. Support Syst., vol. Volume 66, no. October 2014, pp. 170–179, 2014, doi: https://doi.org/10.1016/j.dss.2014.07.003.

[35] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," Balt. J. Mod. Comput., vol. 5, no. 2, pp. 221–232, 2017, doi: 10.22364/bjmc.2017.5.2.05.

[36] C. C. Aggarwal and C. C. Aggarwal, Machine Learning for Text: An Introduction. 2018.

[37] A. C. Haury, P. Gestraud, and J. P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," PLoS One, vol. 6, no. 12, pp. 1–14, 2011, doi: 10.1371/journal.pone.0028210.

[38] M. Eklund, "Comparing Feature Extraction Methods and Effects of Pre-Processing Methods for Multi-Label Classification of Textual Data," Comp. Featur. Extr. Methods Eff. Pre- Process. Methods Multi-Label Classif. Textual Data, 2018.

[39] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," Inf. Sci. (Ny)., vol. 471, pp. 216–232, 2019, doi: 10.1016/j.ins.2018.09.001.

[40] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, "Recurrent neural network based language model," Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010, no. September, pp. 1045–1048, 2010.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc., pp. 1–12, 2013.

[42] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-Term memory networks," ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf., vol. 1, pp. 1556–1566, 2015, doi: 10.3115/v1/p15-1150.

[43] R. Johnson and T. Zhang, "Effective Use of Word Order for Text Categorization with Convolutional Neural Networks," pp. 103–112, 2014, [Online]. Available: http://arxiv.org/abs/1412.1058.

[44] A. K. Kolya, A. Ekbal, and S. Bandyopadhyay, "A Hybrid Approach for Event Extraction," Polibits, vol. 46, no. 46, pp. 55–59, 2012, doi: 10.17562/pb-46-6.

[45] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., no. 1, pp. 3645–3650, 2020, doi: 10.18653/v1/p19-1355.

[46] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," Technometrics, vol. 49, no. 3, pp. 291–304, 2007, doi: 10.1198/004017007000000245.

[47] Yi Wang and Xiao-Jing Wang, "A new approach to feature selection in text classification," no. August, pp. 3814-3819 Vol. 6, 2005, doi: 10.1109/icmlc.2005.1527604.

[48] A. Bhardwaj, Y. Narayan, Vanraj, Pawan, and M. Dutta, "Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty," Procedia Comput. Sci., vol. 70, pp. 85–91, 2015, doi: 10.1016/j.procs.2015.10.043.

[49] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," World Wide Web, vol. 16, no. 4, pp. 449–475, 2013, doi: 10.1007/s11280-012-0178-0.

[50] M. N. Elagamy, C. Stanier, and B. Sharp, "Stock market random forest-text mining system mining critical indicators of stock market movements," 2nd Int. Conf. Nat. Lang. Speech Process. ICNLSP 2018, pp. 1–8, 2018, doi: 10.1109/ICNLSP.2018.8374370.