# FEATURE EXTRACTION FROM NEWS ARTICLES

**A Project Report**

*Submitted by*

**Ashutosh Karbhari Wakhure 121722018**

*in partial fulfilment for the award of the degree*

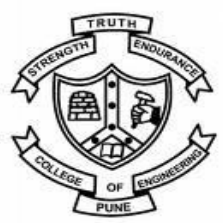
*of*

## M.Tech. (Computer Engineering)

Under the guidance of

**Prof. Vaibhav K. Khatavkar**

College of Engineering, Pune




**DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY,**

**COLLEGE OF ENGINEERING, PUNE-5**

June, 2019

# DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY, COLLEGE OF ENGINEERING, PUNE

## CERTIFICATE

Certified that this project, titled "FEATURE EXTRACTION FROM NEWS ARTICLES" has been successfully completed by

**Ashutosh Karbhari Wakhure 121722018**

and is approved for the partial fulfilment of the requirements for the degree of "M.Tech. Computer Engineering".

SIGNATURE

**Prof. Vaibhav K. Khatavkar**
**Project Guide**
**Department of Computer Engineering**
**and Information Technology,**
**College of Engineering Pune,**
**Shivajinagar, Pune - 5.**

SIGNATURE

**Dr. Vahida Z. Attar**
**Head**
**Department of Computer Engineering**
**and Information Technology,**
**College of Engineering Pune,**
**Shivajinagar, Pune - 5.**

## Acknowledgement

I would like to thank my project advisor, Prof. Vaibhav K. Khatavkar for guiding me through the course of project work. Also, I would like to mention along with his support; it was his pure perseverance that pushed me to perform better during the project that significantly contributed towards its completion on schedule.

I would like to express appreciation towards Dr. Vahida Z. Attar, Head of Department, Computer Engineering and Information Technology, College of Engineering Pune and Dr. B. B. Ahuja, Director, College of Engineering, Pune for their invaluable support in during this project work. I would like to thank Mr. Shubhashish Tiwari, Mr. Shivraj Nashi, and Persistent Systems Ltd. for the encouragement and the support for higher education.

Finally, I would like to thank my mother Dr. Padma Karbhari Wakhure, family members, and friends for their perennial encouragement and emotional support. I would dedicate this work to my late father **Karbhari Muralidhar Wakhure** whose dreams for me have resulted in this accomplishment.

**Abstract**

According to literature survey, the document classification is still a challenging problem. The research study initiated with the hypothesis that based on the important features, Entities and Key-Phrases, the news articles can be classified with improved accuracy. It focuses on the feature extraction from news articles and classification. The BBC News dataset is used in the study, which consists of 2225 news articles across five categories. Mainly Entities, Key-Phrases are targeted for the news articles classification. The news features are extracted with AWS Comprehend service. The TF-IDF used as feature weighting method and vectorizer. k-NN, SVM, Random Forest, Logistic Regression  lbfgs, liblinear, Linear Regression with SGD, Multinomial Nave Bayes classification models are trained with the extracted features and tested with cross validation technique. Based on the results, the study evaluates the classification models and the significance of Entities, Key-phrases in news articles classification.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Document classification has always been an important application and research topic since the inception of digital documents. Document classification is the task of grouping documents into categories based upon their content. It performs an essential role in various applications that deals with organizing, classifying, searching, and concisely representing a significant amount of information. Shahnawaz et. al. presents the literature survey on sentiment analysis and open issues in documents classifications. [35] The paper provides a good survey on the various machine learning approaches, their categorization, the differentiation based on methodology, the limitations and open issues in the sentiment analysis and text classification. The sentiment analysis is derived field of documents classification and further choosing what to be considered as window and observed results based on the features and the domain of sentiment analysis.

In every supervised machine learning task, an initial dataset is needed. A document can be assigned to more than one category, but in this study only research on Hard Categorization (assigning a single category to each document) are taken into consideration. The BBC news dataset used in the study which consists 2225 news articles across five categories  business, sports, politics, technology, entertainment. The important features are extracted from the news articles dataset, mainly entities, key-phrases for news

articles classification, and to identity the importance in news classification and their co-relation. The training and test dataset would be BBC news dataset as cross validation technique with 70-30 ratio.

To perform document classification with machine learning, documents needs to be represented such that it is understandable to the machine learning classifier. The Term Frequency-Inverse Document Frequency (TF-IDF) is numerical statistic weighing method, implies a strong relationship with words in document in terms of weight and in most of the research papers it has been proven as significant measure of feature representation. AWS Comprehend service is used in our study to have more insights into the news article documents for extraction of Entities, Key-phrases. AWS Comprehend uses Latent Dirichlet Algorithm (LDA) which recognizes these Topic Modelling features with the richness of dataset features and its evolution. The Stanford Parser, Spacy are used in prior research for Part-of-Speech tagging and Named Entity Recognition alternatively. The number of features and feature sets differs across the parsers.

S. Foroozan et. al. used SVM with n-gram model to validate whether the extracted sentences is event sentence. [32] The SVM accuracy is achieved in a relatively better results way as per the domain considering prior work. The study evaluates SVM linear and RBF SVM method with TF-IDF as weighing method. The combination of feature extraction method (unigram and bigram) using SVM outperforms with 97.03 accuracy in results. [32] The SVM - SVC linear classification model is evaluated in the study considering the prior work and the experimental results of SVM for different domains in literature survey.

The k-NN (k  Nearest Neighbor) is powerful machine learning algorithm, can be widely used for both classification and regression predictive problems. The k-NN algorithm outperforms in the both the domains when evaluated with the three parame-

ters technique  Ease to interpret output, Calculation time, Predictive power and when compared with Logistic Regression, CART (Classification And Regression Trees) and Random Forest. According to the literature survey, k-NN for classification of the news articles should provide result at par and more better results. The k-Means clustering is also experimented on the dataset as Bag-of-Words (BOW). The Random forests is a statistical method for classification. It was first introduced in 2001 by L. Breiman. It is a decision-tree based supervised learning algorithm. The Random forest consists of many individual decision trees. Each decision tree votes for classification of given data. The random forest algorithm then accepts the classification which got a maximum number of votes from individual trees. Collectively the decision tree models represent or form a random forest where each decision tree votes for the result and the majority wins. [45]

According to the objectives of the study, the feature selection is evaluated over classifiers  k-NN, SVM, Random forest, Logistic Regression (lbfgs), Logistic regression (liblinear), Linear Regression with SGD, Multinomial Naive Bayes. The importance of features entities and key-phrases in news articles classification is studied. Based on the classification results and metrics, the performance of individual classifier is measured.

The project report is divided as follows: Chapter 2 discusses about literature survey on feature extraction from news articles, news articles classification and sentiment analysis from feature extraction and selection perspective. In Chapter 3, research hypothesis, problem statement and objectives are described. Chapter 4 provides the description of the dataset used in the study. Chapter 5 elaborates the related terms in the research domain. Followed by Chapter 6, it provides the proposed system and procedure. Chapter 7 highlights the implementation part and Chapter 8 provides detailed analysis on the obtained classification results, exercised classifiers and considered features.

# Chapter 2

# Literature Survey

Shahnawaz et. al. have provided with a brief literature review on approaches and open issues in Sentiment Analysis. [35] The sentiment analysis approaches are categorized into statistical, machine learning based approaches, knowledge or lexicon-based approaches and hybrid approaches as shown in Figure 2.1. In statistical approaches, the SVM, latent semantic analysis, semantic orientation, and Bag-of-Words learning techniques are discussed. In statistical and machine learning-based approaches, SVM and Naive Bayes classification methods are explained. In knowledge and lexicon-based approaches, role of NLP and lexical resources to extract knowledge from the opinionated data to analyse sentiment of the text is explained. The machine learning based approaches are categorized into three learning methods 1. Supervised existing sentiment classification methods like SVM, k-NN, Naive Bayes and their limitations are discussed. 2. Unsupervised the differentiation with supervised, the role of labelled training data, and the limitations of unsupervised in terms of training data and disjointed topics issues is discussed with reference to prior literature. 3. Semi-supervised- the semi-supervised methods make use of either or both supervised and supervised. The supervised learning methods can tackle the problem of unavailability of training data is explained. Supervised learning methods include generative models, self-training, multiview learning, co-training,

and graph-based methods. The sentiment analysis process and open problems and issues in sentiment analysis are discussed. [35]

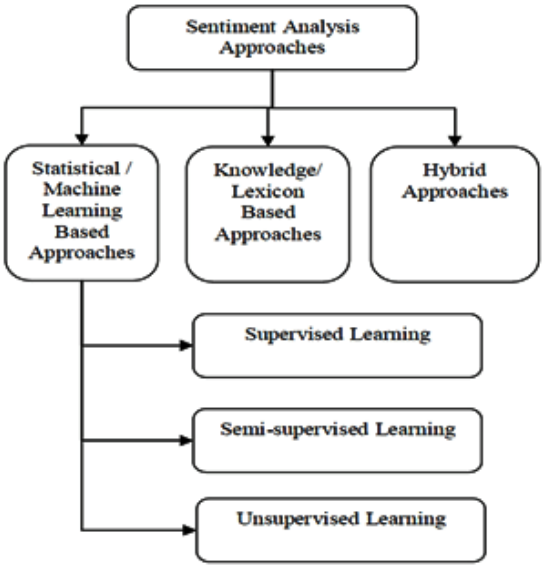



Figure 2.1: Shahnawaz, et. al. - Sentiment analysis approaches [35]

Mr. Nilesh M. Shelke, et. al. have worked on aspect-oriented sentiment analysis using statistical feature based approach. [5] The categorization iterates over various types of text features observed in literature and their importance as shown in Figure 2.2. It also provides the summary of the feature selection methods categorized into four methods 1. NLP and heuristic based methods, 2. Statistical methods, 3. Clustering methods, 4. Hybrid methods, and their underlined techniques as shown in Figure 2.3. The proposed system in research measures the similarity and relatedness between words with WordNet and focused on semantic features of word domains. It used SentiWordNet 3.0 for polarity classification. The accuracy results are obtained camera 0.847 and restaurant domain 90% based on single review.

Masayu Leylia Khodra et. al. focuses on event extraction from Indonesian news articles using multiclass categorization. [1] The 5W1H corpus is constructed by human annotator for all the news documents. It uses tagging sequence under BIO (Begin

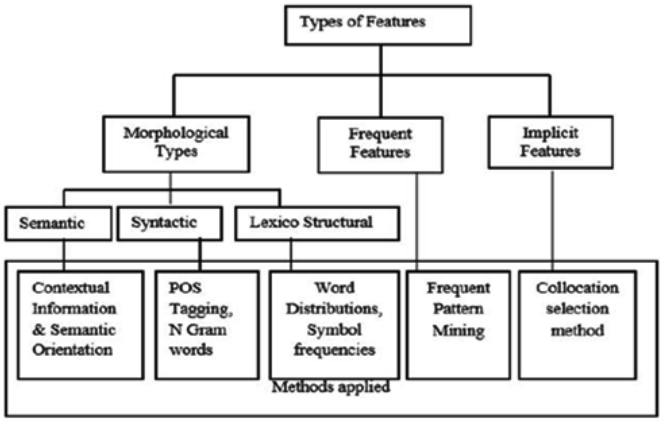



Figure 2.2: Mr. Nilesh M. Shelke, et. al. - Types of features observed in literature [5]

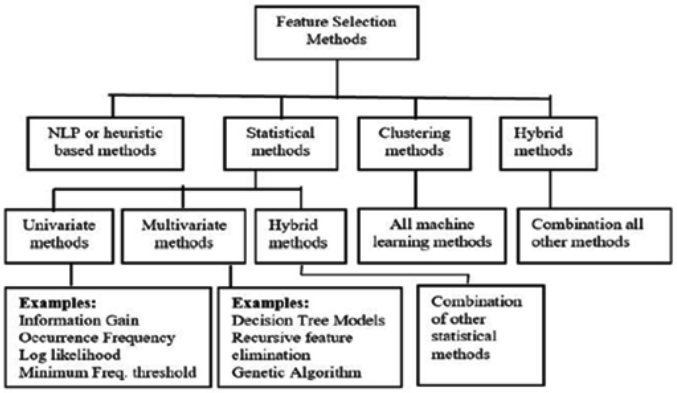



Figure 2.3: Mr. Nilesh M. Shelke, et. al. - Summary of feature selection techniques [5]

Inside Outside) labelling method, statistical learning-based approach for event extraction as a sequence labelling problem and standard techniques of supervised learning like decision tree with adaptive learning to handle imbalanced dataset. The input text tokens are classified into one of the 13 predefined classes. The lexical analysis, POS tagger are used to extract the Named Entities from news events.

Terry Traylor, et. al. worked on classifying fake news articles to identify In-Article Attribution as a Supervised Learning Estimator. [2] The importance of this research is explained in terms of opinion forming, decision making, and voting patterns. The study focuses on the fake news distributed over social media platforms like twitter and facebook and news documents classifier and performance measures. The TextBlob,

NLP, SciPy toolkits are used to develop a fake news detector model which uses quoted attribution in the Bayesian machine learning approach to detect the document is fake. The dataset is locally generated and validated for fake news over 7-month period. Taufik Fuadi Abidin et. al. study focuses on Event Sentence Extraction from Indonesian Online News Articles using the combination of Rule based and machine learning based methods. [3] It focuses on the tropical disease information and incident identification from Indonesian news articles. The proposed method is composed of the two steps: determining candidates of sentence using a combination of rule-based algorithm and contextual and morphological components; the candidates of sentences are further classified using SVM to validate whether the sentences are event sentences. The n-gram model is applied to generate numerical weights which are found in the dictionary. The study is conducted over a collection of sentences derived from 1,863 tropical disease web pages as dataset. The evaluation results show that the accuracy of SVM to classify the sentences that have the incidence date and the number of casualty are 79.51% and 86.99% respectively.

S. Foroozan, et. al. research focuses on the sentiment classification of financial news with the identification of positive and negative news. [32] The news are applied in decision support system to perform stock trend predictions. It explores various types of feature space as different datasets for sentiment classification of the news article. The n-gram approach (unigram, bi-gram and their combination) used as feature extraction with different feature weighting methods, while, document frequency (DF) is used as feature selection method. The experimentation evaluates the classification accuracy of SVM with two kernel methods - linear and Radial Basis Function (RBF). The paper concludes that the combination of feature extraction methods (unigram and bi-gram) boosts classification accuracy upto (97.03%). It is also observed that DF method can be used as a dimensional reduction approach to the number of features.

Raihannur Reztaputra, et. al. research used POS tagger, Dependency parser, syntaxnet, Density based Spatial clustering for Application with Noise (DBSCAN) for Sentence Structure-based Summarization for Indonesian News Articles. [6] It extended the existing summarization using clustering procedure into more derived form. It examines over the differentiation of sentence structure information extractions scores for simple, complex and all the sentences from news articles.

Ismini Lourentzou, et. al. study focuses a novel problem of mining news text and social media jointly to discover controversial points in news. [7] It enables many applications such as highlighting controversial points in news documents for readers, identifying controversies in news and their trending over time, and quantifying the controversy of a news sources. It proposed the controversy scoring function to identify the most controversial statements in news documents by using in twitter and news websites comments.

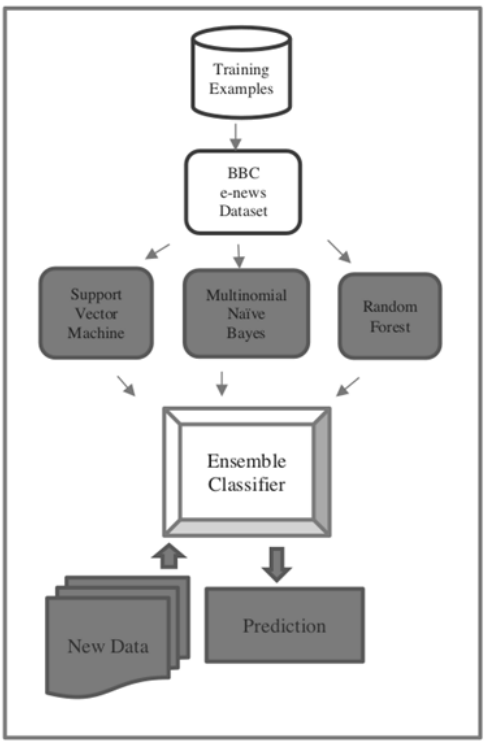



Figure 2.4: Dilini Dandeniya, et. al. - Ensemble classifier [8]

Dilini Dandeniya, et. al. worked on an automated e-news articles content extraction and classification. [8] The BBC News dataset is used as training dataset and tested over RSS feeds and news from URLs dataset. The work combinatorial approach for feature extraction and news classification with SVM, Random Forest and Multinomial Naive Bayes classifiers and the study results in effective results in terms of classification. The concept of ensemble classifier proposed as shown in Figure 2.4. and evaluated in the study. It provides better results as shown in Table 2.1.

| **Classifier** | **Accuracy** |
|---|---|
| RandomForestClassifier | 0.964044 |
| MultinomialNB | 0.957513 |
| GussianNB | 0.921320 |
| BernoulliNB | 0.948310 |
| Support Vector Machine | 0.973033 |
| Ensemble classifier (Hard Voting) (RandomForestClassifier, MultinomialNB, Support Vector Machine) | 0.975280 |
| Ensemble classifier (Soft Voting) (RandomForestClassifier, MultinomialNB, Support Vector Machine) | 0.979775 |

Table 2.1: Dilini Dandeniya, et. al. - Probabilistic result for different classifiers [8]

Pal-Christian S. Njlstad, et. al. worked on evaluating feature Sets and classifiers for sentiment analysis of financial news proposes an experimental on sentiment analysis of financial internet news using five different machine learning models - Artificial Neural Networks (ANN), SVM, Nave Bayes, Random Forrest, and J48. [9] Norwegian financial internet news articles dataset is used for sentiment classifier and the precision has been obtained up to 71%. The financial news features are extracted and classified

into four main categories Textual Feature Category(X), Categorical Feature Category (C), Grammatical Feature Category(G), Contextual Feature Category (X) along with the underlined features. The study evaluated the ANN, SVM, RF, J48 and NB and J8 classification trees yield the highest classification performance and followed by RF. The performance of growing datasets and results are shown in Table 2.2 and graphical representation of different feature subsets and growing datasets shown in Figure 2.5.

| **Features Categories** | **SVM** | **ANN** | **NB** | **RF** | **J48** |
|---|---|---|---|---|---|
| T | 57.4 | 57.7 | **57.8** | 56.7 | 57.3 |
| TC | 57.1 | 57.8 | **59.1** | 57.6 | 57.9 |
| TCG | 57.6 | 58.7 | **59.0** | 58.2 | 56.9 |
| TCGX | 66.5 | 64.0 | 68.2 | 66.9 | **68.7** |
| BEST* | 68.4 | 65.8 | 70.1 | 70.2 | **70.8** |

Table 2.2: Pal-Christian S. Njlstad, et. al. - Classification Precision Results (%) [9]

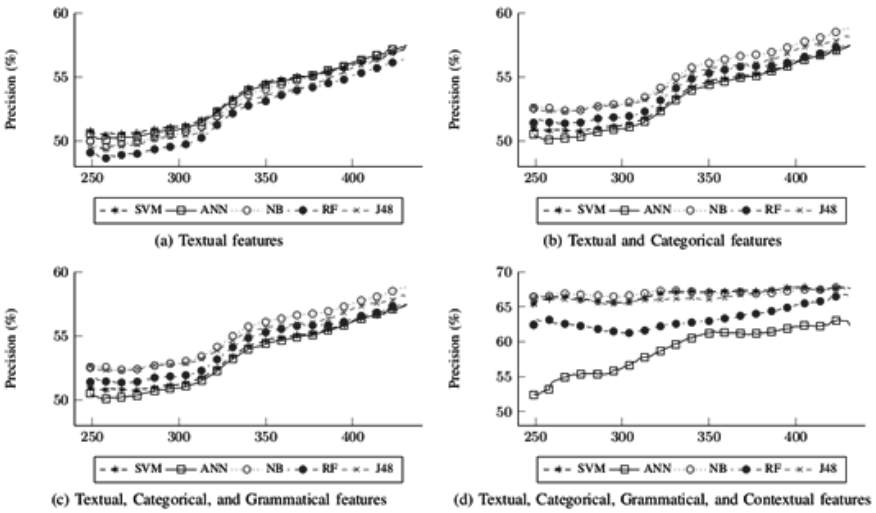



Figure 2.5: Pal-Christian S. Njlstad, et. al. - Classification precisions with different feature subsets and growing datasets (smoothed over 10 steps) [9]

Adhy Rizaldy1, et. al. provided with a comparative study on the SVM classifier and SVM classifier with feature selection for information gain. [10] Information

Gain (IG) is often used as discriminant of attributes of attributes in machine learning field. The IG of a term is measured by counting number of bits of information extracted from predicted category by the presence and absence of terms in the document. [10] The SVM model with feature selection is applied and it improved the categorization accuracy 2.9% from 95.1% to 98.06% with an average of 65.16 seconds computational time as compared to previous work SVM without feature selection. The study used cross-validation technique for testing the results, where this process divides data randomly into 10 sections. The study concluded that the SVM classifier with feature selection for information gain obtains increased accuracy as shown in Figure 2.6.

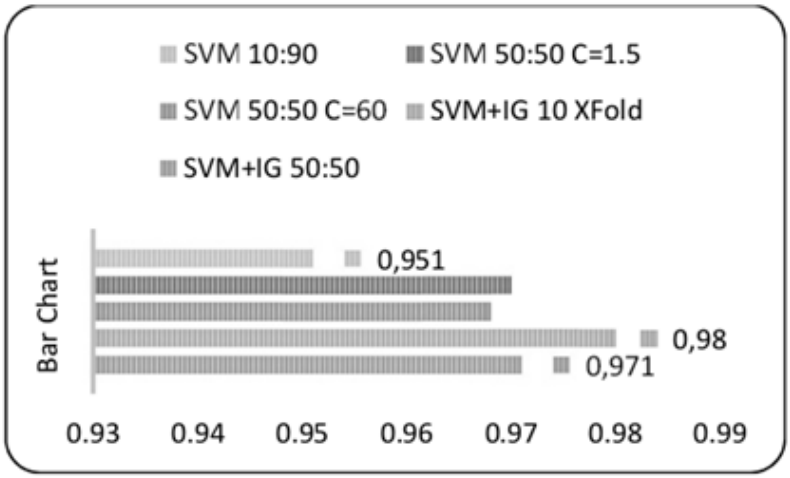



Figure 2.6: Adhy Rizaldy1, et. al. - SVM news classification comparison [10]

Qiang Pan, et. al. research proposes a method to extract a small group of company-specific key phrases from news media that can accurately describe a company, based on prior research work. [11] The key-phrase extraction techniques supervised and unsupervised are discussed i.e. graph based approaches such as PageRank or RandomWalk. The proposed work uses semantic similarity between words to construct the adjacency matrix of PageRank. Further constructs relationship between company-entity with word in the document by co-occurrence relationship so that it can give higher weight

to keywords to some keywords which are relevant to a company. The study proposed EntityRank model and used the semantic similarity between words by using Googles word2vec tool. The similarity between words dont depend on the frequency of words, but the semantic relationship between words. [11] According to the results, concludes entity-based PageRank model for key phrase extraction performs better as shown in Figure 2.7. The experiments on TextRank and TF-IDF is compared to proposed model. The observations are: 1. TF-IDF algorithm has the advantages of simplicity and quickness. However, it simply uses word frequency to measure the importance of a word, not comprehensive, because sometimes important words may not occur many times. 2. TextRank is graph-based model for key phrases extraction, it uses the local vocabulary to sort the subsequent keywords and extract them directly from the text itself. 3. Although TextRank considers the relationship between the words, but still tends to select the frequent words. So, although a word appears many times, it will not be considered as a key phrase. [11] The description of corpus in the study: 1. The 600,000 financial news dataset is crawled from the internet; each news describes something relevant with one or more companies. 2. The corpus for company names as named entity has been used and the company-entity is extracted from the corpus. [11]

Mihail Minev, et. al. research study concerned about the financial news articles, which reflect the monetary policy during the US subprime mortgage crisis. [12] The official announcements conducted by Federal Reserve and its leading representatives is used as dataset in the study. It aims derive a prototype for news classification which should reveal the impact of the events. The significant aspect of the study is the identification, extraction, and representation of topic-related features and the corresponding instances. 1. The combination of NLP techniques and statistical measures with adoption of domain specific terminology. 2. Analyse these features by determining their conditional

| | Metrics | TF-IDF | TextRank | EntityRank |
|---|---|---|---|---|
| Top-6 | Precision(P) | 0.308 | 0.347 | **0.358** |
| | Recall(R) | 0.234 | 0.259 | **0.269** |
| | F1 | 0.266 | 0.297 | **0.307** |
| Top-8 | Precision(P) | 0.274 | 0.290 | **0.318** |
| | Recall(R) | 0.276 | 0.289 | **0.320** |
| | F1 | 0.275 | 0.290 | **0.319** |
| Top-10 | Precision(P) | 0.248 | 0.256 | **0.268** |
| | Recall(R) | 0.311 | 0.319 | **0.327** |
| | F1 | 0.276 | 0.284 | **0.295** |
| Top-12 | Precision(P) | 0.226 | 0.228 | **0.254** |
| | Recall(R) | 0.339 | 0.340 | **0.265** |
| | F1 | 0.272 | 0.273 | **0.300** |

Figure 2.7: Qiang Pan, et. al. - The results of the experiment [11]

instances in each document. 3. The multi-instance (MI) classifier is trained, which correlates the fiscal policy decisions with abnormal stock market movements. The proposed work workflow includes 1. Document Retrieval 2. Information Extraction 3. Classifier Setup and 4. Model Evaluation as shown in Figure 2.8. The description of dataset used 1. Financial news related to the monetary policy of the United States between period 2007 2012, which captures the beginning and the development of the subprime mortgage crisis (financial crisis) 2. The collection includes 174 documents with 1.225.719 tokens. Four document types are enclosed in the set. The paper concludes that this model enabled an explicit tracking of the monetary policy conducted by the Federal Reserve in conjunction with abnormal stock market movements. Furthermore, it facilitates economic surveys by providing a model for extracting financial variables and their conditional values.

K. Ohtsuki, et. al. paper introduced the Topic-Extraction model for topic extraction in Japanese broadcast news speech. [13] The study started with the evaluation and conformance - that using continuous speech recognition, the extraction of several topic-words from broadcast-news. A combination of multiple topic-words represents the content of the news. This is a more detailed and more flexible approach than using a single

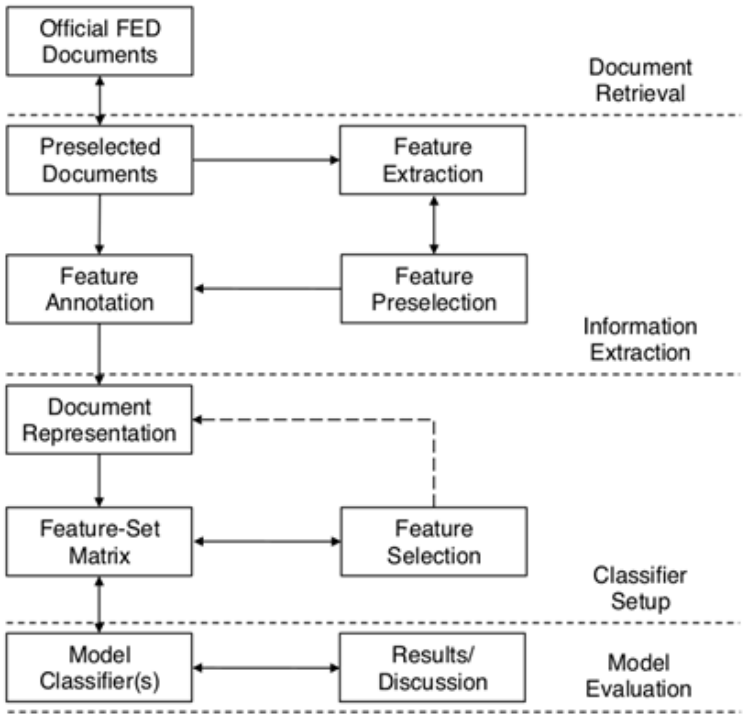



Figure 2.8: Mihail Minev, et. al. - The research workflow [12]

word or a single category. A topic-extraction model shows the degree of relevance between each topic-word and each word in the article. For all words in an article, topic-words which have high total relevance score are extracted. The dataset used: the topic-extraction model is trained with five years of newspapers 900k news articles, using the frequency of topic-words taken from headlines and words from the articles. The distinct topic-words after stemming was about 70k and the degree of relevance between topic-words and words in articles is calculated based on statistical measures, i.e., mutual information or the X2-value. [13] Large vocabulary continuous speech recognition (LVCSR) system has context-dependent phoneme HMMs and statistical n-gram language models. The phoneme HMMs are triphone models designed using tree-based clustering. The study evaluated the transcribed speeches against the model. For mis-recognized words the model uses N-gram approach and hypothesis is used. It has been concluded that N-best hypothesis gives the better results as per experiment conducted over 10-best hypothesis,

the results as shown in Table 2.2.

| Number of topic-words extracted | | 1 | 5 | 10 |
|---|---|---|---|---|
| N-best | 1 | 89.7 | 74.5 | 66.2 |
| | 5 | 93.1 | 75.9 | 69.3 |
| | 10 | 93.1 | 76.6 | 69.3 |
| | 15 | 93.1 | 75.9 | 69.0 |
| | 20 | 93.1 | 75.9 | 69.3 |

Figure 2.9: K. Ohtsuki, et. al. - Topic extraction results with N-best approach (precision [%]) [13]

Benjamin D. Horne, et. al. paper highlights two main questions: 1. Can we predict the type of community interested in a news article using only features from the article content? and 2. How well do these models generalize over time? [14] The corpus used in study: the well-studied content-based features on more than 60K news articles from 4 communities on reditt.com. To evaluate the features which degrades the performance, the model train and test models over three different time periods between 2015 and 2017. The predictions are both community-pair dependent and feature group dependent.

The linear-kernel SVM and Random Forest classifiers is used in the proposed work. Each algorithms hyper-parameters are tuned using 10-fold cross validation. Further, each model is trained using balanced class weights. The classes are natural imbalanced due to varying posting behaviour in each community. Features groups in study are described - Style, Complexity, Bias, Named Entity, Sentiment, Entity Slant, Source. The paper recommends that for prediction of community-interest the hierarchical binary models should be used over standard multi class models, where multiple binary classifiers can be used to separate community pairs, rather than a traditional multi-class model. And, these models should be retrained over time based on accuracy goals and the availability of training data.

Sneha Pasarate, et. al. worked on concept-based document clustering using K prototype algorithm and proposed a system as shown in Figure. 2.9. to increase the efficiency of the system using features for clustering which reduces data dimensionality. [15] The study used Stanford dictionary and POS tagger, LDA for topic modelling to assign each word k-topics randomly. The dataset used: Reuters 21578 dataset, self-created dataset, News article dataset, web dataset for processing. The proposed k-prototype algorithm compared to fuzzy clustering and results have been analysed with respect to time and space complexity across large set of real time documents. The k-prototype algorithm considers the number of mismatches and works on the features for clustering performs better than fuzzy clustering as k-prototype.

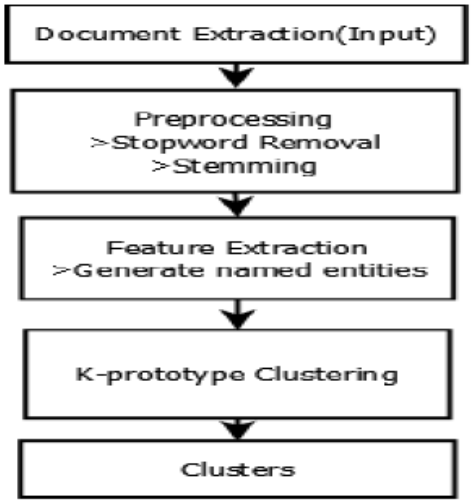



Figure 2.10: Sneha Pasarate, et. al. - Proposed System [15]

Mahsa Afsharizadeh, et. al. worked on Query-oriented Text Summarization using Sentence Extraction Technique with extraction of the most informative sentences. [16] Further, the number of features are extracted from the sentences to evaluate the important of sentences from an aspectand the proposed procedure is as shown Figure 2.11. To evaluate the automated generated summaries, the ROUGE criterion has been used. The process of manual text summarization specialized people requirement for manual text summarization is discussed. Extractive summarization extracts important sentences from source documents and group them together to generate summary. Abstractive summarization creates a brief useful summary by generating new sentences. The TF-IDF is used as numerical criterion which depicts the importance of word in document among the corpus of documents. For extractive summarization, fuzzy logic, graph-based methods, LSA can be used. The dataset in the study: DUC 2007 corpus which has 45 clusters. Each of them has 25 text documents. The proposed technique experimentation results are compared with the previous methods, the better results in summarization are obtained.

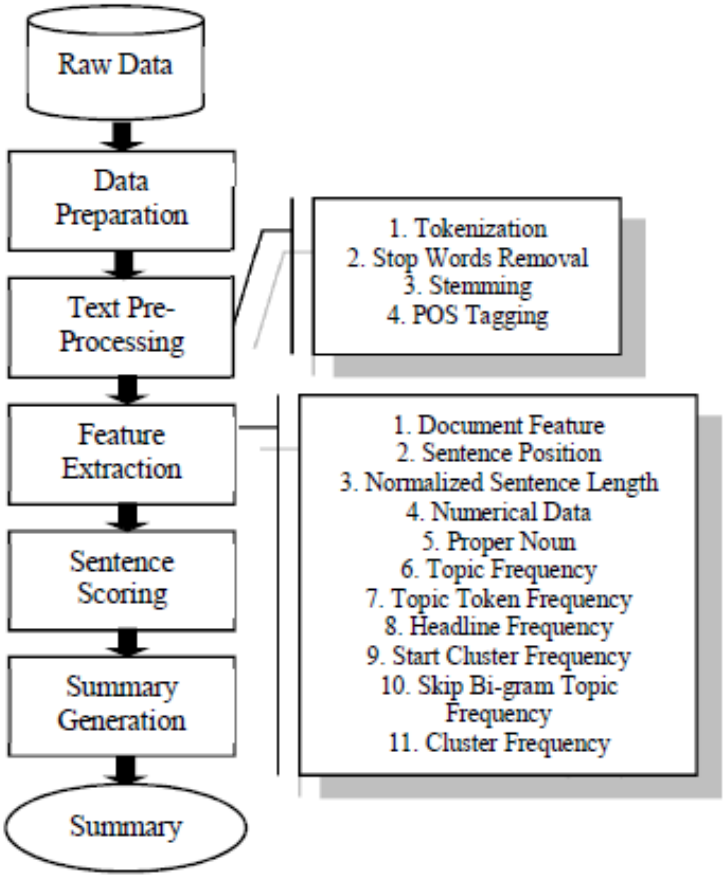



Figure 2.11: Mahsa Afsharizadeh, et. al. - The proposed summarization method [16]

Hairon Sato, et. al. research focuses on the dynamic field of stock market short-term predictions and automated purchase with online news articles. [17] The proposed method as shown in Figure 2.12. can detect the relationships between news articles and exchange rate fluctuations. The method has been applied with foreign exchange forecasting to automatic purchasing systems and the usefulness of proposed methods support for exchange investment has been verified. The TF-IDF, SVM linear SVC implemented with scikit-learn and applied on the preprocessed data. The technical terms are selected from stemmed words. Then, the method scored news articles from TF-IDF, and polarity discrimination was carried out with SVM. The experimentation results show proposed method can effectively predict short-term currency exchange rates.

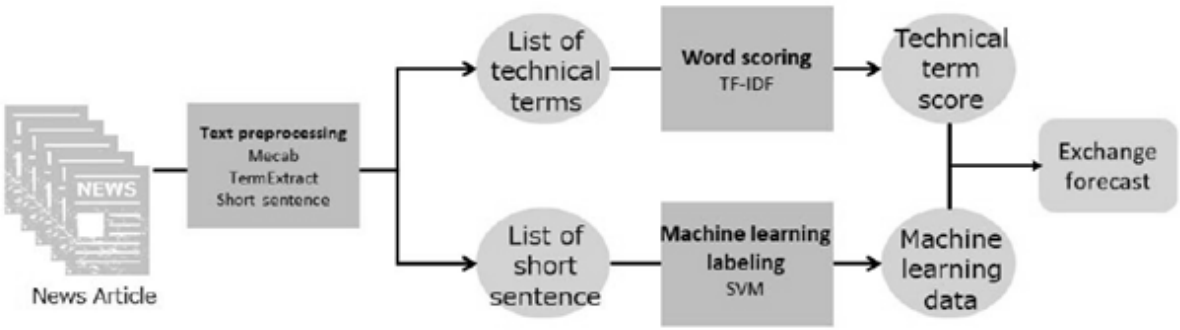



Figure 2.12: Hairon Sato, et. al. - The processing flow of short-term exchange forecasting [17]

Sungjick Lee, et. al. introduces a keyword extraction method that can be used to track topics over time. [18] The keywords are considered most important as keywords provides high level description of the contents to readers. The automated process establishment for keyword extraction from news articles, for rapid use of keywords, is the major step involved. The paper proposes an unsupervised keyword extraction technique that includes several variants of the conventional TF-IDF model with reasonable heuristics. The threshold for taking words is calculated and has been taken as Table Term Frequency (TTF). The paper concluded that the TTF technique is more precise for extracting keywords compared to the TF-IDF variants experimentation results.

Taishi Saito, et. al. presented study on category classification of news using distributed representation of sentences. [19] The proposed method specifically classifies articles by extracting words with similar meanings from sentence vectors of each category. The Paragraph Vector (PV) convention is used in the study and the flow of proposed method is as shown in Figure 2.13. Distributed representation: Distributed representation is a technique of expressing a word as a high-dimensional real vector and it is constructed by two-layer neural network. It can be learned to include linguistic properties of words and phrases and similar words to have similar vectors. To create a distributed representation of a word, there are a method of creating a co-occurrence frequency vector based on a

distribution hypothesis that words of the same meaning appear on the same context, and a method of obtaining from a learning by a neural network. The skip-gram model is obtained from learning neural network. [19]

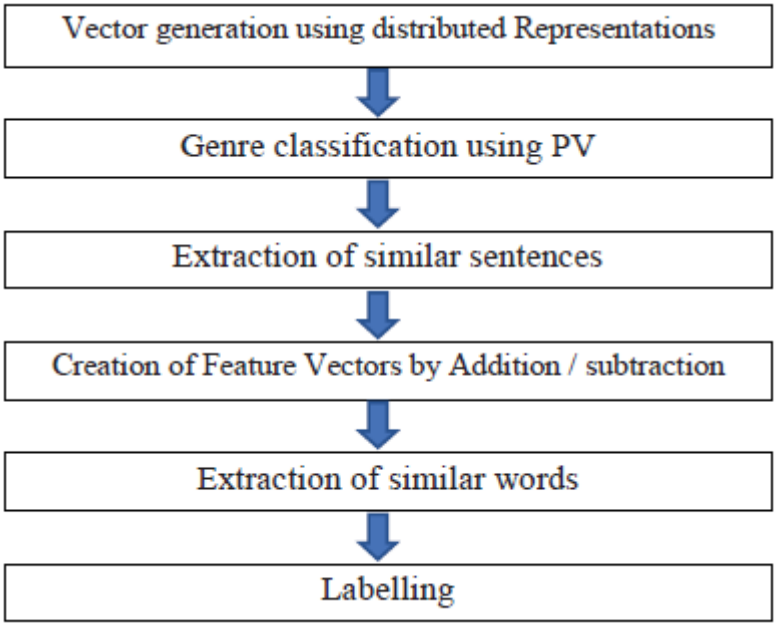



Figure 2.13: Taishi Saito, et. al. - Flow of the proposed method [19]

Naoya OKUMURA Takao MIURA, et. al. proposed a new methodology to estimate headlines for news articles. [20] The limitations of conventional approaches are discussed. The LSA is used to extract semantics and analyse their relationship based on word co-occurrences with each document. The Bag-of-Words (BOW), TF-IDF, LSA, Latent Semantic Indexing (LSI) assuming Single Value Decomposition (SVD) are used. The work is categorized into two parts headline estimation and feature word selection. Corpus: 17,615 articles from January to June in Mainichi news corpus 2012 in Japanese. 1,761 articles as queries from them. Processed the dataset in terms of morphological analysis and to extract all the independent words in advance.

Maryam Bahojb Imani, et. al. worked on automatic identification of geolocation from the online news articles provides vital information for understanding some

events associated with the location. [21] The number of open-source, commercial tools exists for geolocation extraction still they lack in reliable identification of fine-grained location  mostly at country level is identified. The paper proposes a method to solve the problem of location identification in more fine-grained level. Also it focus on news articles describing an event. A set of locations directly associated with the event are called focus locations. However, an event can occur only at single location. The paper aims to extract this location among focus locations, call this as primary focus location, a high-level schema of Profile as shown in Figure 2.14. The NER is used to identify potential sentences containing focus locations and then proceed to identify primary focus geolocation using supervised classification mechanism over sentence embeddings. In the study, feature extraction using sentence embedding is consists of three steps location named entity extraction, resolution and event location extraction. Corpus: Atrocities Event Data is a collection of recent news reports on atrocities and mass killings in several locations. The original size of Atrocity dataset is about 15K reports, and almost 5K of them are annotated. Another dataset the New York Times (NYT) news reports is used in the study. The New York Times Annotated Corpus includes more than 1.8 million articles composed and published by the New York Times between January 1, 1987 and June 19, 2007 with article metadata. Like the Atrocity Event dataset, the study selected political news articles that contain special keywords such as kill, die, injure, dead, death, wounded and massacre in their title. [21]

The paper has proposed a focus location extraction method executable on unstructured text-based news reports. First, the features are extracted using sentence embedding algorithm. The word2vec model is used in this algorithm. Then SVM model is trained to detect the sentences that contain Focus or non-focus locations. The key contributions in this work are extracting the exact focus location at the locality level

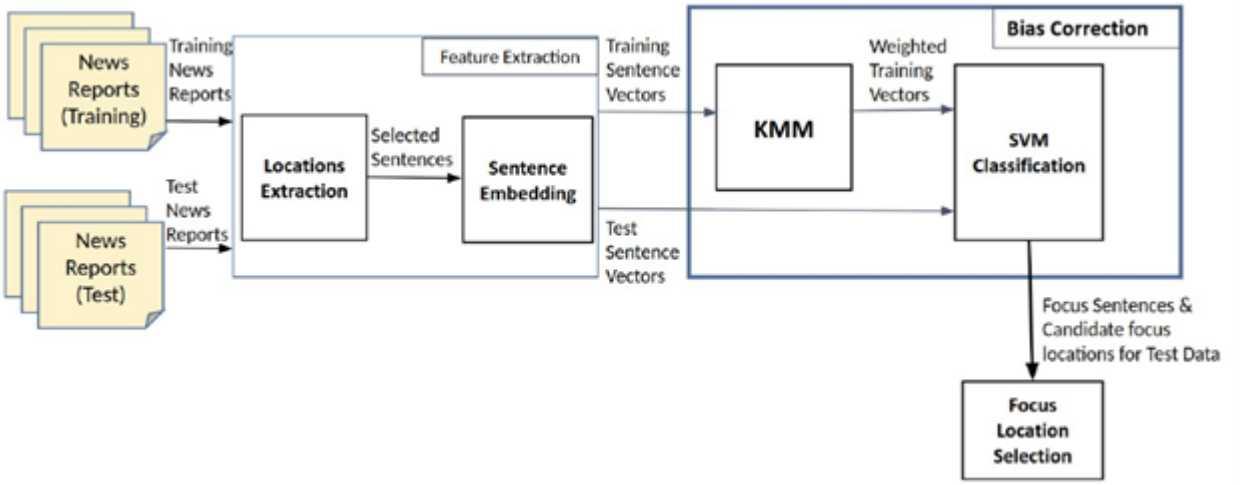



Figure 2.14: Maryam Bahojb Imani, et. al. - A high-level schema of Profile [21]

where an event happened and works based on the semantic relationship among the words in the sentences. The performance of the proposed method exceeds the other approaches considerably as compared to prior work. [21]

Gisel Bastidas Guacho, et. al. worked on semi-supervised content based detection of fake information using Tensorflow assuming a small number of labels, made available by manual fast-checkers or automated sources. [22] The proposed method is shown in Figure 2.15. The paper claims that this is more realistic considering the large amount of content which cannot be manually or easily fast-checked. The previous work on fake news detection has been highlighted and Fact-checking websites such as Snopes.com, PolitiFact.com, and FactCheck.org can be used to assess claims of real or fake detection, although there are some manual interventions by domain experts are required and considering large amount of data it is more time consuming. Dataset: News articles from Twitter tweets during a 3-month period from June-August 2017. These URLs were filtered based on website domain. The news article content are crawled from these URLs using web API boilerpipe, Python library Newspaper3k, and Diffbot. The real news from Alexa and fake news from BSDetector across 367 domains.

The proposed method consists of following three steps: Step 1: Tensor Decomposition The tensor-based article embeddings are needed to build, specifically, proposed the use of binary based tensor construction method. Step 2: k-NN graph of news articles The k-nearest-neighbors of a point in n-dimensional space are defined using a closeness relation where proximity is often defined in terms of a distance metric such as Euclidean distance. Step 3: Belief Propagation Considering small set of news articles and their graphical representation as obtained in above steps, the ground truth tables are obtained. The belief propagation algorithm is used which assumes homophily, because news articles that are connected in k-NN graph are likely to be of same type due to the construction method of the tensor embeddings. The study concludes based-on the experimentation results on over 63K real articles, that the proposed method, leveraging tensor-based article embeddings and guilt-by-association, is performing at par or better than state-of-the-art. [22]

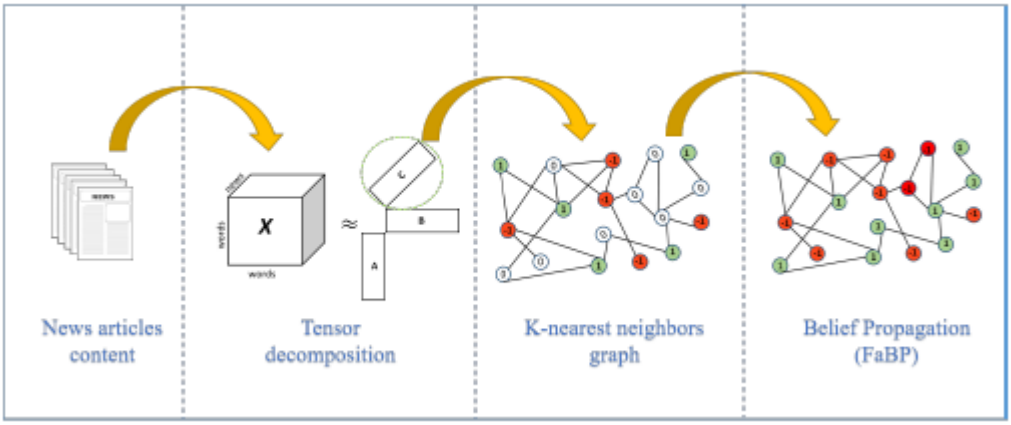



Figure 2.15: Gisel Bastidas Guacho, et. al. - Proposed method discerns real from misinformative news articles via leveraging tensor representation and semi-supervised learning in graphs [22]

Yu Shuqi,Wu Bin, et. al. tackles the event detection task, which aggregate news articles report on same event into tightly coupled, topic centric news sets. [23] For this purpose, the paper proposes Dual-Level Clustering Model on the news representation

| Methods | F1 | NMI |
|---|---|---|
| TFIDF | 0.205 | 0.172 |
| TFIDF with Embedding | 0.405 | 0.39 |
| Text Rank with Embedding | 0.389 | 0.372 |
| Doc2vec | 0.385 | 0.242 |
| TFIDF with Embedding & Time2vec | 0.616 | 0.67 |
| Text Rank with Embedding & Time2vec | 0.599 | 0.648 |
| Doc2vec with Time2vec | 0.679 | 0.744 |
| DC TFIDF with Embedding & Time2vec | 0.72 | 0.708 |
| DC Text Rank with Embedding & Time2vec | 0.691 | 0.621 |
| Our Method (DC NR-Time) | 0.756 | 0.842 |

Table 2.3: Yu Shuqi, et. al. - Experimentation results based on evaluation metrics [23]

with Time2vec. The use of key-entities has also been taken into account while Dual-Level Clustering. Each detected eventis driven by a typical news topic, with clear structured characteristics: 4W+H: [What, Who, Where, When, How], which is a classical schema in Journalism. The research evaluates the performance of proposed methodology with identified baselines based on the two evaluation metrics f1-measure and Normalized Mutual Information (NMI) is used for the quality of cluster set. It is normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation). [The proposed Dual level clustering methodology (DC NR with Time) performed with much better results as shown in Table.2.3.

Souneil Park, et. al. worked on very rare topic Disputant Relation-Based Classification for Contrasting Opposing Views of Contentious News Issues. [24] Contentious issues continuously arise in various domains, such as politics, economy, environment; each issue involves diverse participants and their different complex arguments. However, news articles are frequently biased and fail to fairly deliver conflicting argu-

ments of the issue. It is difficult for ordinary readers to analyze the conflicting arguments and understand the contention; they mostly perceive the issue passively, often through a single article. In this paper, a disputant relation-based method is presented for classifying news articles on contentious issues. The observation on disputants of a contention, i.e. people such as politicians, companies, stakeholders, civic groups, experts, commentators, and so on, are important feature for understanding discourse. Corpus: 14 contentious issues from Naver News (a popular news portal in Korea) issue archive. The randomly sampled about 20 articles per issue, for a total of 250 articles. The selected issues range over diverse domains including politics, local, diplomacy, economy. [24]

In the study, the initial significant part is to identify the contention and the impacting key opponents. The paper has developed key opponent-based partitioning method for disputant partitioning. To identify the key opponents of the issue, the study searches for the disputants who frequently criticize, and are also criticized by other disputants. The sentence is considered to express the disputants criticism to another disputant if the following holds: 1. the sentence is a quote, 2. the disputant is the subject of the quote, 3. another disputant appears in the quote, and 4. a negative lexicon appears in the sentence. The HITS algorithm is designed to rate webpages regarding the link structure. The feature of the algorithm is that it separately models the value of outlinks and inlinks. Each node, i.e., a webpage, has two scores: The authority score, which reflects the value of inlinks toward itself, and the hub score, which reflects the value of its outlinks to others. The hub score of a node increases if it links to nodes with high authority score, and the authority score increases if it is pointed by many nodes with high hub score. Due to above feature, it enables separately measure the significance of a disputants criticism (using the hub score) and criticism about disputant (using the authority score).

The developed method performs disputant extraction, disputant partitioning, and article classification in sequence. [24] The study applies a modified version of HITS algorithm to identify the key opponents of an issue and used disputant extraction techniques combined with an SVM classifier for an article analysis. The paper depicts that the method achieves acceptable performance for practical use with basic language resources and tools, i.e., named entity recognizer, POS tagger, and a translated positive/negative lexicon. To deal with non-English (Korean) news articles, it is difficult to obtain rich resources and tools, for example, WordNet, dependency parser, annotated corpus such as MPQA. When applied to English, the study believes that the method could be further improved by adopting them. The study conducts an accuracy analysis and an upper-bound analysis for evaluation of the method.

Zhenzhong Li, et. al. paper proposed a multi-class text classification model based on Softmax regression as shown in Figure 2.16. [25] The Softmax regression basically compute the probability that the sample belongs to the class, and then the model selects the one with the greatest probability as the final class result. Latent Dirichlet Allocation (LDA) is a kind of topic model algorithm based on probability model. The algorithm thinks that each article is composed of a plurality of topic mixture. It can identify potential hidden information topic in large-scale document set. The algorithm assumes that each word in the corpus in an article is through by "with a certain probability to choose a topic, and then from this subject with a certain probability to select a word".

Logistic Regression is classic regression algorithm for classification. But it is just used for two-category. Softmax Regression is the extention of the Logistic Regression. Multi-class classification of Softmax Regression is to solve Multi-class problems. And the improved model makes full use of the Softmax Regression to news text classification

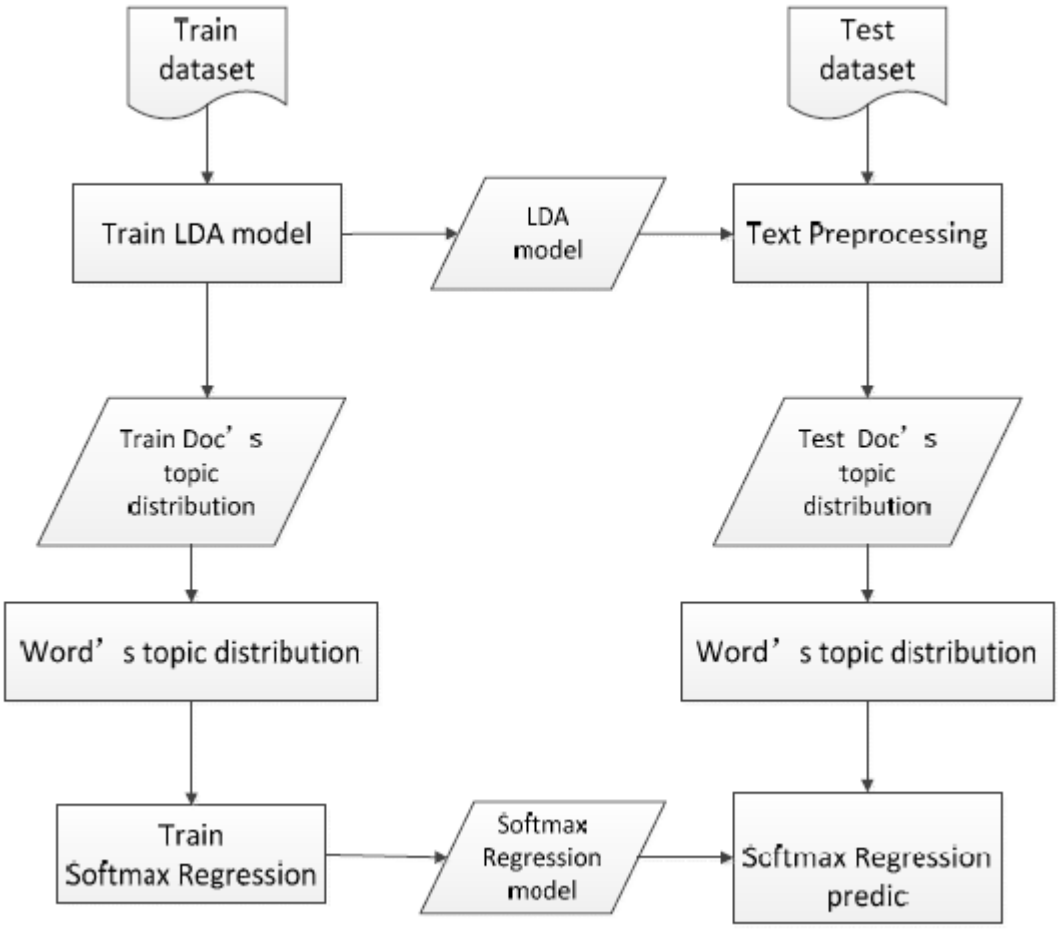



Figure 2.16: Zhenzhong Li, et. al. - The proposed classification process [25]

model for text mining. The model uses Softmax Regression with topic vector for news text. It finds out that result of text classification is very effective. [25]

Chenbin Li, Guohua Zhan, Zhihua Li et. al. focuses on text classification problem of NLP by using the Bi-LSTM-CNN method. [26] The traditional document classification methods require a large of amount of artificial intelligence and human participation. For the purpose of improving the accuracy of text classification, a kind of comprehensive expression is employed to accurately express semantics. The Bi-LSTM-CNN model utilizes the loop structure to obtain the context information and constructs the left and right contexts of each word through the Convolutional Neural Network (CNN) to construct the textual expression of the word, which is more accurately expressed the

semantics of the text. The 96.45% accuracy obtained with Bi-LSTM-CNN. the proven classification methods are used as baselines and experimental results are as shown in Table. 2.4.

| **Model** | **The Comparison experimental results** | | |
|---|---|---|---|
| | **Accuracy** | **Loss** | **F1** |
| TF-IDF | 90.27 | 0.31 | 0.86 |
| SVM | 93.49 | 0.25 | 0.92 |
| LSTM | 94.26 | 0.21 | 0.94 |
| CNN | 95.61 | 0.15 | 0.96 |
| Bi-LSTM- CNN | 96.45 | 0.11 | 0.99 |

Table 2.4: Chenbin Li, et. al. - The Comparison experiment results [26]

Priya P. Raut, et. al. worked on classification of controversial news documents based on disputant relation using SVM. [27] The method suggests, Opponent-based frame to understand controversial issues. For better performance of Disputant relation-based method, Naive Byes Classifier can be used to classify articles. The disputant relation-based method uses the opponent-based frame for classification. The aim of method is to recognize the two groups of the issue. The method usually works on two groups of disputants. They try to control readers' understanding, estimate of the issue, and obtained support from them. The method consists of three steps: Disputant Mining, Disputant Partitioning, Article Classification.

Syafruddin Syarifm, et. al. researched on topic trending topic trediction by Optimizing k-NN. [28] The research aimed at assisting the government of Makassar City to predict the trending topic which would happen by analysing the historical stack in the data mining. The research concludes the K-NN method obtains 81.13% by optimizing the value of K. Dataset used in the study: The news and conversation taken from the online and social media related to Makassar City Government with 393.667 raw data.

| **Method** | **Accuracy** | **Training Time (s)** | **Testing Time (s)** |
|---|---|---|---|
| SVM | 88.54% | 694.577 | 378.638 |
| ELM with kernels | 88.74% | 5.21 | 5.175 |

Table 2.5 Xueying Zhang, et. al. - Comparison between ELM and SVM [42]

Xueying Zhang, et. al. worked on Sentiment Analysis with Chinese language. Due to the huge difference between English and Chinese in syntax, semantics and pragmatics etc., there are problems in the processing of Chinese text.[42] The study uses SVM and ELM with Kernel classifiers with TF-IDF weighing method and have used hotel BBS comments dataset. Extreme Learning Machine was first proposed by Huang in 2006, ELM with kernels is a single-layer feedforward network, and it has more effectively to regression prediction compared with the basic ELM algorithm. The study started with the assumption, that as for the support vector machine algorithm, ELM with kernels can get better or similar predictive accuracy with less time. Compared with SVM, although the accuracy for ELM with kernels is similar to SVM, but the training and testing time in ELM with kernels will be far less. The experiment results show that ELM with kernels method of emotional polarity analysis of Chinese text is more effective. The study results are as shown in Table 2.5.

Vishal S. Shirsat, et. al. worked on document level sentiment analysis from news articles. [36] Sentiment analysis has been divided mainly in three levels - document level, sentence level, entity and aspect level. The paper proposed ontology to effectively find the polarity of any news articles as Positive, Negate and Neutral. The BBC news dataset is used in the study for sentiment analysis of news articles.

# Chapter 3

# Research Hypothesis

The purpose of the project work was to determine the best features for the news articles classification. Entity and Key-phrases are the identified features which play an important role in almost every domain of text document and can majorly contribute to the news documents classification. The second purpose of the work was to achieve improved accuracy in the news articles classification based on the selected features. To conduct the comparative study of the considered features and the classifiers accuracy based on obtained experimental results.

For the objective mentioned above, the Count and TF-IDF vectorizer is studied and hence identified the term frequency can evaluated for each document and can be selected for the further work when it compared to count.

For example, we have an article that contains a word credit 15 times and the word banking 2 times. Here, if we just used term frequency, more weight to the word credit will be given compared to the word banking, since it occurs more frequently in the article. However, the word credit might frequently be occurring across multiple categories whereas the word banking might be occurring in very few categories that may related to business, or any other. Thus, the word banking is a more distinguishing feature in the document. In inverse document frequency, we determine the distinguishability of the word which

then multiply with term frequency to get the new weight of each word in the document. So, considering the inverse document frequency along with term frequency should help in better news articles classification. Further, train and evaluate the classification models based on the extracted features from news articles with k-Nearest Neighbor, Support Vector Machine (linear SVC), Random Forest, Logistic Regression lbfgs & liblinear, Linear Regression with SGD, Multinomial Naive Bayes.

## 3.1 Alternative Hypothesis

Considering selected features Entities, Key-phrases out of news articles corpus, TF-IDF for feature representation of each news articles, and the applying classifiers on the extracted features should result in accuracy improvement of text document classification model by up to 5 percent.

## 3.2 Null Hypothesis

Considering selected features Entities, Key-phrases out of news articles corpus, TF-IDF for feature representation of each news articles, and the applying classifiers on the extracted features should result in accuracy improvement of text document classification model by 5 percent.

# Chapter 4

# Dataset

A high-quality dataset should contain:

- balanced taxonomy,
- sufficient amount of data,
- high quality information in data and labels,
- minimum data and label errors,
- relevant to your problems
- diversify. [48]

| **Category** | **News Articles** |
|---|---|
| Business | 510 |
| Entertainment | 386 |
| Politics | 417 |
| Sports | 511 |
| Technology | 401 |
| **Total** | **2225** |

Table 4.1: BBC News Dataset [49]

The selected dataset for the study is BBC News Dataset which consists of 2225 news articles from five categories Business, Entertainment, Politics, Sports, Technology. The train and test dataset splitted into 70-30 (%) ratio using cross validation technique. The description of BBC News Dataset is given in Table 4.1. [49]

# Chapter 5

# Terms

**i. Entity**

An entity is a textual reference to the unique name of a real-world object such as people, places, and commercial items, and to precise references to measures such as dates and quantities. [50]

**ii. Key-phrases**

A key phrase is a string containing a noun phrase that describes a particular thing. It generally consists of a noun and the modifiers that distinguish it. For example, "day" is a noun; "a beautiful day" is a noun phrase that includes an article ("a") and an adjective ("beautiful"). [51]

**iii. TF-IDF**

Term Frequency - Inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is used as a weighting factor in searches of information retrieval, text mining, and user modelling. [54]

### iv. Entity Types

| Type | Description |
|---|---|
| COMMERCIAL_ITEM | A branded product |
| DATE | A full date (for example, 11/25/2017), day (Tuesday), month (May), or time (8:30 a.m.) |
| EVENT | An event, such as a festival, concert, election, etc. |
| LOCATION | A specific location, such as a country, city, lake, building, etc. |
| ORGANIZATION | Large organizations, such as a government, company, religion, sports team, etc. |
| PERSON | Individuals, groups of people, nicknames, fictional characters. |
| QUANTITY | A quantified amount, such as currency, percentages, numbers, bytes, etc. |
| TITLE | An official name given to any creation or creative work, such as movies, books, songs, etc. |
| OTHER | Entities that don't fit into any of the other entity categories. |

Table 5.1: Entity Types in AWS Comprehend [50]

### v. K-Means

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster) [56]

**vi. k-NN**

A k-nearest-neighbor is a data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it. An algorithm, looking at one point on a grid, trying to determine if a point is in group A or B, looks at the states of the points that are near it. The range is arbitrarily determined, but the point is to take a sample of the data. If the majority of the points are in group A, then it is likely that the data point in question will be A rather than B, and vice versa. The k-nearest-neighbor is an example of a "lazy learner" algorithm because it does not generate a model of the data set beforehand. The only calculations it makes are when it is asked to poll the data point's neighbors. [56]

**vii. LDA**

In natural language processing, Latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a topic model. [58]

**viii. Topic Modelling**

Topic modeling is the process of identifying topics in a set of documents. This can be useful for search engines, customer service automation, and any other instance where knowing the topics of documents is important. LDA (Latent Dirichlet Allocation) is a form of unsupervised learning that views documents as bags of words (ie order does not matter). LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words. [57]

**ix. SVM**

This linear classification method can be used for multiclass classification other than the binary classification. SVM which is doing the classification using linear decision boundaries is called as linear SMV and as well as with the little enhancement of the algorithm SVM can be modified for nonlinear classification which uses the non-linear decision boundaries. SVM is a supervised learning algorithm and for a given set of training data this algorithm generates an optimal hyper plane which can use to categorize new data items. SVM is commonly recognized to be a more accurate algorithm [8]

**x. Random Forest**

The Random forest is a statistical method for classification. It was first introduced in 2001 by Leo Breiman. It is a decision-tree based supervised learning algorithm. The Random forest consists of many individual decision trees. Each decision tree votes for classification of given data. The random forest algorithm then accepts the classification which got a maximum number of votes from individual trees. Collectively the decision tree models represent or form a random forest where each decision tree votes for the result and the majority wins. [45]

**xi. Multinomial Naive Bayes**

The Naive Bayes (NB) method is known to be a robust, effective and efficient technique for text classification. More importantly, it can accommodate new incoming training data in classification models incrementally and efficiently. This classifier is suitable to classify discrete features. It is a probabilistic classifier based on text features. Naive Bayes classifier can be trained very efficiently by requiring a relatively trivial quantity of trained data. [8]

**xii. Regression**

A regression problem is when the output variable is a real or continuous value, such as salary or weight. Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points. [52]

**xiii. Classification**

A classification problem is when the output variable is a category, such as red or blue or disease and no disease. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. [52]

**xiv. Clustering**

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. [53]

**xv. Linear Regression**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression

technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. [59]

**xvi. Logistic Regression**

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for classification problems. Despite the name logistic regression this is not a algorithm for regression problems (where the task is to predict a real-valued output). Logistic Regression is a little bit similar to Linear Regression in the sense that both have the goal of estimating the values for the parameters/coefficients, so the at the end of the training of the machine learning model we got a function that best describe the relationship between the known input and the output values. [60]

**xvii. Stochastic Gradient Descent**

Gradient Descent is a very popular optimization technique in Machine Learning and Deep Learning and it can be used with most, if not all, of the learning algorithms. A gradient is basically the slope of a function; the degree of change of a parameter with the amount of change in another parameter. Mathematically, it can be described as the partial derivatives of a set of parameters with respect to its inputs. The more the gradient, the steeper the slope. Gradient Descent is a convex function. Gradient Descent can be described as an iterative method which is used to find the values of the parameters of a function that minimizes the cost function as much as possible. The parameters are initially defined a particular value and from that, Gradient Descent is run in an iterative fashion to find the optimal values of the parameters, using calculus, to find the minimum possible value of the given cost function. [61]

**xviii. LBFGS**

Limited-memory BFGS (L-BFGS or LM-BFGS) is an optimization algorithm in the family of quasi-Newton methods that approximates the BroydenFletcherGoldfarb-Shanno (BFGS) algorithm using a limited amount of computer memory. It is a popular algorithm for parameter estimation in machine learning. The algorithm's target problem is to minimize *f(x)* over unconstrained values of the real-vector $x$ where $f$ is a differentiable scalar function. Like the original BFGS, L-BFGS uses an estimation to the inverse Hessian matrix to steer its search through variable space, but where BFGS stores a dense $n * n$ approximation to the inverse Hessian (n being the number of variables in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. [62]

**xix. Cohen's kappa coefficient**

The kappa statistic is frequently used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured. Measurement of the extent to which data collectors (raters) assign the same score to the same variable is called interrater reliability. While there have been a variety of methods to measure interrater reliability, traditionally it was measured as percent agreement, calculated as the number of agreement scores divided by the total number of scores. In 1960, Jacob Cohen critiqued use of percent agreement due to its inability to account for chance agreement. He introduced the Cohens kappa, developed to account for the possibility that raters actually guess on at least some variables due to uncertainty. Like most correlation statistics, the kappa can range from 1 to +1. While the kappa is one of the most commonly used statistics to test interrater reliability, it has limitations. Judgments about what level of kappa should be acceptable for health research are questioned.

Cohens suggested interpretation may be too lenient for health related studies because it implies that a score as low as 0.41 might be acceptable. Kappa and percent agreement are compared, and levels for both kappa and percent agreement that should be demanded in healthcare studies are suggested. [46] This is essentially a measure of how well the classifier performed as compared to how well it would have performed simply by chance. In other words, a model will have a high Kappa score if there is a big difference between the accuracy and the null error rate.

**xx. Confusion Matrix**

A confusion matrix is a table that is often used to describe the performance of a classification model (or classifier) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix.

**xxi. Underfitting**

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. (Its just like trying to fit undersized pants!) Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have less data to build an accurate model and also when we try to build a linear model with a non-linear data. In such cases the rules of the machine learning model are too easy and flexible to be applied on such a minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection. [64]

### xxii. Overfitting

A statistical model is said to be overfitted, when we train it with a lot of data (just like fitting ourselves in an oversized pants!). When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too much of details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees. [64]

The commonly used methodologies are:

- **Cross-Validation:** A standard way to find out-of-sample prediction error is to use 5-fold cross validation.
- **Early Stopping**: Its rules provide us the guidance as to how many iterations can be run before learner begins to over-fit.
- **Pruning:** Pruning is extensively used while building related models. It simply removes the nodes which add little predictive power for the problem in hand.
- **Regularization:** It introduces a cost term for bringing in more features with the objective function. Hence it tries to push the coefficients for many variables to zero and hence reduce cost term. [64]

### xxiii. Good fit

The case when the model makes the predictions with 0 error, is said to have a good fit on the data. This situation is achievable at a spot between overfitting and underfitting. In order to understand it we will have to look at the performance of our model with the passage of time, while it is learning from training dataset. With the passage of time, our model will keep on learning and thus the error for the model on the training and testing data will keep on decreasing. If it will learn for too long, the model will become more prone to overfitting due to presence of noise and less useful details. Hence the performance of our model will decrease. In order to get a good fit, we will stop at a point just before where the error starts increasing. At this point the model is said to have good skills on training dataset as well our unseen testing dataset. [64]

# Chapter 6

# Proposed System

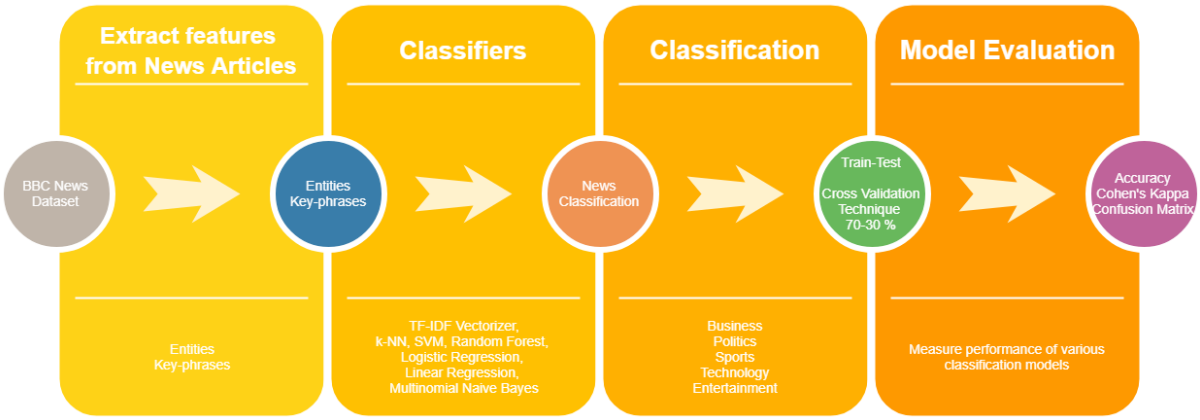



Figure 6.1: The Proposed System - Classification

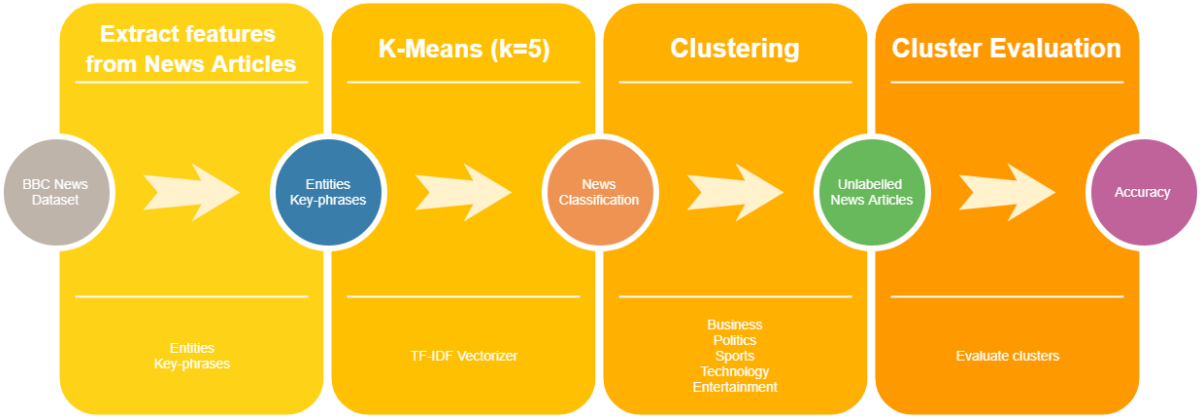



Figure 6.2: The Proposed System - Clustering

**PROCEDURE**

1. Extract Entities and Key-Phrases from BBC News dataset for each news article with AWS Comprehend.

2. Apply the following machine learning - regression and classification models for news articles classification. Split the dataset into 70-30 ratio and perform the classification based on the extracted features - entities and key-phrases from BBC News Articles.

   (a) k-Nearest Neighbor

   (b) Support Vector Machine (linear SVC)

   (c) Random Forest

   (d) Logistic Regression (lbfgs)

   (e) Logistic Regression (liblinear)

   (f) Linear Regression with SGD

   (g) Multinomial Naive Bayes

3. Evaluate each classifiers performance using - Accuracy, Cohens Kappa score, Confusion Matrix.

4. Evaluate significance of Entities and Key-phrases in News Articles Classification.

# Chapter 7

# Implementation

### i. News Article

Japan narrowly escapes recession

Japan's economy teetered on the brink of a technical recession in the three months to September, figures show.

Revised figures indicated growth of just 0.1% - and a similar-sized contraction in the previous quarter. On an annual basis, the data suggests annual growth of just 0.2%, suggesting a much more hesitant recovery than had previously been thought. A common technical definition of a recession is two successive quarters of negative growth.

The government was keen to play down the worrying implications of the data. "I maintain the view that Japan's economy remains in a minor adjustment phase in an upward climb, and we will monitor developments carefully," said economy minister Heizo Takenaka. But in the face of the strengthening yen making exports less competitive and indications of weakening economic conditions ahead, observers were less sanguine. "It's painting a picture of a recovery... much patchier than previously thought," said Paul Sheard, economist at Lehman Brothers in Tokyo. Improvements in the job market apparently have yet to feed through to domestic demand, with private consumption up just 0.2% in the third quarter.

**ii. Entities**

631, b-006.txt, japan japan three months september 0.1% 0.2% two successive quarters japan heizo takenaka paul sheard lehman brothers tokyo 0.2% third quarter, business

**iii. Key-phrases**

1250, b-006.txt, japan recession japan economy the brink a technical recession the three months september figures revised figures growth just 0.1% a similar-sized contraction the previous quarter an annual basis the data annual growth just 0.2% a much more hesitant recovery a common technical definition a recession two successive quarters negative growth the government the worrying implications the data the view japan economy a minor adjustment phase an upward climb developments economy minister heizo takenaka the face the strengthening yen exports indications economic conditions observers a picture a recovery much patchier paul sheard economist lehman brothers tokyo improvements the job market domestic demand private consumption just 0.2% the third quarter, business

**iii. Entity-with-Type**

"Japan","LOCATION"

"Japan","LOCATION"

"three months","QUANTITY"

"September","DATE"

"0.1%","QUANTITY"

"0.2%","QUANTITY"

"two successive quarters","QUANTITY"

"Japan","LOCATION"

"Heizo Takenaka","PERSON"

"Paul Sheard","PERSON"

"Lehman Brothers","ORGANIZATION"

"Tokyo","LOCATION"

"0.2%","QUANTITY"

"third quarter","DATE"

**iv. K-Means Clustering (Entities)**

Top 10 terms per cluster:

Cluster 0

microsoft million china 2004 2005 uk japan appl europ month

Cluster 1

uk bbc british eu week london lord oscar yuko month

Cluster 2

blair brown howard tori dem lib lib dem labour kennedi britain

Cluster 3

olymp open athen australian roddick set second kenteri australian open holm

Cluster 4

england wale chelsea ireland arsenal robinson liverpool franc cup Scotland

**v. K-Means Clustering (Key-phrases)**

Top 10 terms per cluster:

Cluster 0

game player england club team match cup final injuri champion

Cluster 1

music mobil peopl phone game technolog user servic comput digit

Cluster 2

compani bank firm mr market growth economi govern sale price

Cluster 3

film award best oscar actor festiv star nomin director actress

Cluster 4

mr labour elect parti blair tori brown mr blair govern tax

**vi. Entity based Classification using Machine Learning Models**

No. of features extracted: 6862

Train size: (1557, 6862)

Test size: (668, 6862)

**1. k-Nearest Neighbors (k-NN)**

Accuracy score: 0.8293413173652695

Kappa score: 0.7852047623211379

**2. SVM - SVC (Linear)**

Accuracy score: 0.9356287425149701

Kappa score: 0.9190248247943483

**3. Random Forest**

Accuracy score: 0.8697604790419161

Kappa score: 0.8361757426440326

**4. Logistic Regression (lbfgs)**

Accuracy score: 0.9176646706586826

Kappa score: 0.89630783646329

**5. Logistic Regression (liblinear)**

Accuracy score: 0.9161676646706587

Kappa score: 0.8944189485924596

**6. Linear Regression with SGD**

Accuracy score: 0.938622754491018

Kappa score: 0.9228074328990054

**7. Multinomial Naive Bayes**

Accuracy score: 0.9161676646706587

Kappa score: 0.8943775111741071

The two confusion matrices are obtained for each of the entity based classification models respectively.

1. Instance based Confusion Matrix

2. Normalized Confusion Matrix

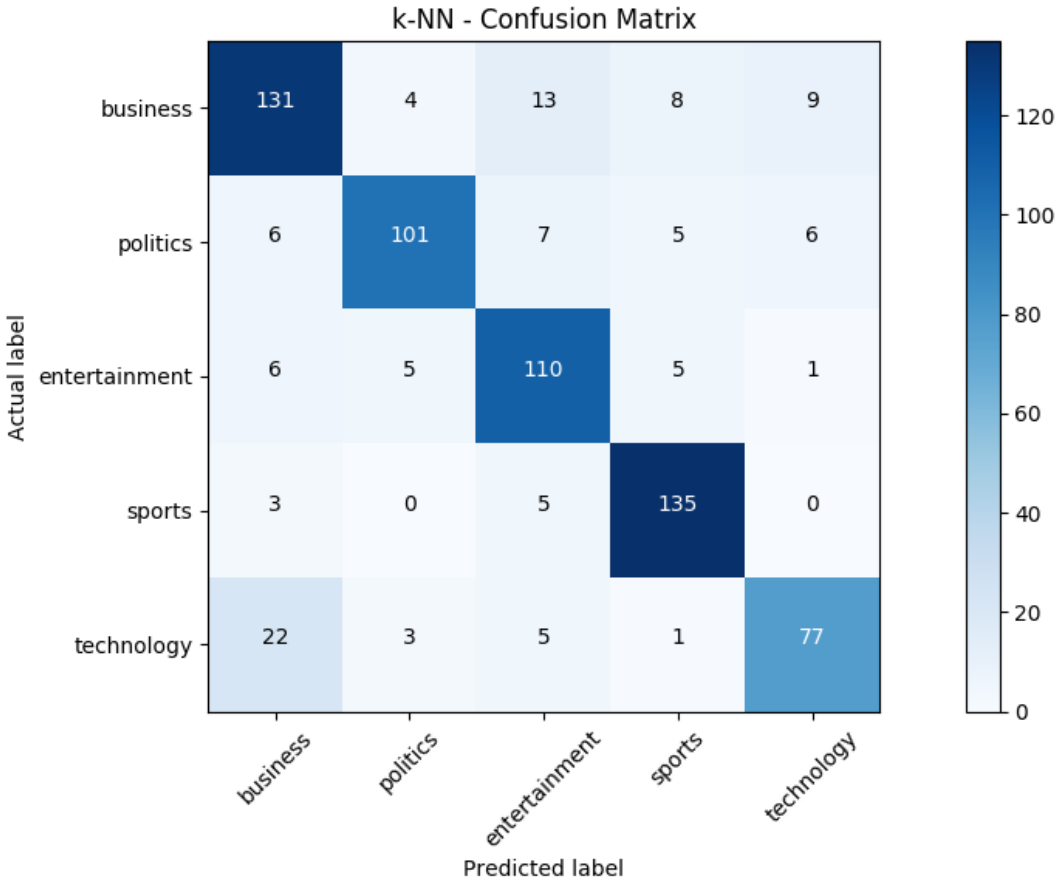



Figure 7.1: Entity based k-NN classification - Confusion Matrix

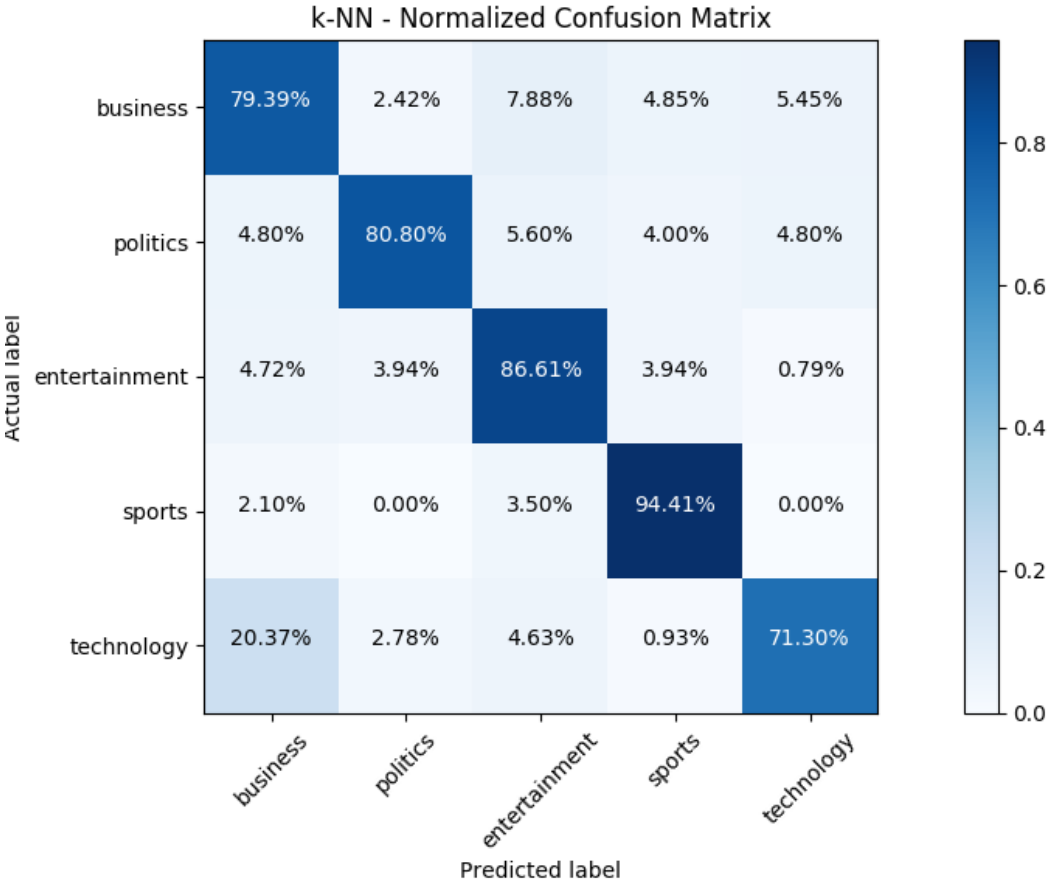



Figure 7.2: Entity based k-NN classification - Normalized Confusion Matrix

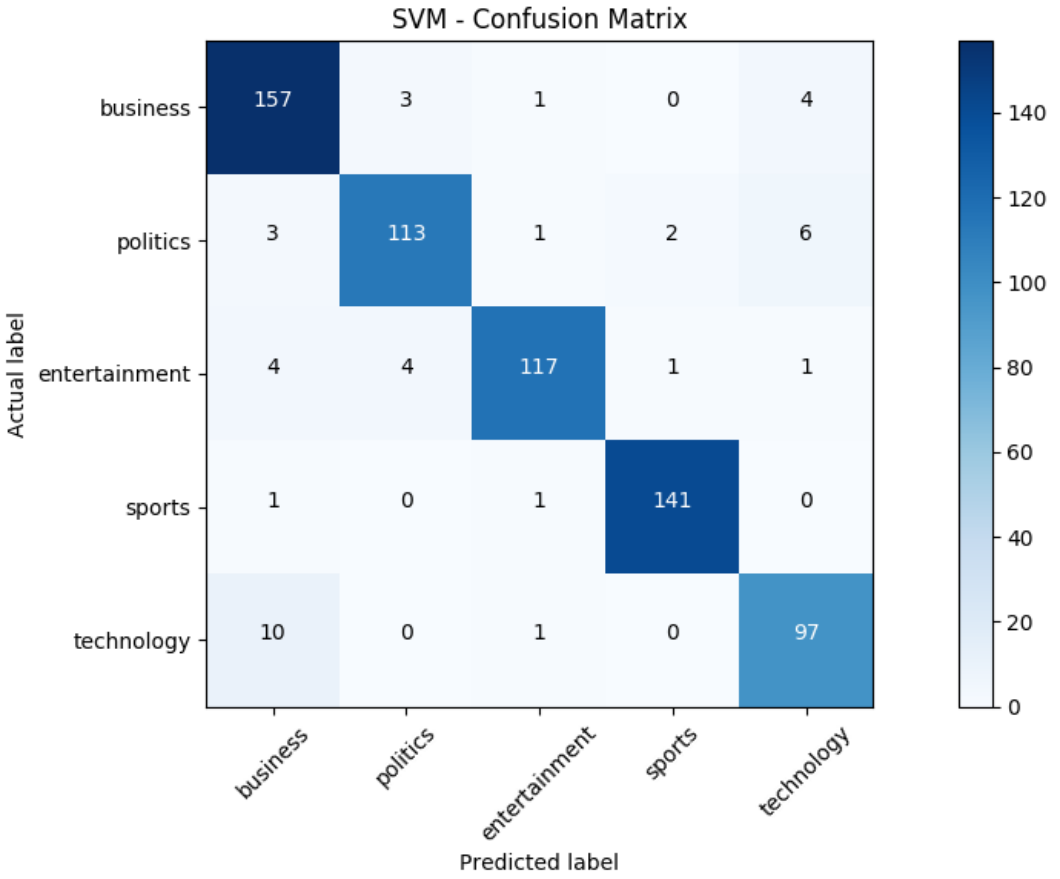



Figure 7.3: Entity based SVM classification - Confusion Matrix

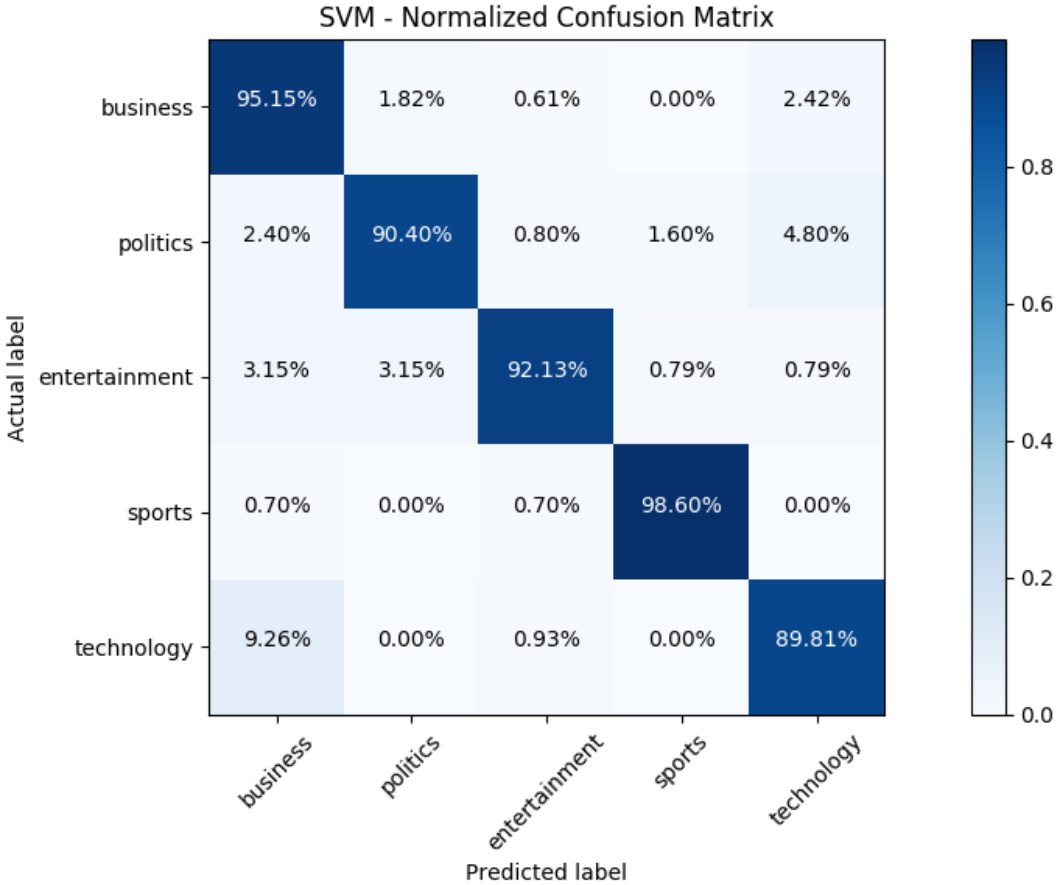



Figure 7.4: Entity based SVM classification - Normalized Confusion Matrix

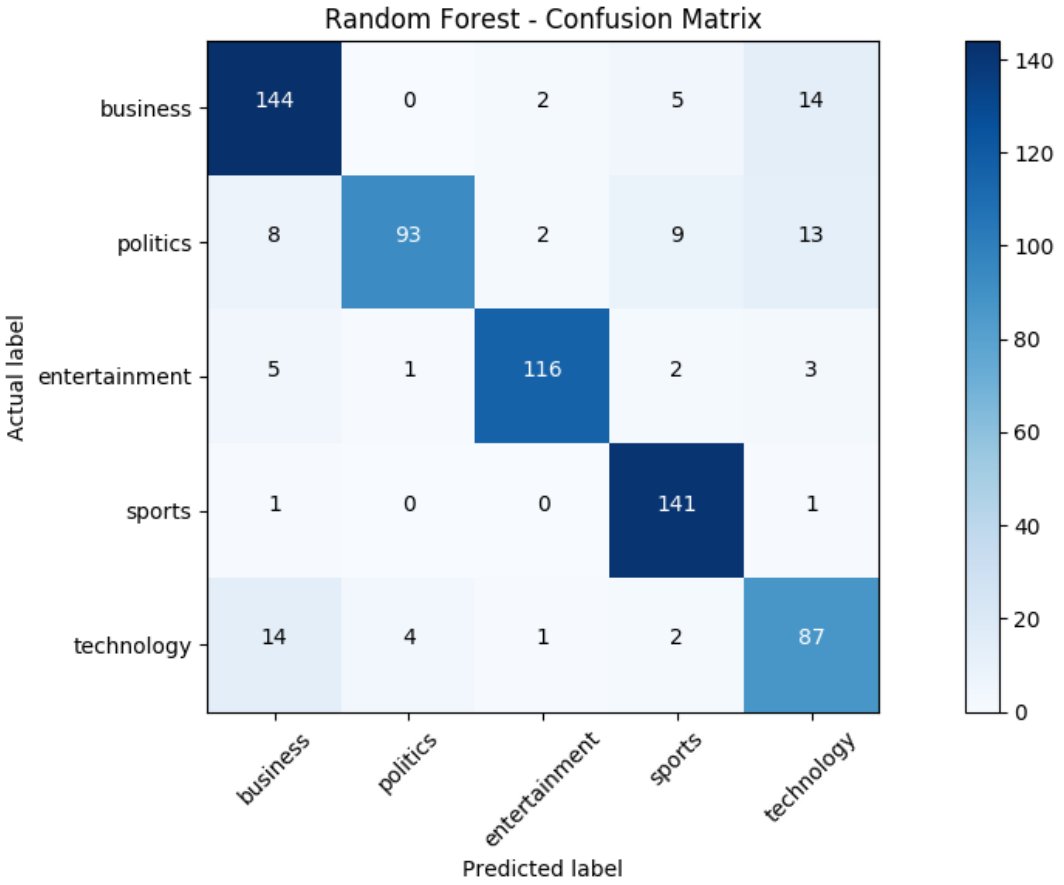



Figure 7.5: Entity based Random Forest classification - Confusion Matrix

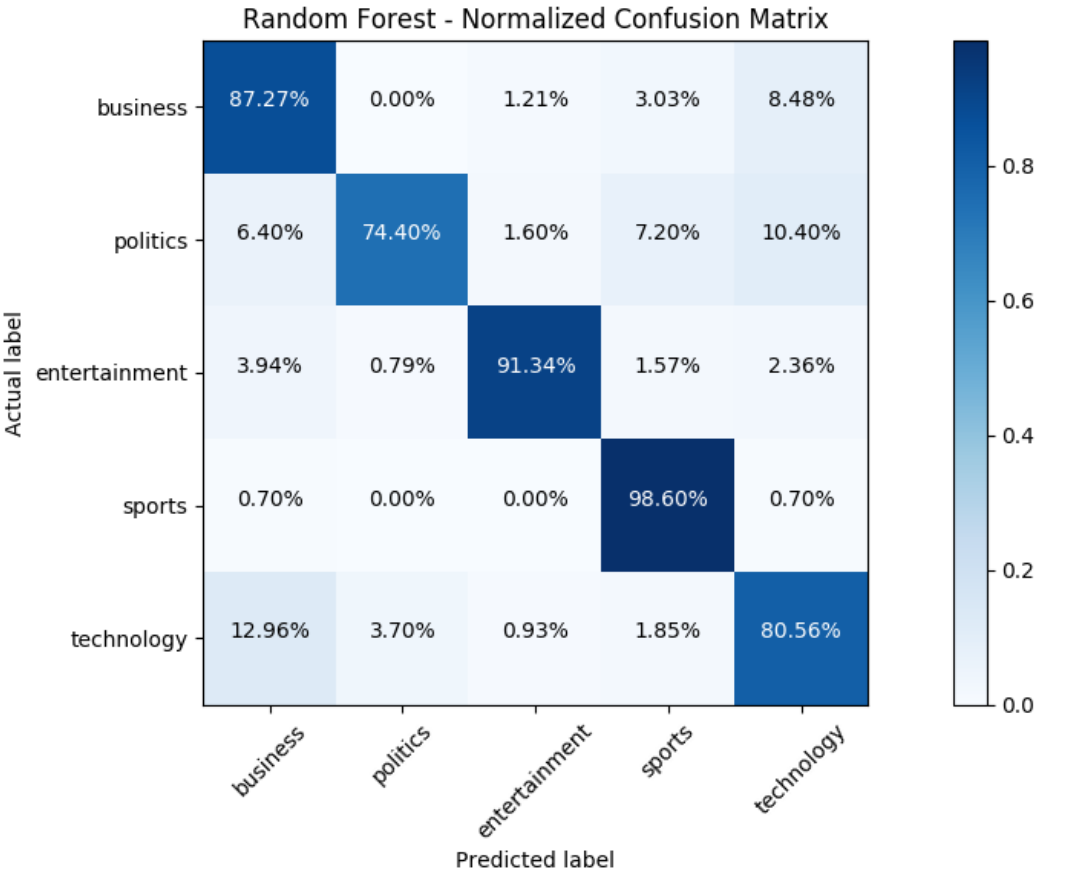



Figure 7.6: Entity based Random Forest classification - Normalized Confusion Matrix

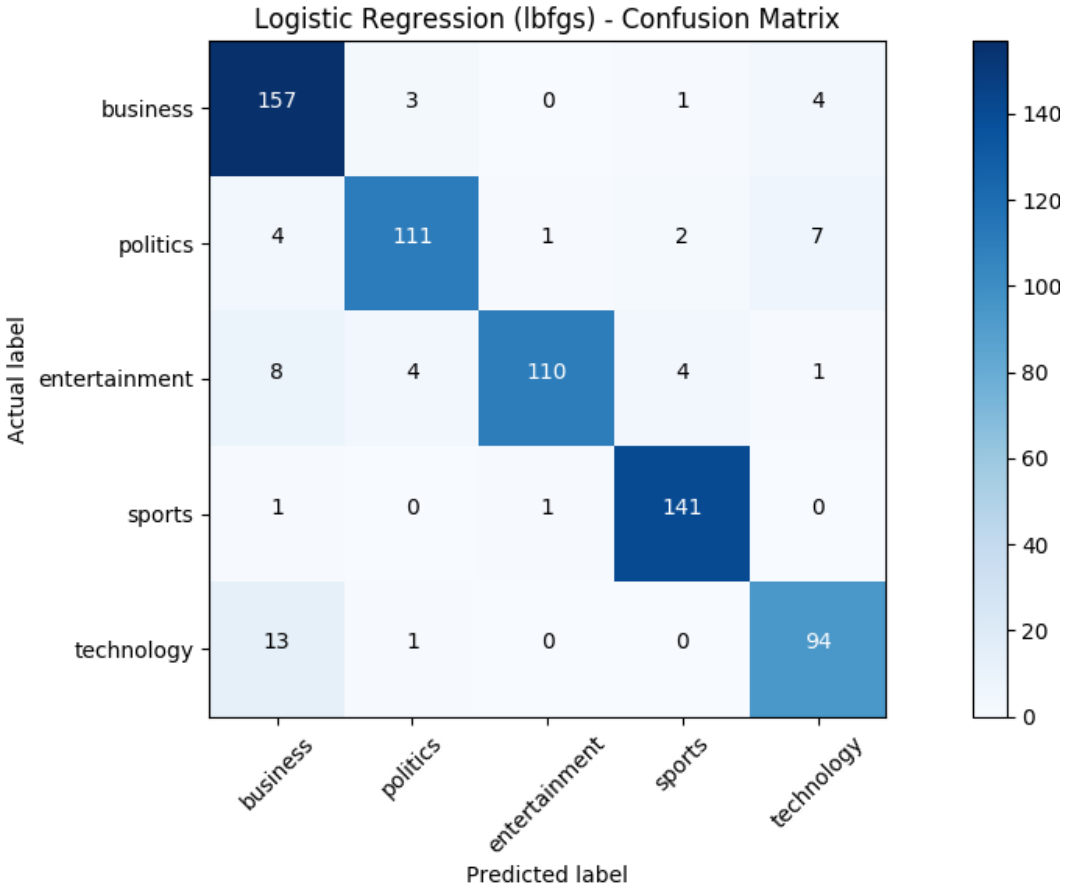



Figure 7.7: Entity based Logistic Regression (lbfgs) - Confusion Matrix

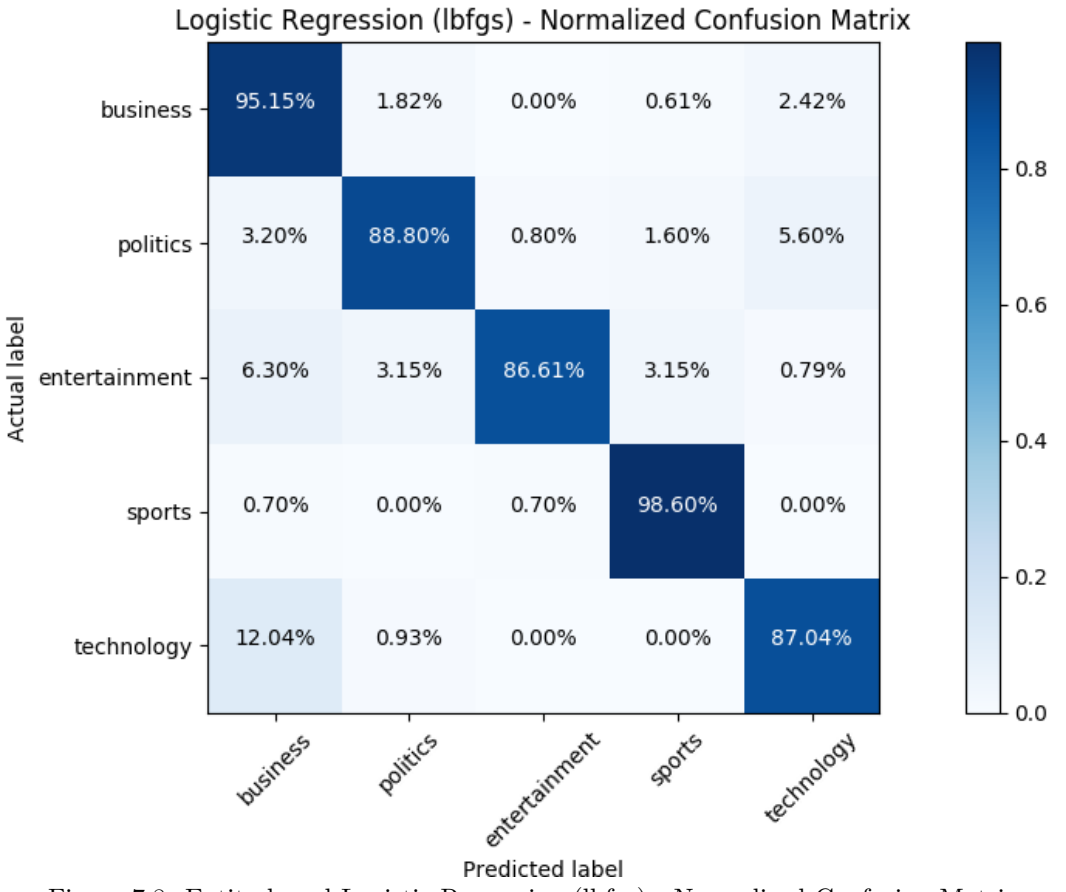



Figure 7.8: Entity based Logistic Regression (lbfgs) - Normalized Confusion Matrix

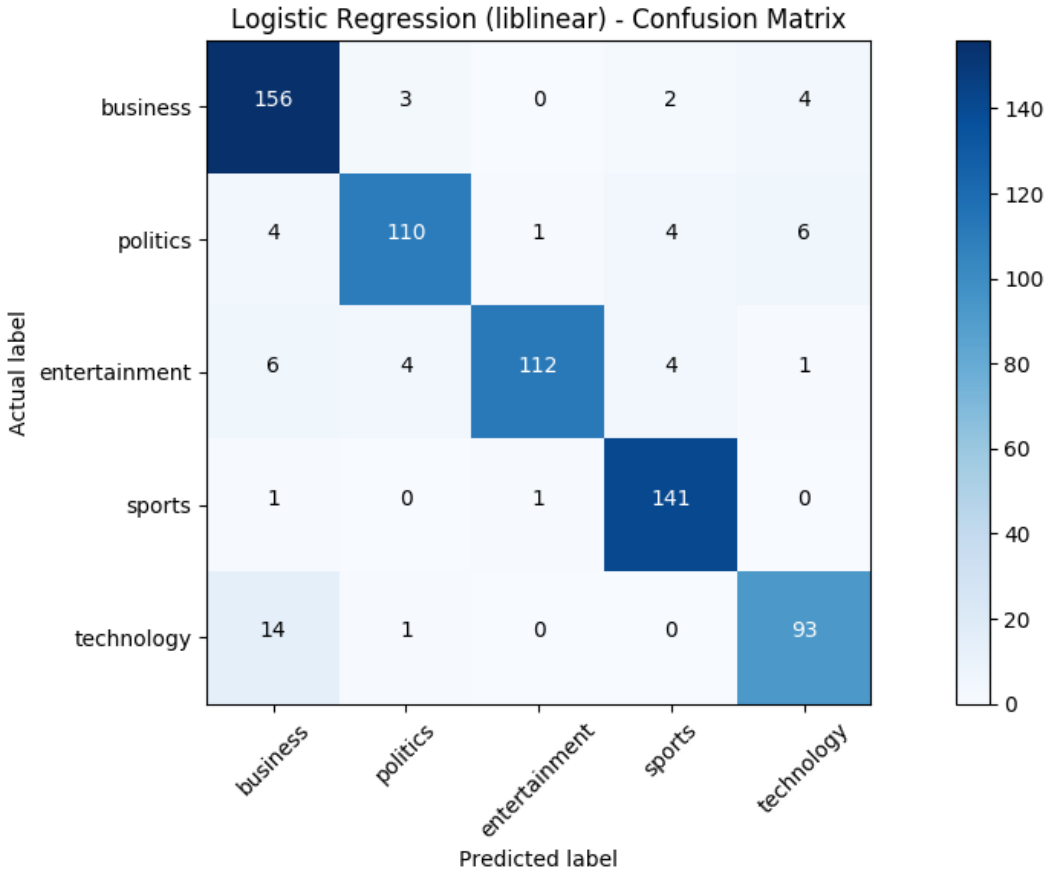



Figure 7.9: Entity based Logistic Regression (liblinear) - Confusion Matrix

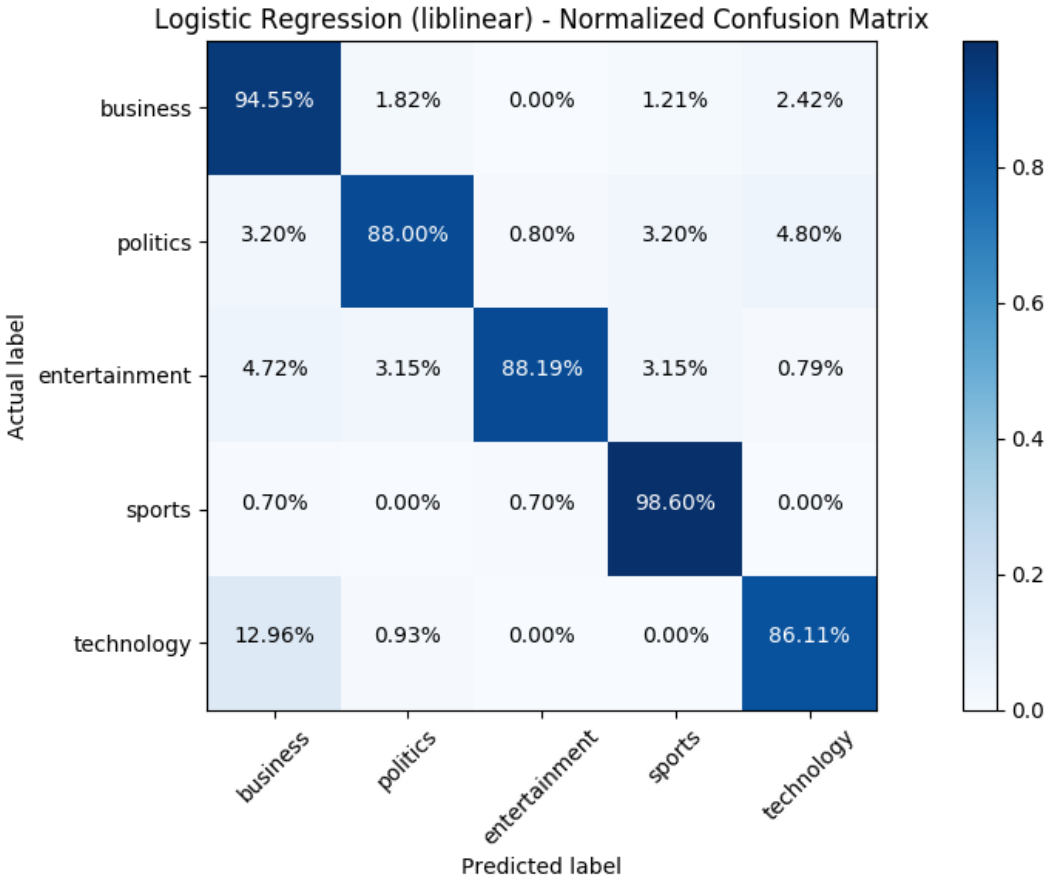



Figure 7.10: Entity based Logistic Regression (liblinear) - Normalized Confusion Matrix

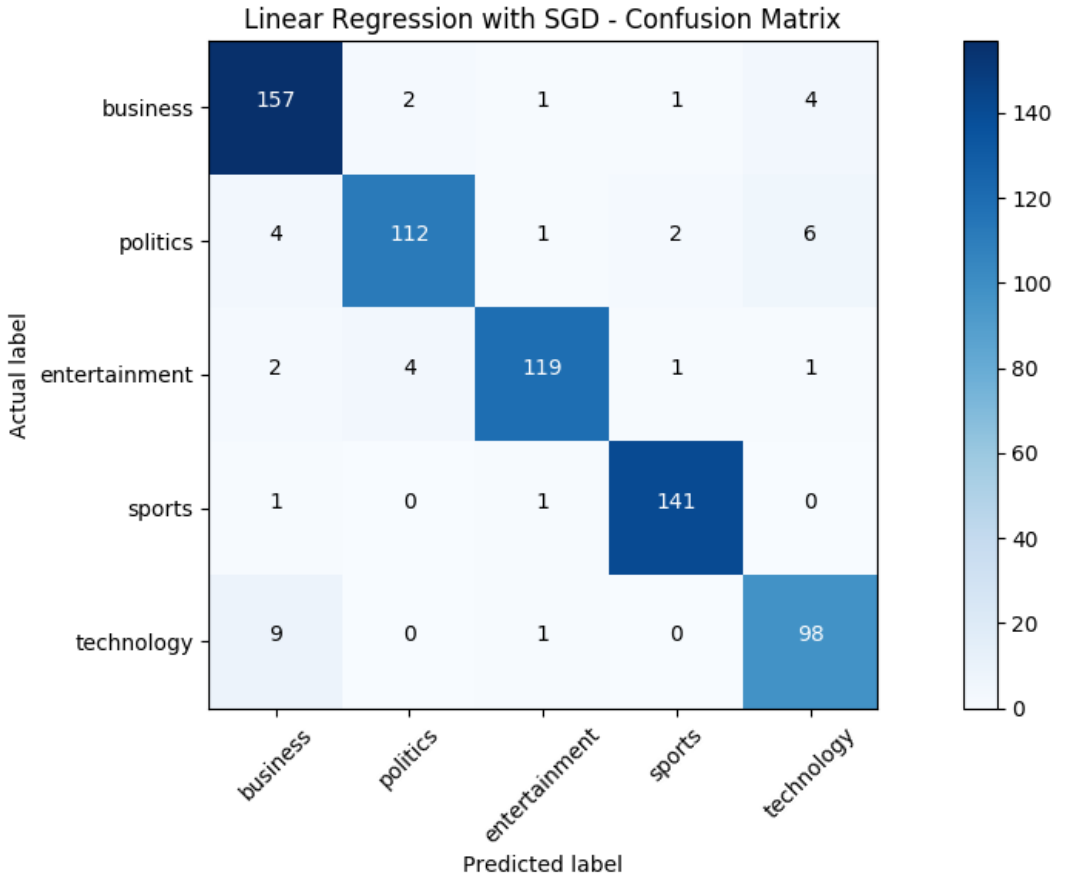



Figure 7.11: Entity based Linear Regression with SGD - Confusion Matrix

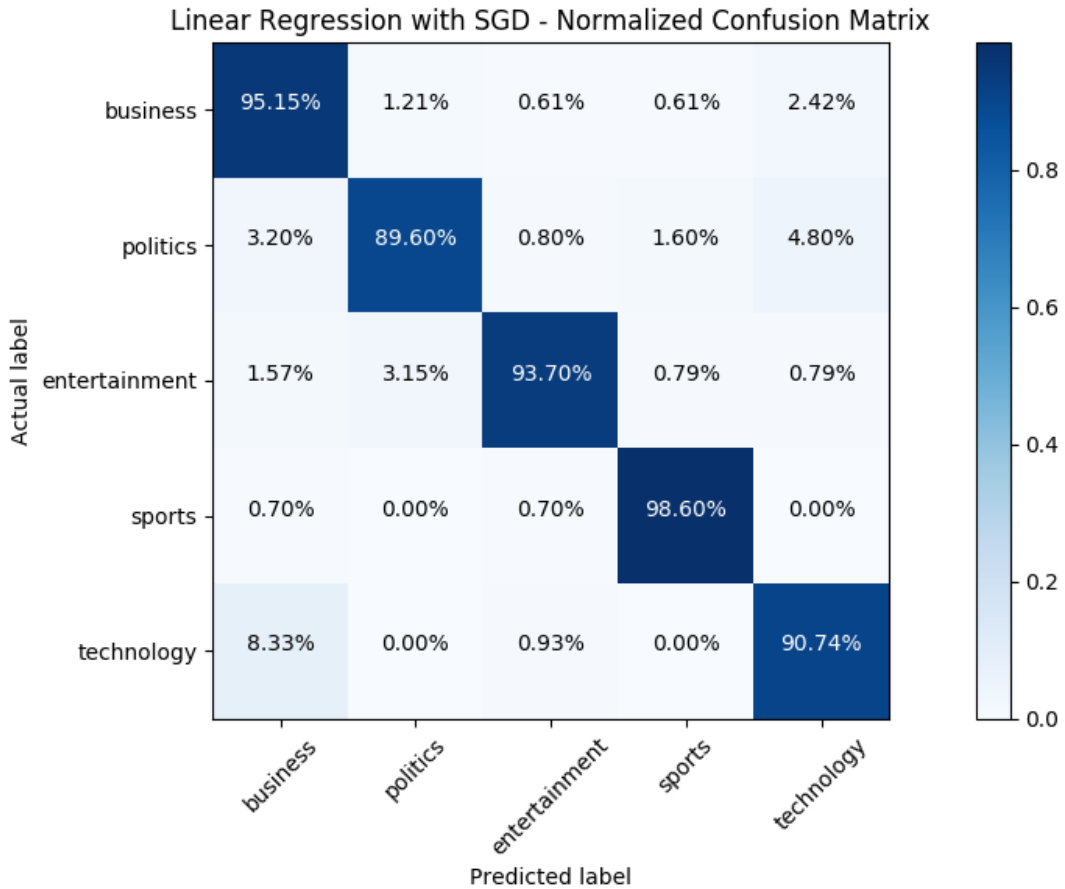



Figure 7.12: Entity based Linear Regression with SGD - Normalized Confusion Matrix

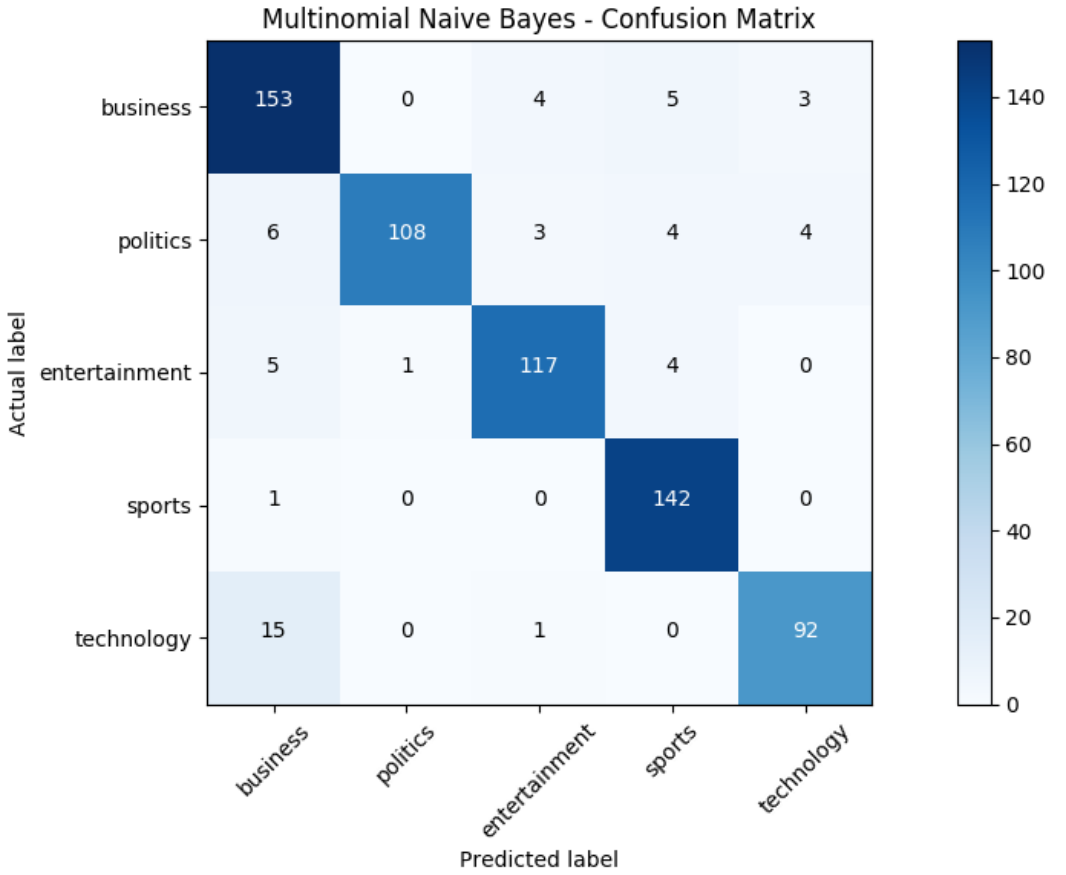



Figure 7.13: Entity based Multinomial Naive Bayes classification - Confusion Matrix

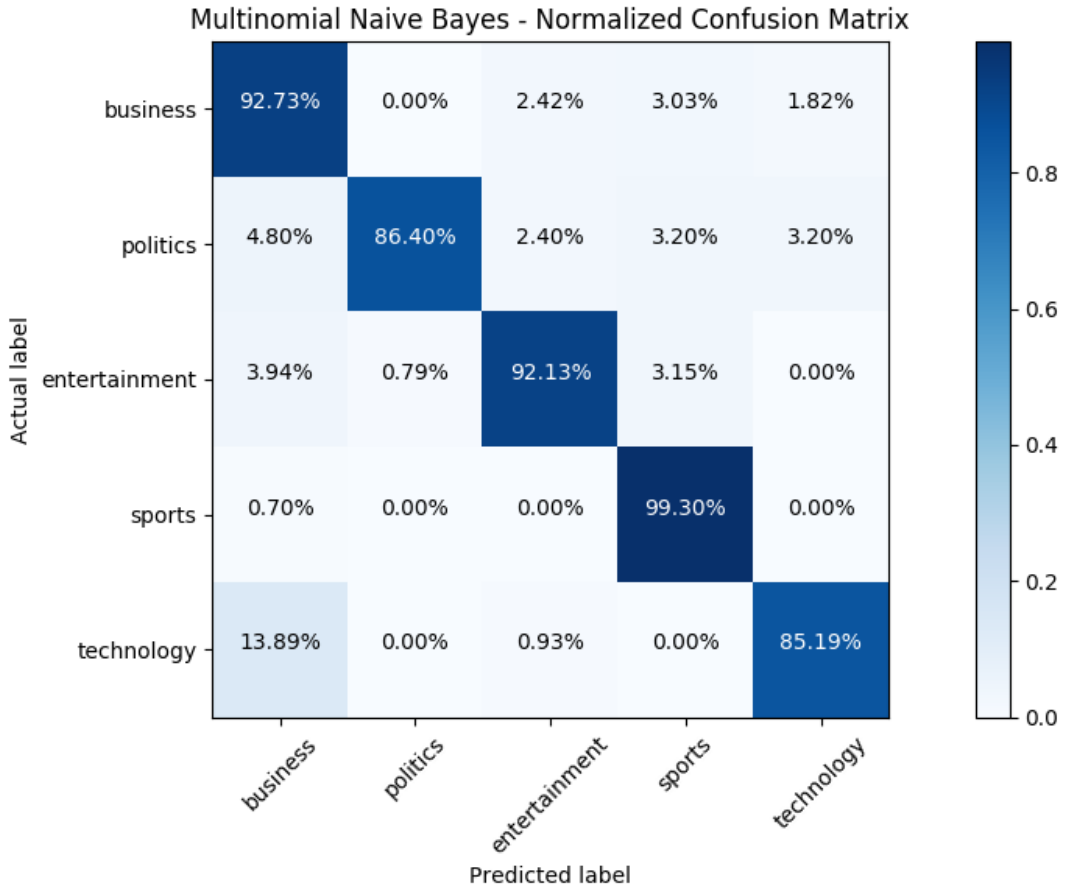



Figure 7.14: Entity based Multinomial Naive Bayes classification - Normalized Confusion Matrix

### vii. Key-phrase based Classification using Machine Learning Models

No of features extracted: 12279

Train size: (1556, 12279)

Test size: (668, 12279)

#### 1. k-Nearest Neighbors (k-NN)

Accuracy score: 0.9281437125748503

Kappa score: 0.9099803755871181

#### 2. SVM - SVC (Linear)

Accuracy score: 0.9760479041916168

Kappa score: 0.9699603987667125

#### 3. Random Forest

Accuracy score: 0.968562874251497

Kappa score: 0.9605461883534193

#### 4. Logistic Regression (lbfgs)

Accuracy score: 0.9760479041916168

Kappa score: 0.9699651260492395

#### 5. Logistic Regression (liblinear)

Accuracy score: 0.9745508982035929

Kappa score: 0.9680840001236618

**6. Linear Regression with SGD**

Accuracy score: 0.9760479041916168

Kappa score: 0.9699678266855867

**7. Multinomial Naive Bayes**

Accuracy score: 0.9655688622754491

Kappa score: 0.956821713922777

The two confusion matrices are obtained for each of the key-phrase based classification models respectively.

1. Instance based Confusion Matrix

2. Normalized Confusion Matrix

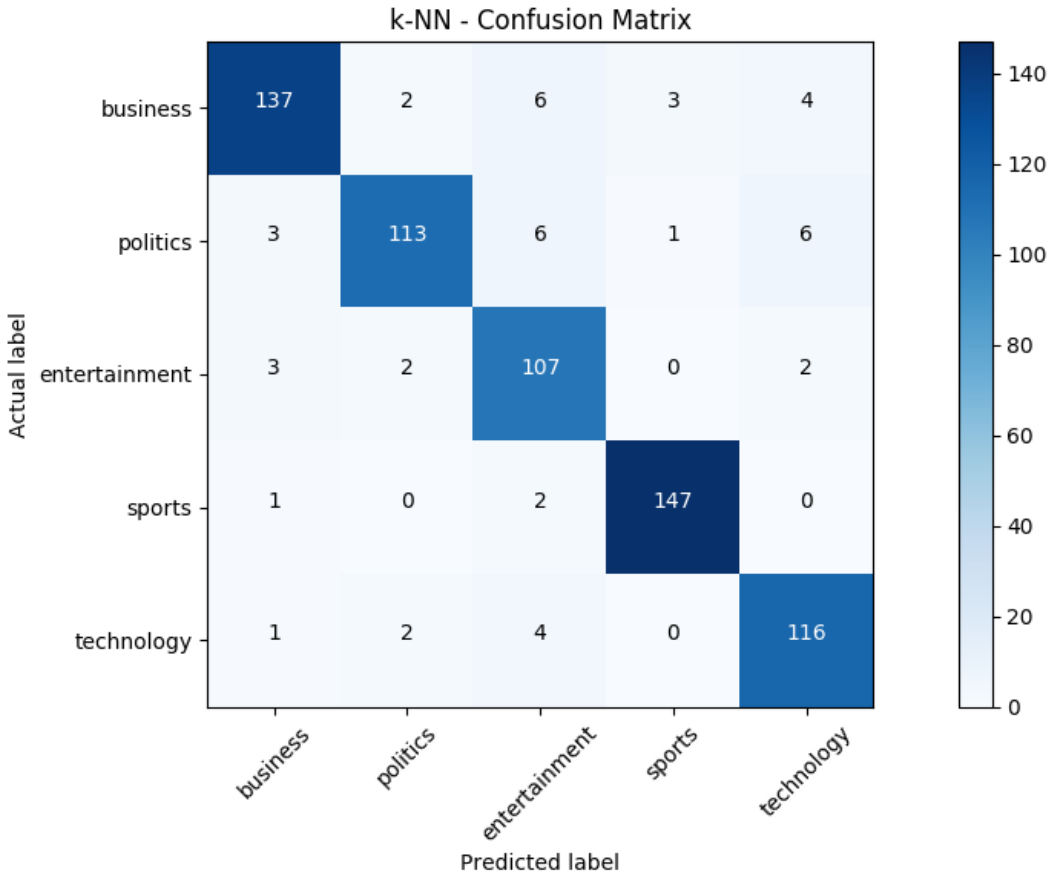



Figure 7.15: Key-phrase based k-NN classification - Confusion Matrix

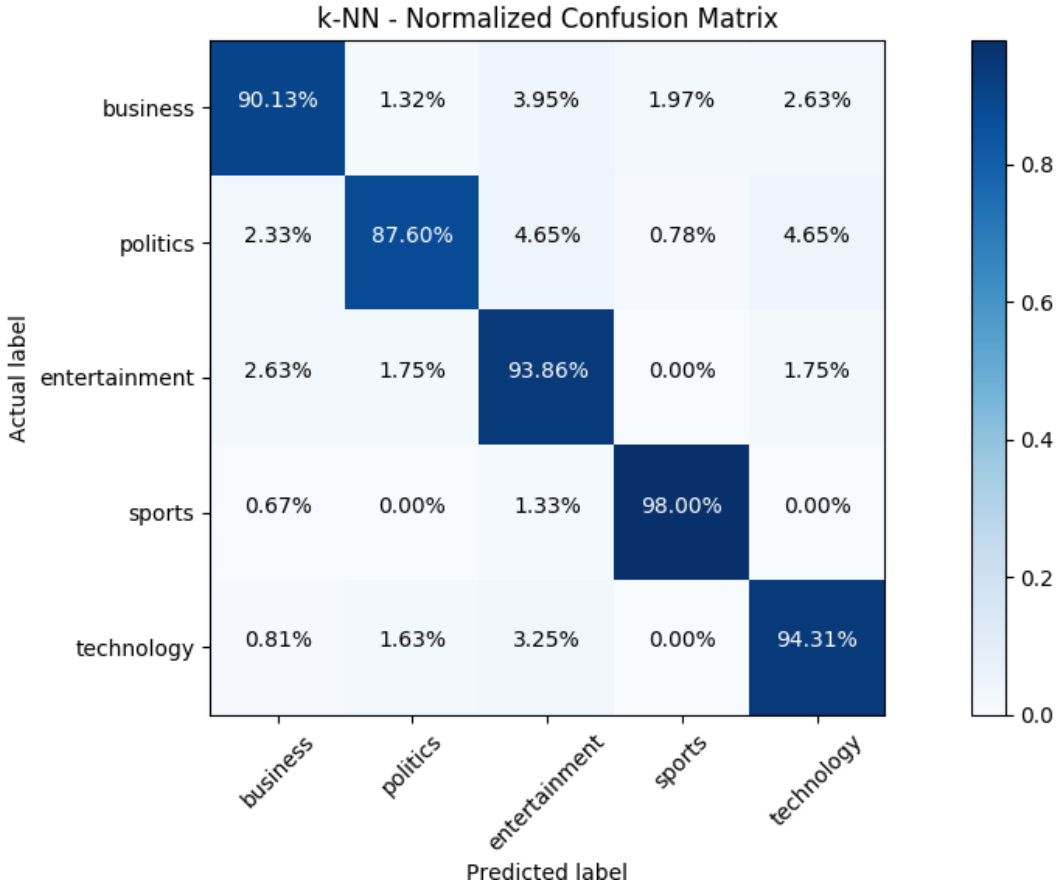



Figure 7.16: Key-phrase based k-NN classification - Normalized Confusion Matrix

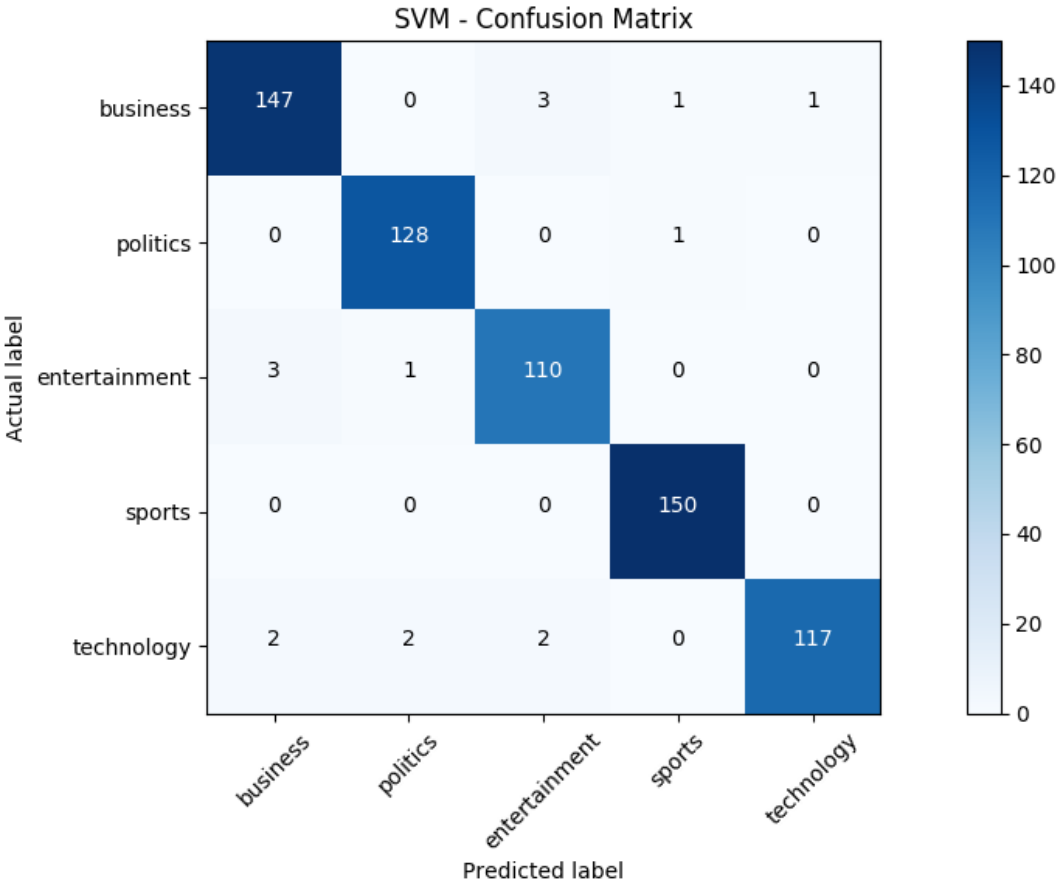



Figure 7.17: Key-phrase based SVM classification - Confusion Matrix

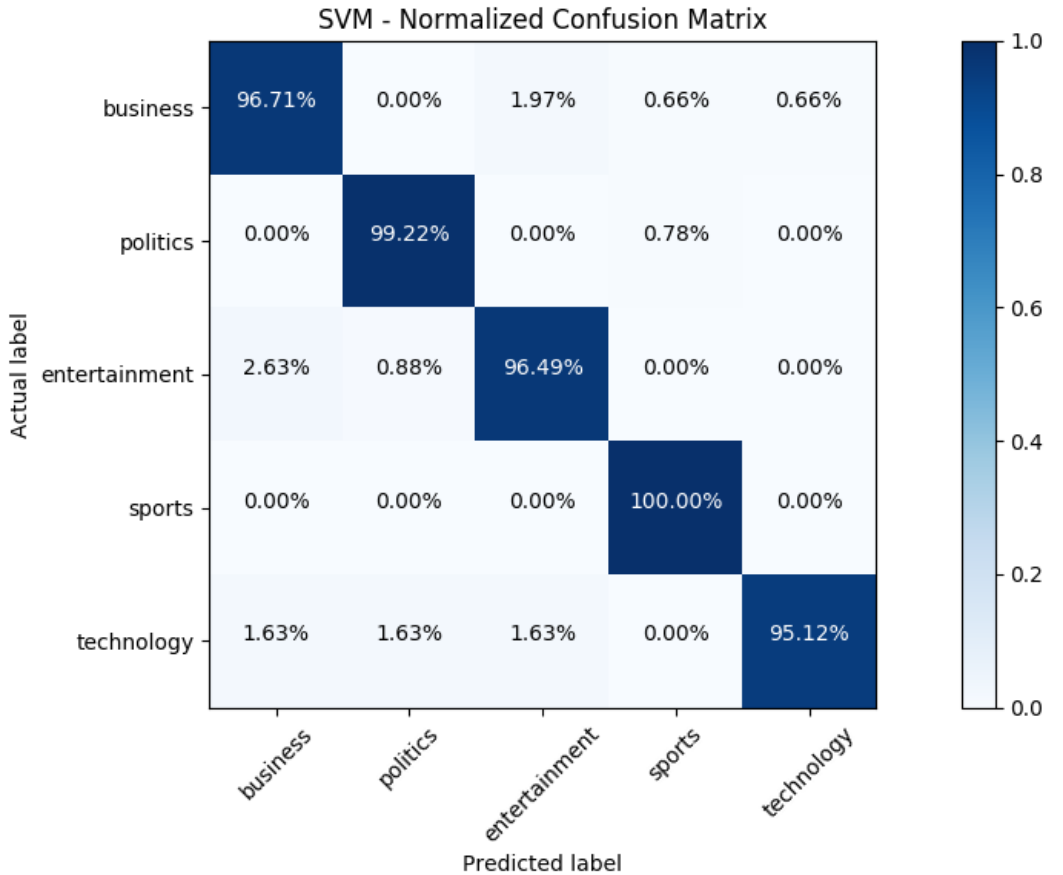



Figure 7.18: Key-phrase based SVM classification - Normalized Confusion Matrix

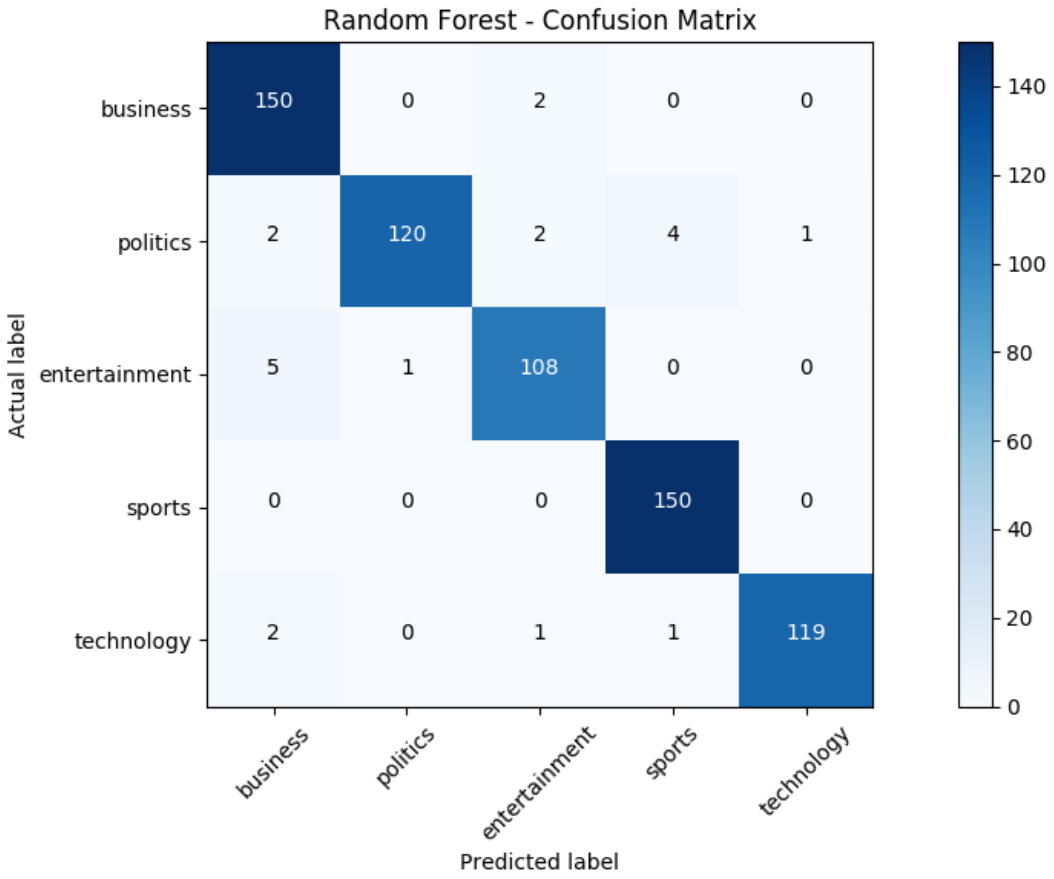



Figure 7.19: Key-phrase based Random Forest classification - Confusion Matrix

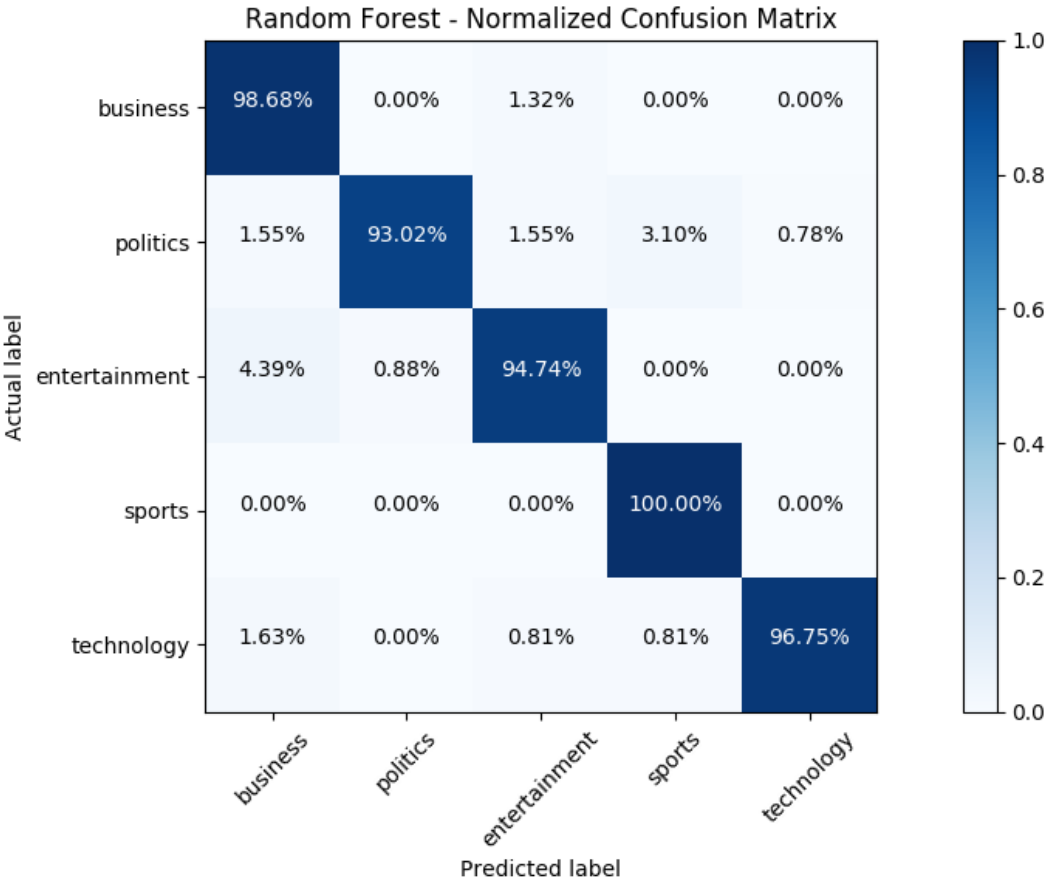



Figure 7.20: Key-phrase based Random Forest classification - Normalized Confusion Matrix

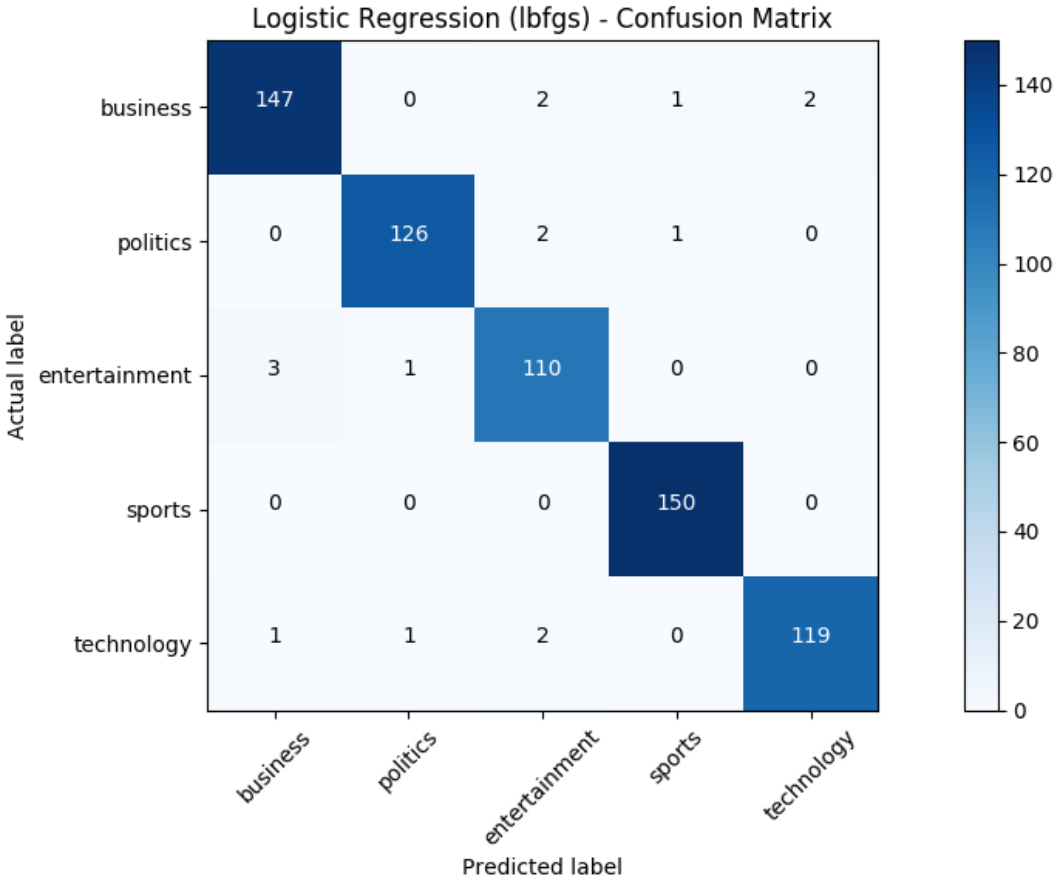



Figure 7.21: Key-phrase based Logistic Regression (lbfgs) - Confusion Matrix

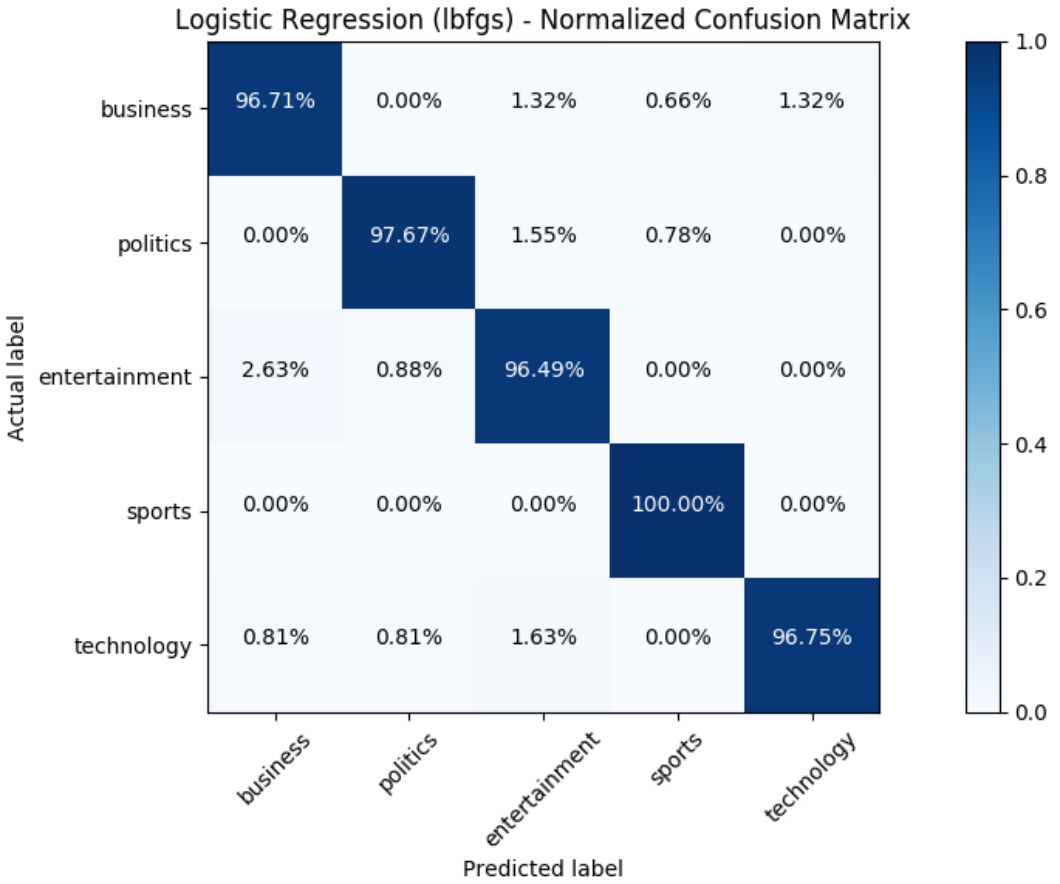



Figure 7.22: Key-phrase based Logistic Regression (lbfgs) - Normalized Confusion Matrix

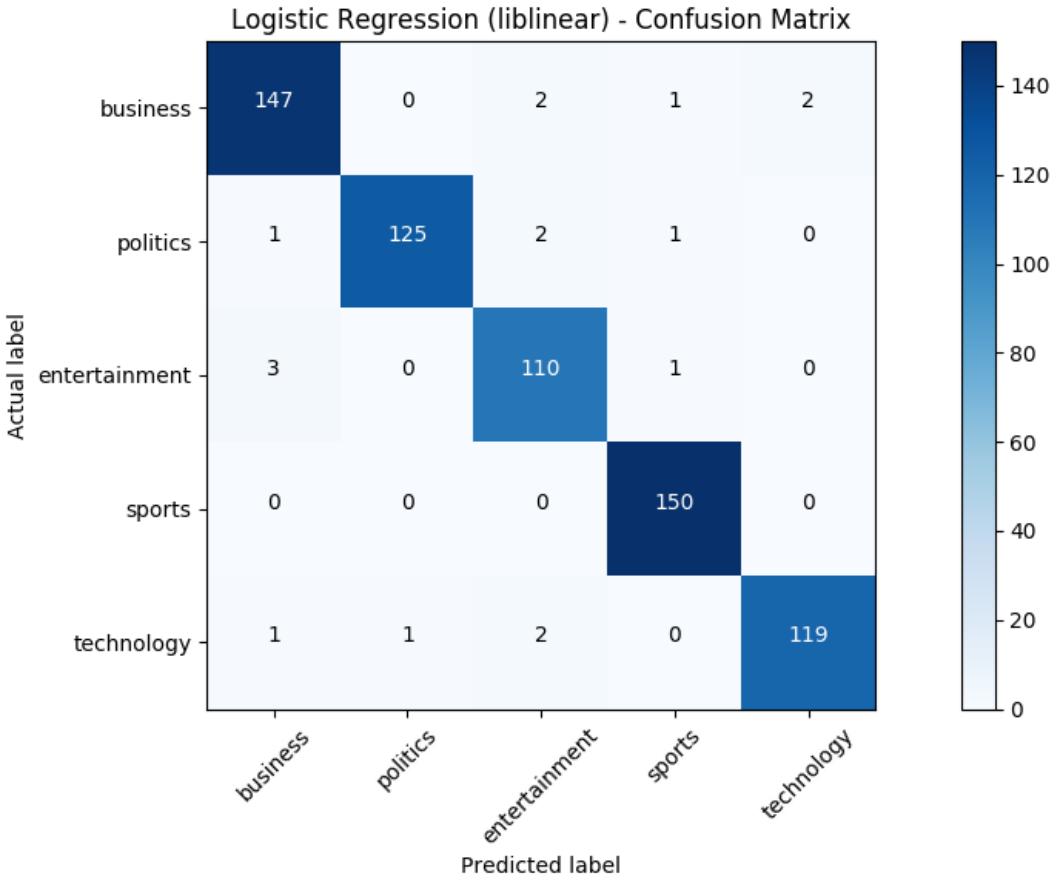



Figure 7.23: Key-phrase based Logistic Regression (liblinear) - Confusion Matrix

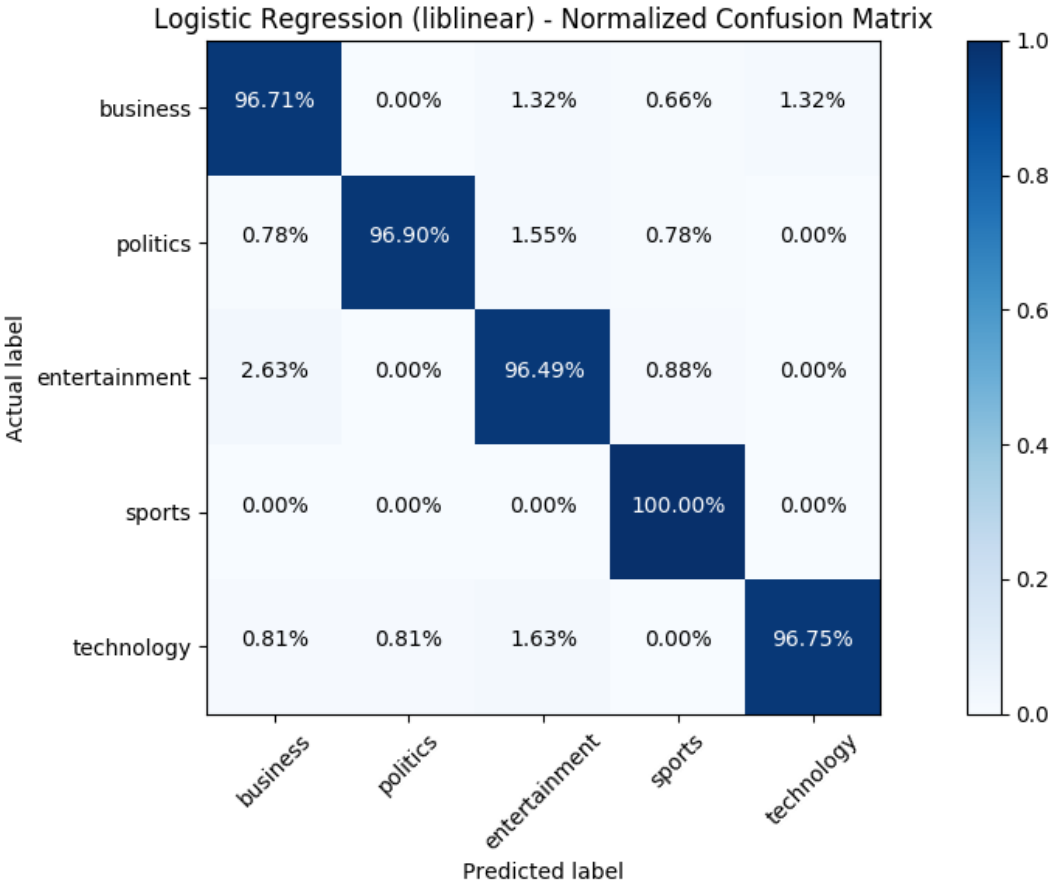



Figure 7.24: Key-phrase based Logistic Regression (liblinear) - Normalized Confusion Matrix

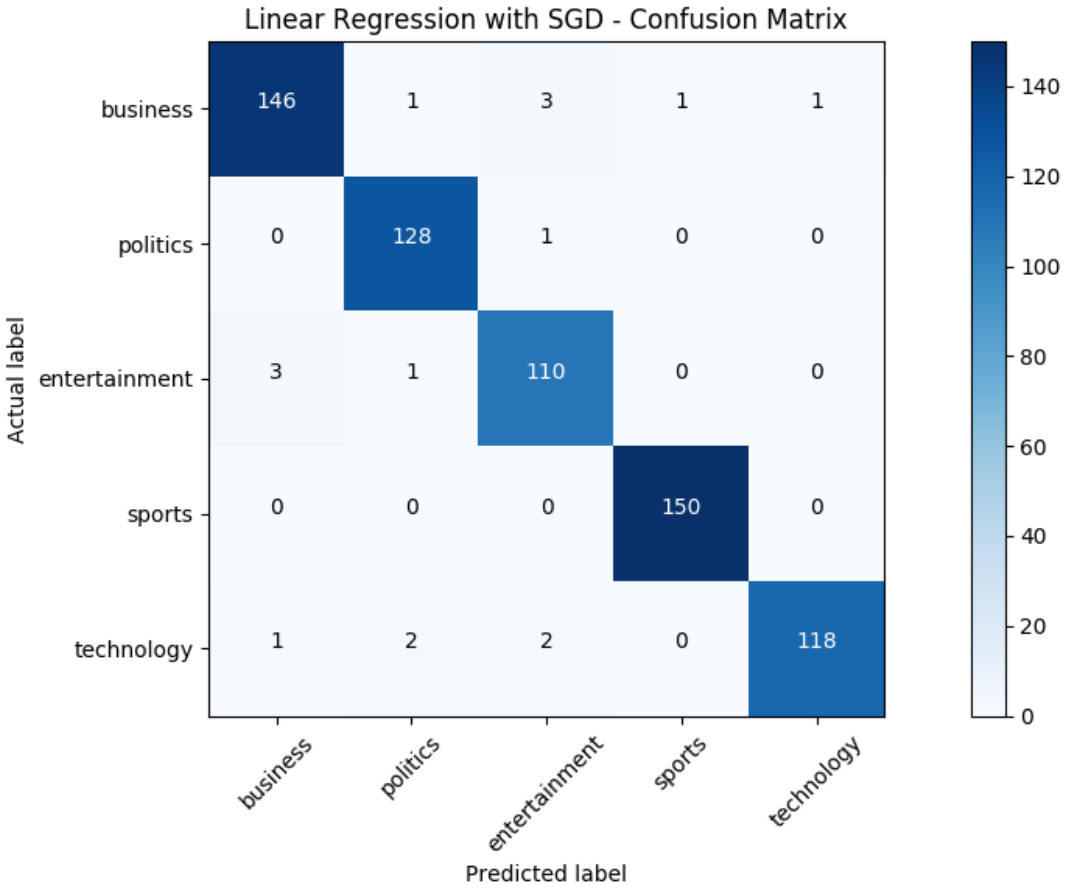



Figure 7.25: Key-phrase based Linear Regression with SGD - Confusion Matrix

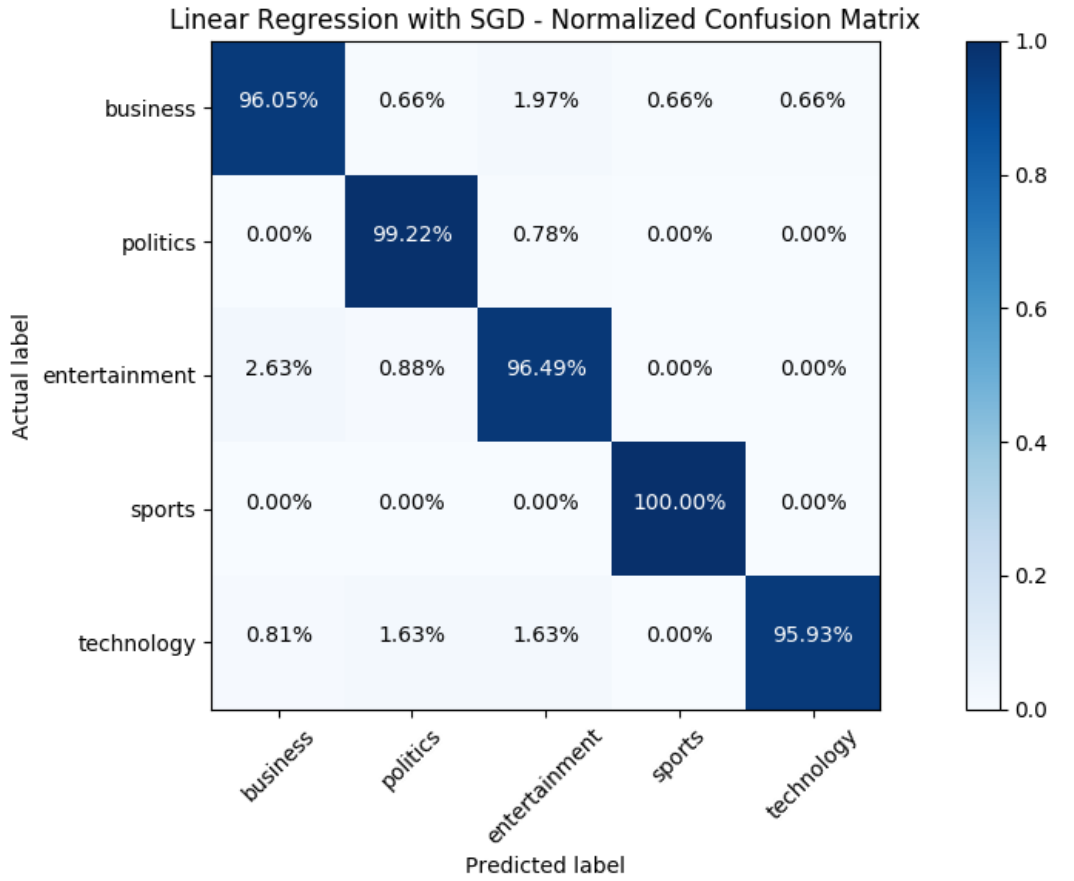



Figure 7.26: Key-phrase based Linear Regression with SGD - Normalized Confusion Matrix

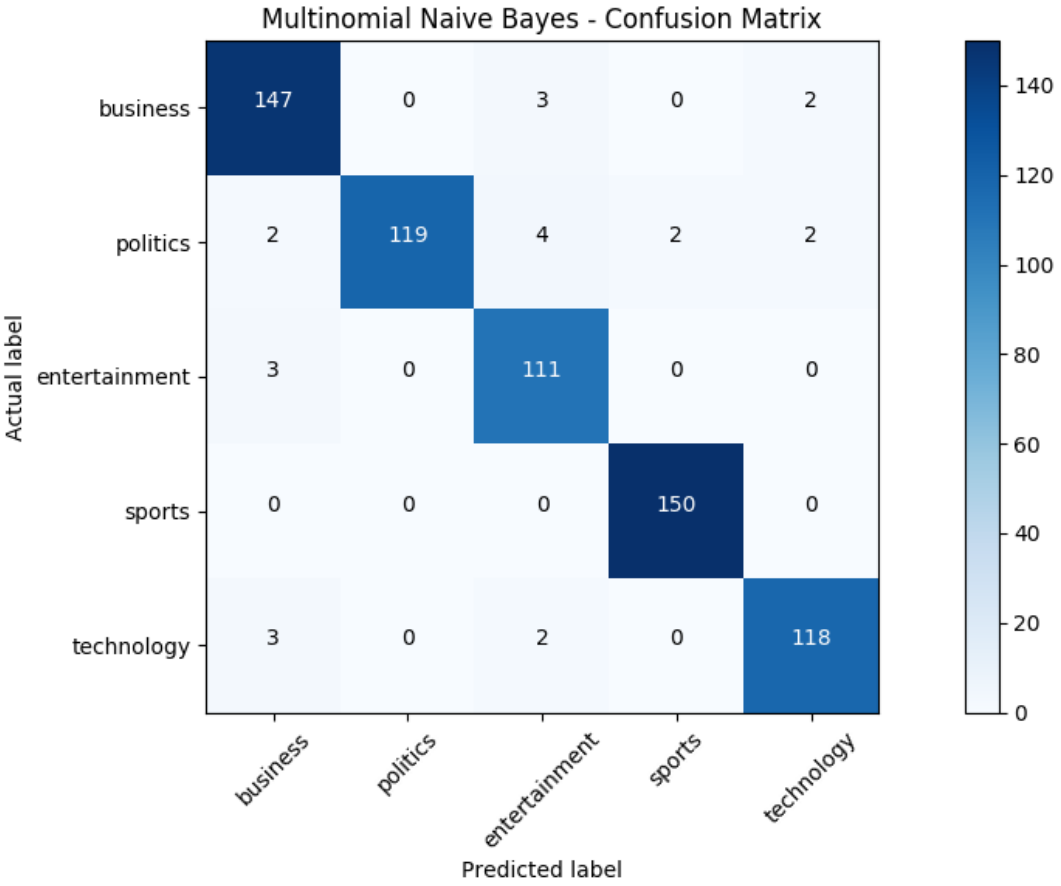



Figure 7.27: Key-phrase based Multinomial Naive Bayes classification - Confusion Matrix

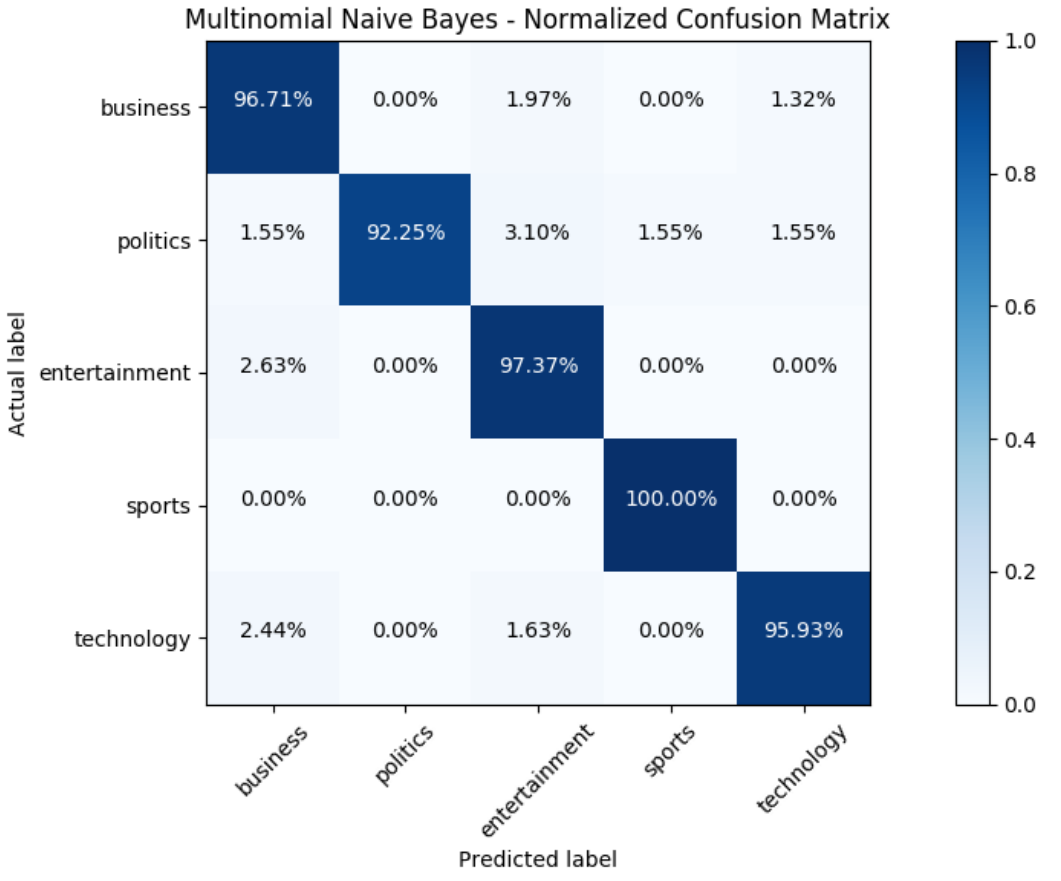



Figure 7.28: Key-phrase based Multinomial Naive Bayes classification - Normalized Confusion Matrix

# Chapter 8

# Results and Analysis

## 8.1. Entity based Classification - Consolidated Confusion Matrix

| **Classifiers** **Classes** | **k-NN** | **SVM** | **RF** | **LR** **(lbfgs)** | **LR** **(liblinear)** | **LR** **(SGD)** | **MultiNB** | **Mean Accuracy** **(Per Class)** |
|---|---|---|---|---|---|---|---|---|
| **Business** | 79.39 | 95.15 | 87.27 | 95.15 | 94.55 | 95.15 | 92.73 | 91.34 |
| **Politics** | 80.80 | 90.40 | 74.40 | 88.80 | 88.00 | 89.60 | 86.40 | 85.49 |
| **Entertainment** | 86.61 | 92.13 | 91.34 | 86.61 | 88.19 | 93.70 | 92.13 | 90.10 |
| **Sports** | 94.41 | 98.60 | 98.60 | 98.60 | 98.60 | 98.60 | 99.30 | 98.10 |
| **Technology** | 71.30 | 89.81 | 80.56 | 87.04 | 86.11 | 90.74 | 85.19 | 84.39 |
| **Mean Accuracy** **(Classifier)** | 82.502 | 93.218 | 86.434 | 91.24 | 91.09 | 93.558 | 91.15 | 89.89 |

Table 8.1: Entity based Classification  Consolidated Confusion Matrix (%)

## 8.2. Key-phrase based Classification - Consolidated Confusion Matrix

| **Classifiers** **Classes** | **k-NN** | **SVM** | **RF** | **LR (lbfgs)** | **LR (liblinear)** | **LR (SGD)** | **MultiNB** | **Mean Accuracy (Per Class)** |
|---|---|---|---|---|---|---|---|---|
| **Business** | 90.13 | 96.71 | 98.68 | 96.71 | 96.71 | 96.05 | 96.71 | 95.96 |
| **Politics** | 87.60 | 99.22 | 93.02 | 97.67 | 96.90 | 99.22 | 92.25 | 95.13 |
| **Entertainment** | 93.86 | 96.49 | 94.74 | 96.49 | 96.49 | 96.49 | 97.37 | 95.99 |
| **Sports** | 98.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 99.71 |
| **Technology** | 94.31 | 95.12 | 96.75 | 96.75 | 96.75 | 95.93 | 95.93 | 95.93 |
| **Mean Accuracy (Classifier)** | 92.78 | 97.508 | 96.638 | 97.524 | 97.37 | 97.538 | 96.452 | 96.544 |

Table 8.2: Key-phrase based Classification Consolidated Confusion Matrix (%)

## 8.3. Entity based Classification

| **No.** | **Entity based Classification** | **Accuracy score** | **Cohen Kappa score** |
|---|---|---|---|
| 1. | k-Nearest Neighbors | 0.8293 | 0.7852 |
| 2. | Support Vector Machine (SVC linear) | 0.9356 | 0.9190 |
| 3. | Random Forest | 0.8698 | 0.8362 |
| 4. | Logistic Regression (lbfgs) | 0.9177 | 0.8963 |
| 5. | Logistic Regression (liblinear) | 0.9162 | 0.8944 |
| 6. | Linear Regression (SGD) | 0.9386 | 0.9228 |
| 7. | Multinomial Naive Bayes | 0.9162 | 0.8944 |

Table 8.3: Entity based classification results using cross validation technique (70:30 ratio)

## 8.4. Key-phrases based Classification

| No. | Key-phrases based classification | Accuracy score | Cohen Kappa score |
|---|---|---|---|
| 1. | k-Nearest Neighbors | 0.9281 | 0.9100 |
| 2. | Support Vector Machine (SVC linear) | 0.9760 | 0.9700 |
| 3. | Random Forest | 0.9686 | 0.9605 |
| 4. | Logistic Regression (lbfgs) | 0.9760 | 0.9700 |
| 5. | Logistic Regression (liblinear) | 0.9746 | 0.9681 |
| 6. | Linear Regression with SGD | 0.9760 | 0.9700 |
| 7. | Multinomial Naive Bayes | 0.9656 | 0.9568 |

Table 8.4: Key-phrases based classification results using cross validation technique (70:30 ratio)

## 8.5. Classifiers Performance for Entity and Keyphrase based Classification

| No. | Classifiers | Avg. Accuracy | Avg. Cohen Kappa |
|---|---|---|---|
| 1. | k-Nearest Neighbors | 0.8787 | 0.8476 |
| 2. | Support Vector Machine (SVC linear) | 0.9558 | 0.9445 |
| 3. | Random Forest | 0.9192 | 0.8984 |
| 4. | Logistic Regression (lbfgs) | 0.9469 | 0.9332 |
| 5. | Logistic Regression (liblinear) | 0.9454 | 0.9314 |
| 6. | Linear Regression (SGD) | 0.9573 | 0.9454 |
| 7. | Multinomial Naive Bayes | 0.9409 | 0.9256 |

Table 8.5: Average scores - Entity and Key-phrases based classification results

## 8.6. Feature selection based on average scores

| No. | Features | Avg. Accuracy score | Avg. Cohen Kappa score |
|---|---|---|---|
| 1. | Entities | 0.9033 | 0.8783 |
| 2. | Key-phrases | 0.9664 | 0.9579 |

Table 8.6: Avg. scores - Accuracy and Cohens Kappa for Entities and Key-phrases features

## 8.7. Representation and interpretation

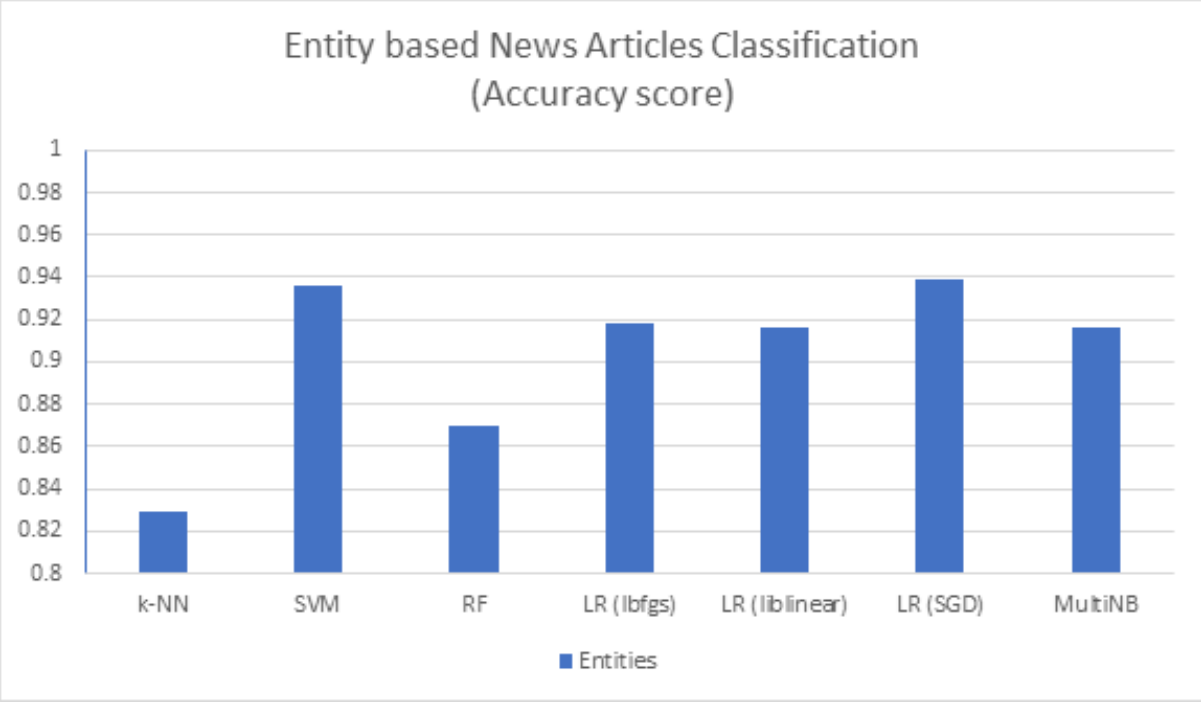



Figure 8.1: Entity based News Articles Classification (Accuracy score)

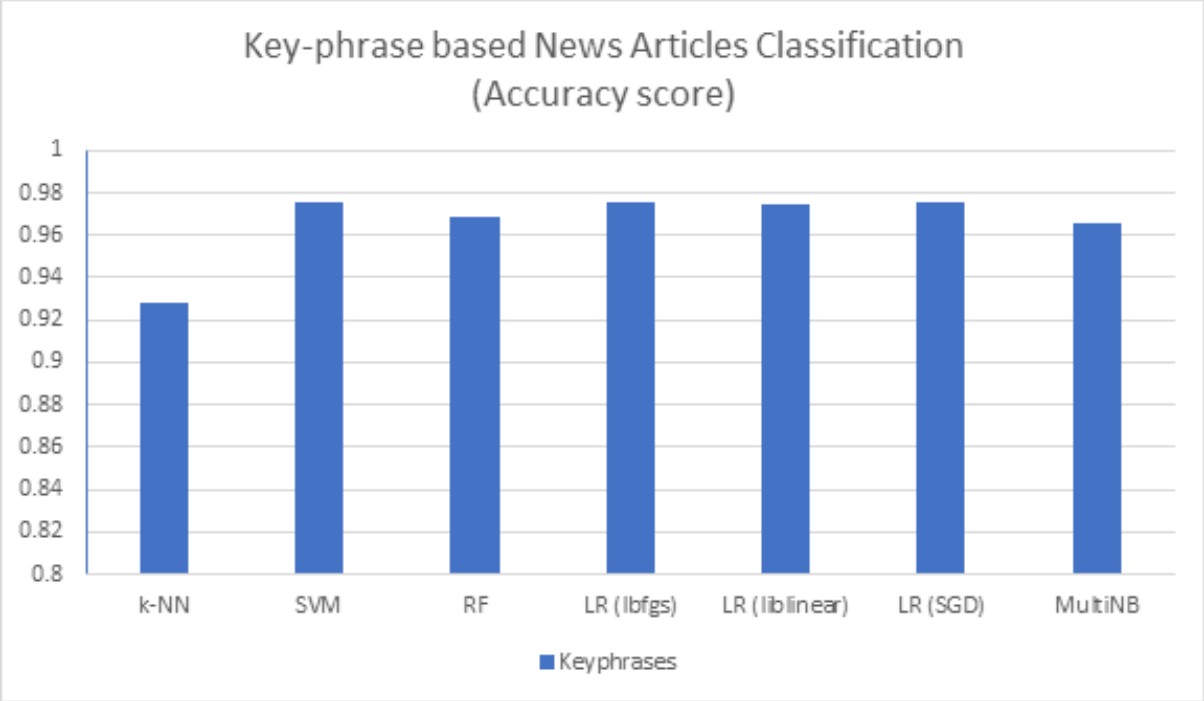



Figure 8.2: Key-phrase based News Articles Classification (Accuracy score)

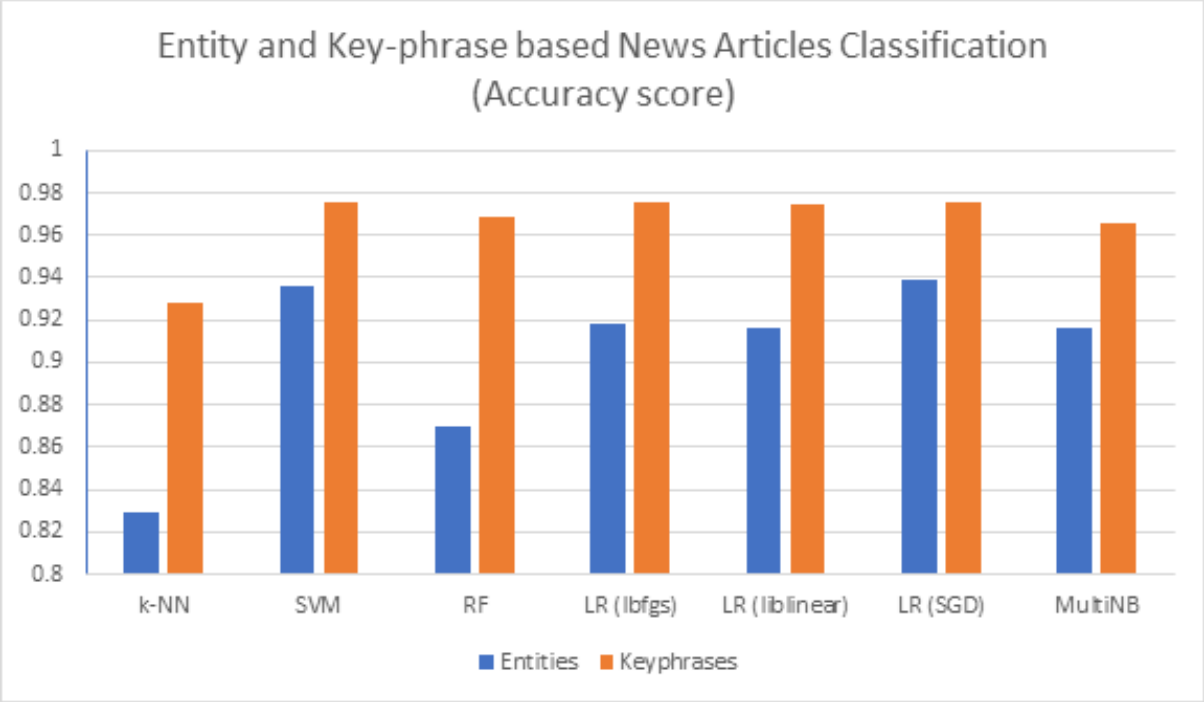



Figure 8.3: Entity and Key-phrase based News Articles Classification (Accuracy score)

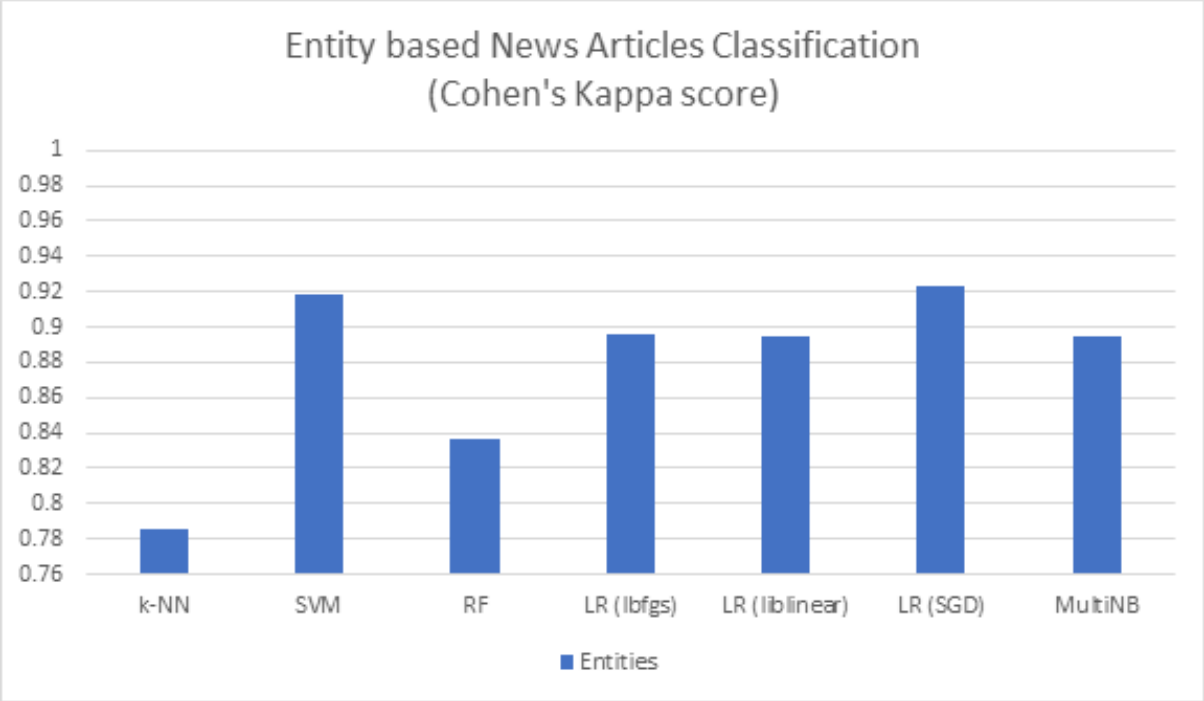



Figure 8.4: Entity based News Articles Classification (Cohens Kappa score)

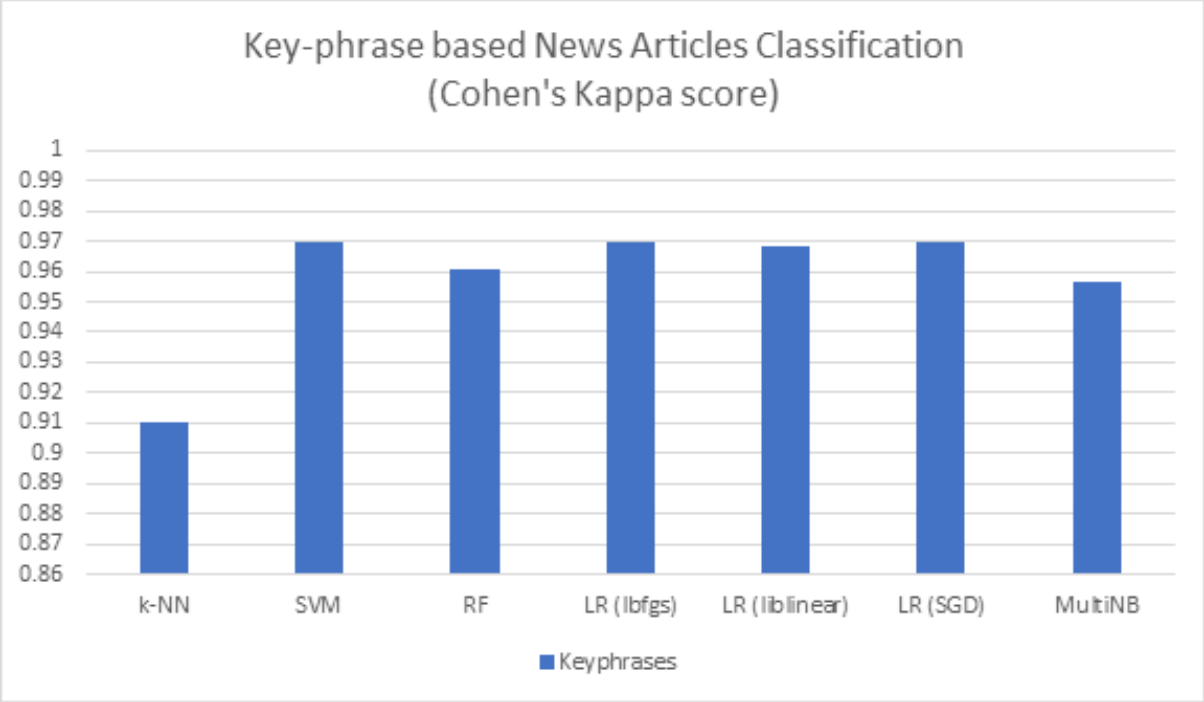



Figure 8.5: Key-phrase based News Articles Classification (Cohens Kappa score)

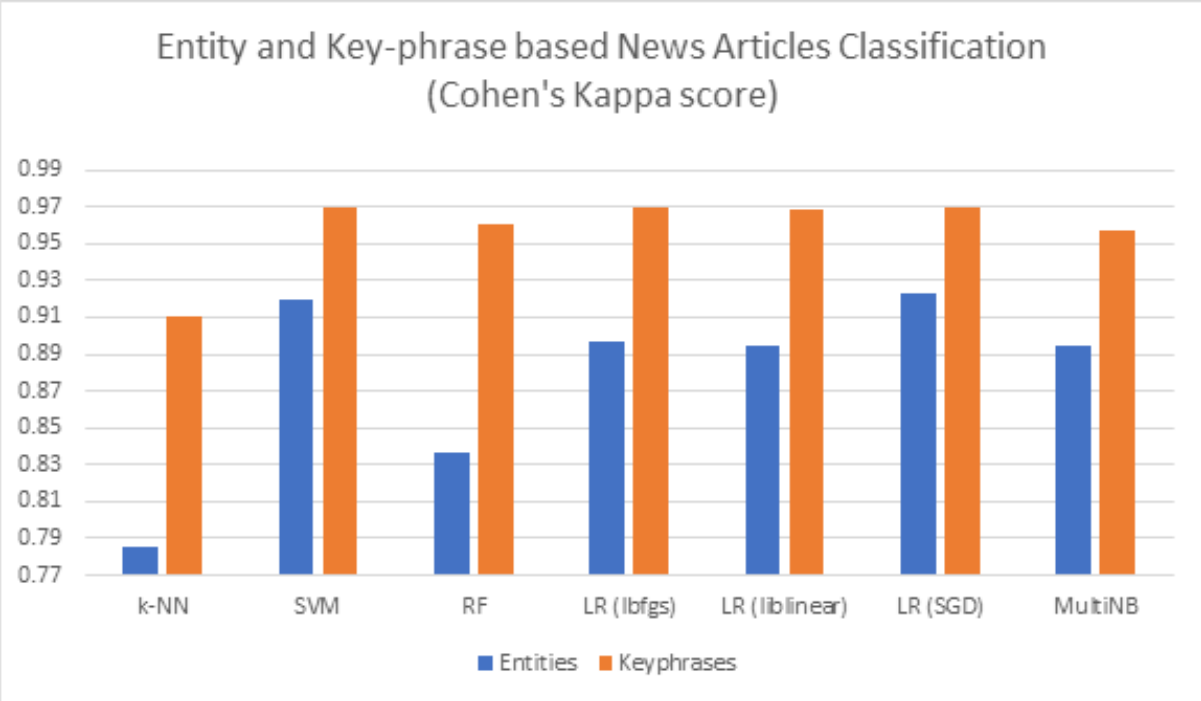



Figure 8.6: Entity and Key-phrase based News Articles Classification (Cohens Kappa score)

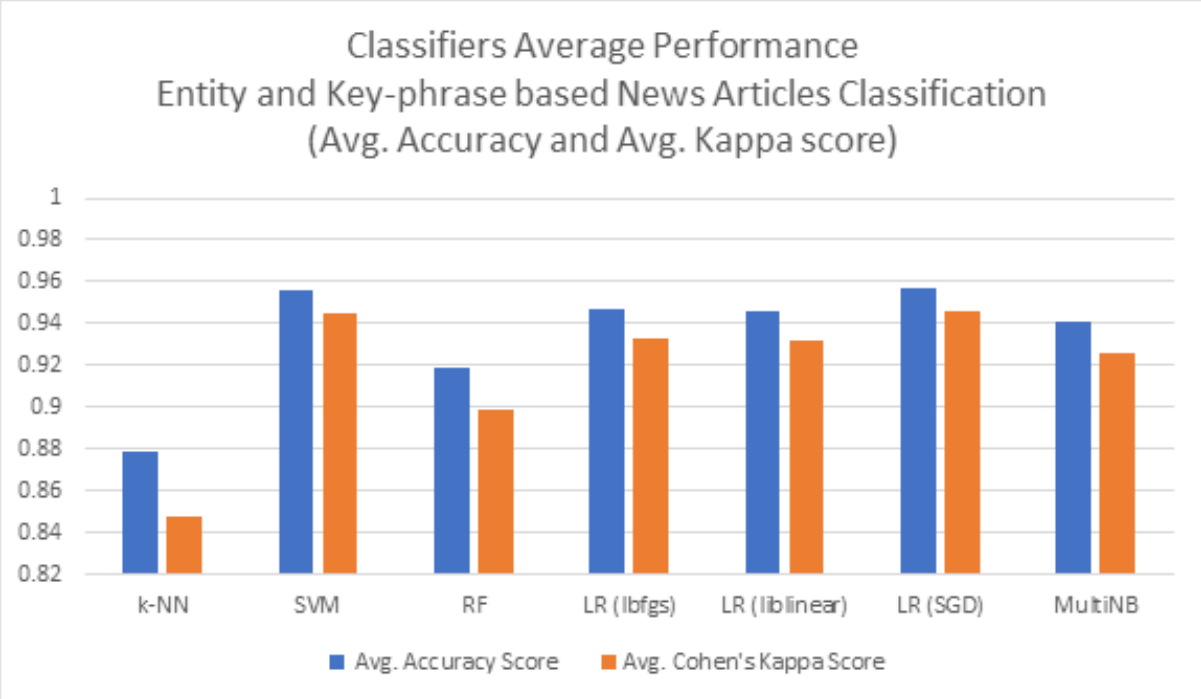



Figure 8.7: Classifiers Avg. Performance for Entity and Key-phrase based News Articles Classification - Avg. Accuracy and Avg. Cohens Kappa score

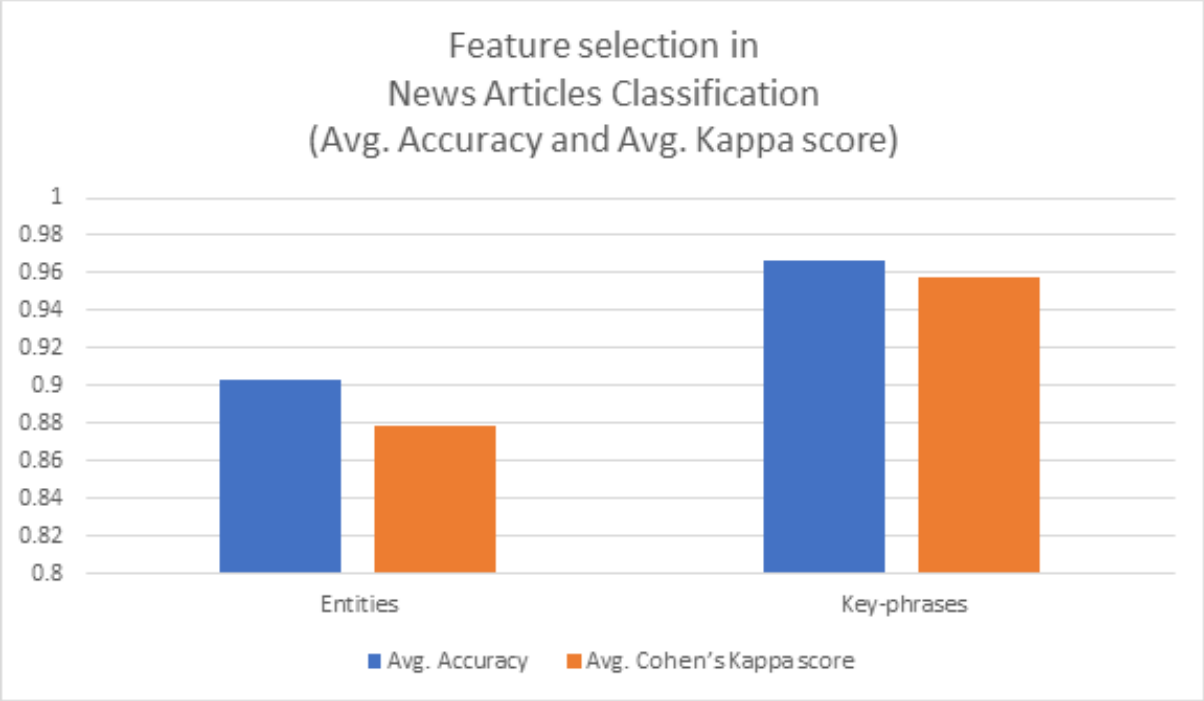



Figure 8.8: Feature selection in News Articles Classification

The instance based and normalized confusion matrix for entity-based classification are given in Figures [7.1 - 7.14]. Similarly, the confusion matrix for Key-phrase based classification are given in Figures [7.15 - 7.28]. Consolidated confusion matrix for both the classification methods are described in Table [8.1, 8.2]. Entity-based and Key-phrase based classification with accuracy score and Cohen Kappa's score are shown in Table [8.3, 8.4]. Average performance for both the classification approaches by individual classifiers given in Table 8.5. Average Classifier Performance is calculated as arithmetic mean of individual classifiers in both the approaches. Feature selection - Entity and Key-phrases comparison described in Table 8.6. We observed key-phrases has higher accuracy and avg. accuracy scores. The classification results for individual classifiers performance are represented in graph as shown in Figure [8.1, 8.2] and in Figure [8.4, 8.5]. Comparison graph for accuracy scores, Cohen Kappa's, avg. scores of individual classifiers are shown in Figure [8.3, 8.6, 8.7], respectively. Figure 8.8 represents feature selection with respect to avg. accuracy and Cohen Kappa score.

# Conclusion and Future scope

The research study conducted with the hypothesis that based on the important features, Entities and Key-Phrases, the news articles can be classified with improved accuracy. It focuses on the feature extraction from news articles and classification. The BBC News dataset is used in the study, which consists of 2225 news articles across five categories - business, entertainment, sports, politics, technology. Mainly Entities, Key-Phrases are targeted for the news articles classification. The news features are extracted with AWS Comprehend service. The TF-IDF used as feature weighting method and vectorizer. k-NN, SVM, Random Forest, Logistic Regression lbfgs, Logistic Regression liblinear, Linear Regression with SGD, Multinomial Naive Bayes classification models are trained with the extracted features and tested with cross validation technique in 70:30 ratio. Based on the results, the study evaluates the classification models and the significance of Entities, Key-phrases in news articles classification.

In this work, the confusion matrices and graphs given above depict detailed results on classifier performance for entity-based and key-phrase based news articles classification. The average accuracy is arithmetic mean, where as accuracy score and cohen's kappa score measures are classification metrics. Based on the results, SVM and Logistic Regression with SGD classifiers performed with higher accuracy in both entity-based and key-phrase based classification compared to obtained classifiers' results. As part of feature selection, we can infer that key-phrase based classification provides rich set of

features and hence improved accuracy in news or document analysis and classification.

In this study, key-phrase based classification provides improved classification results than entity-based classification. In entity-based news articles classification, accuracy score obtained Linear Regression (SGD) is 0.9386 and nearer to that accuracy score obtained with SVM is 0.9356. In key-phrase based news articles classification, exactly same accuracy score is obtained with SVM, Logistic Regression (lbfgs), Logistic Regression with SGD i.e. 0.9760. The average classifiers performance on entity-based and key-phrase-based classification with SVM and Logistic Regression with SGD achieved as 0.9558 and 0.9573, respectively,. In terms of feature selection, the average accuracy scores for both the cases entity and key-phrase based classification obtained results as 0.9033 and 0.9664, respectively. Hence, key-phrases based classification provides better results compared to entity-based classification. In this research, the SVM and Logistic Regression with SGD classifier performance in news articles classification have obtained higher accuracy scores.

With reference to research work in the literature survey, our approach have obtained improved accuracy and remarkable results. S. Foroozan, et. al. worked on improving sentiment classification accuracy of financial news using N-gram approach and feature weighting methods. [32] The linear SVM and RBF SVM classifiers are trained with unigram, bigram, combination of unigram-bigram with binary, TF, TF-IDF weighing methods. It implies document frequency (DF) can be used as a dimensional reduction method in order to reduce number of features while improving the classification accuracy. The N-gram approach results in 0.9703 classification accuracy. In our approach, we have achieved better classification accuracy with 0.9760 accuracy score than N-gram approach. The key-phrase based news articles classification approach have proven with better classification accuracy. The feature selection, identification and extraction are

| No. | Research work | Method | Accuracy |
|---|---|---|---|
| 1. | S. Foroozan, et. al. [32] | N-gram approach | 97.03% |
| 2. | Chenbin Li, et. al. [26] | Improved Bi-LSTM-CNN model | 96.45% |
| 3. | Xueying Zhang, et. al. [42] | SVM and ELM with kernels | 88.74% |
| 4. | Dilini Dandeniya, et. al. [8] | Ensemble Classifier | 97.98% |
| 5. | K. Ohtsuki, et. al. [13] | Topic-extraction model | 76.60% |
| 6. | Our research study | Key-phrase based classification approach | 97.60% |

Table 8.7: Classification accuracy comparison with research work

important aspects in document classification.

Chenbin Li, et. al. research focuses on text classification problem of NLP by using the Bi-LSTM-CNN method. [26] The Bi-LSTM-CNN model utilizes the loop structure to obtain the context information and constructs the left and right contexts of each word through the Convolutional Neural Network (CNN) to construct the textual expression of the word, which is more accurately expressed the semantics of the text. The data set used in this experiment is a subset of THUCNews for training and testing. It selects news from the ten categories of sports, finance, real estate, home, education, technology, fashion, politics, games and entertainment as experimental data. The 96.45% accuracy obtained with Bi-LSTM-CNN. In our study, we have achieved better accuracy with key-phrase based news articles classification with 97.60%.

Xueying Zhang, et. al. worked on Sentiment Analysis with Chinese language [42]. Due to the huge difference between English and Chinese in syntax, semantics and pragmatics etc., there are problems in the processing of Chinese text. The study uses SVM and ELM with Kernel classifiers with TF-IDF weighing method and have used hotel BBS comments dataset. Extreme Learning Machine was first proposed by Huang in 2006, ELM with kernels is a single-layer feedforward network, and it has more effectively to regression prediction compared with the basic ELM algorithm. The study started with

the assumption, that as for the support vector machine algorithm, ELM with kernels can get better or similar predictive accuracy with less time. Compared with SVM, although the accuracy for ELM with kernels is similar to SVM, but the training and testing time in ELM with kernels will be far less. The experiment results show that ELM with kernels method of emotional polarity analysis of Chinese text is more effective. The accuracy score results with SVM and ELM with kernels are 88.54% and 88.74% respectively. As Chinese language sentiment analysis is challenging because of lack in research of Chinese language processing, this work has shown better results. Compared to our study, key-phrase based news articles classification has performed with far better accuracy i.e. 97.60%.

Dilini Dandeniya, et. al. worked on An Automatic e-news Article Content Extraction and Classification with Ensemble classifier methodology. [8] The Multinomial Naive Bayes, Support Vector Machine, Random Forest classifiers are used with TF-IDF weighing method. The study proposed ensemble classifier by combining three supervised learning algorithms. The experiments are conducted on various classifiers with BBC News dataset. The ensemble classifier approach generates weighted average probability for each class and aggregate probabilities from different classes and get the maximum value class as predicted class for e-news item. Without Ensemble classifier approach, the study has obtained classification accuracy using SVC with linear kernel is 0.973033. With Ensemble classifier (Hard Voting), the accuracy is 0.975380, whereas with Ensemble classifier (Soft Voting), the accuracy obtained is 0.979775. This research implies the ensemble classifier algorithm shows higher accuracy rather than individual classifier. In our research, key-phrase based content extraction and news articles classification approach have proved higher accuracy i.e. 0.9760 as compared to the results in SVC with linear kernel without ensemble classifier accuracy i.e. 0.973033. Also as compared to Ensemble classifier (Hard

Voting) results accuracy i.e. 0.975380, our study has still obtained higher accuracy score i.e. 0.9760. Ensemble classifier (Soft Voting) results in better accuracy. In our study, hard categorization was only considered still the accuracy obtained at par as compared to Ensemble classifier (Soft Voting).

In prior research study, K. Ohtsuki, et. al. conducted on topic extraction with multiple topic-words in broadcast-news speech. [13] The topic-words are extracted from news articles on the basis of relevance scores between topic-words and articles. It proposes topic extraction methodology. The dataset in study trained the topic-extraction model with newspaper articles and headlines extending back about five years and speech data with combination of speakers is used as evaluation data. The speech data used has background noise, human sound problems. The experiment is carried on transcribed news speech and speech recognized news speech, then word-error rates are calculated in experiment. The study shows that the topic extraction model achieved better performance which yielded precision to 74.5% for speech recognized news speech. The study employed N-best approach in order to compensate performance degradation caused by speech recognition errors, and achieved improved precision 76.6% with 10-best hypothesis. In comparison with this research, we have achieved much better results with key-phrase based news articles classification approach and proved to be significant improvement in news articles classification with accuracy 97.60%.

Hence, based on Key-phrases feature selection from news article, we observe better classification results and resulted in higher accuracy than literature work. Thus, we have achieved significant improvement with Key-phrase based classification approach for news articles classification.

## Future Scope

1. The research study can be extended to explore feature extraction for document analysis using different NLP cloud services like Google Cloud Natural Language, Microsoft Azure Text Analytics, IBM Watson Natural Language Understanding.

2. The research study can further be evaluated with Open Source NLP platforms like Apache NLTK, Apache OpenNLP, SpaCy, TextBlob, CogCompNLP, Stanford CoreNLP using similar benchmark datasets.

3. The study can further be targeted for different document types which contains structured or unstructured data for analysis with Apache TIKA framework. Apache TIKA is open source content detection and analysis framework which helps in bulk document analysis across different content types.

4. The feature extraction techniques can be experimented with different approach to extract the features like events, controversy and the inter-relation of two news domain articles.

5. The further research can be conducted for extraction and analysis of healthcare or clinical records with Apache cTAKES framework. The clinical text or record analysis over digital healthcare platform records can be conducted with various disease or medical lab use cases.

6. The research study can be scoped up for sentiment analysis with product or event analysis with Readers perspective and Writers perspective. Furthermore, the research study can be exercised with multi-dimensional and multi-perspective document analysis.

# Bibliography

[1] Masayu Leylia Khodra - Event Extraction on Indonesian News Article Using Multi-class Categorization (2015) - 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA) 2015 IEEE

[2] Terry Traylor, Jeremy Straub, Gurmeet, Nicholas Snell - Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator (2019) - IEEE 13th International Conference on Semantic Computing (ICSC)

[3] Taufik Fuadi Abidin, Rahmad Dimyathi and Ridha Ferdhiana - Rule-Based and Machine Learning Approach for Event Sentence Extraction in Indonesian Online News Articles (2014) - International Conference on Information Technology Systems and Innovation (ICITSI)

[4] Danang Tri Massandy, Masayu Leylia Khodra - Guided Summarization for Indonesian News Articles (2014) - International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)

[5] Mr. Nilesh M. Shelke, Dr. Shrinivas Deshpande, Dr. Vilas Thakare - Statistical Feature based Approach for Aspect Oriented Sentiment Analysis (2017) - International Conference on Inventive Communication and Computational Technologies (ICICCT)

[6] Raihannur Reztaputra, Masayu Leylia Khodra - Sentence Structure-based Summarization for Indonesian News Articles (2017) - International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)

[7] Ismini Lourentzou, Graham Dyer, Abhishek Sharma and ChengXiang Zhai - Hotspots of News Articles: Joint Mining of News Text  Social Media to Discover Controversial Points in News (2015) - IEEE International Conference on Big Data (Big Data)

[8] Dilini Dandeniya - An Automatic e-news Article Content Extraction and Classification (2018) 18th International Conference on Advances in ICT for Emerging Regions (ICTer)

[9] Pal-Christian S. Njlstad , Lars S. Hyster, Wei Wei and Jon Atle Gulla - Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News (2014) - IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

[10] Adhy Rizaldy1 , Heru Agus Santoso2 - Performance Improvement Of Support Vector Machine (SVM) With Information Gain On Categorization Of Indonesian News Documents (2017) - International Seminar on Application for Technology of Information and Communication (iSemantic)

[11] Qiang Pan, Xin Xin, Junshuai Liu, Ping Guo - Extracting Company-Specific Keyphrases from News Media (2017) - 13th International Conference on Computational Intelligence and Security

[12] Mihail Minev - Quantification of Financial News for Economic Surveys (2013) - IEEE 13th International Conference on Data Mining Workshops

[13] K. Ohtsuki, T. Matsutoka2, S. Matsunagal, S. Furui - Topic extraction with multiple topic-words in broadcast-news speech (1998) - Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)

[14] Benjamin D. Horne, William Dron, Sibel Adal - Models for Predicting Community-Specific Interest in News Articles - (2018) - IEEE Military Communications Conference (MILCOM)

[15] Sneha Pasarate, Rajashree Shedge - Concept based document clustering using K prototype algorithm (2018) - International Conference on Control, Power, Communication and Computing Technologies (ICCPCCT)

[16] Mahsa Afsharizadeh , Hossein Ebrahimpour-Komleh, Ayoub Bagheri - Query-oriented Text Summarization using Sentence Extraction Technique (2018) - 4th International Conference on Web Research (ICWR)

[17] Hairon Sato, Nana Ogasawara, Yuanyuan Wang - Predicting Short-Term Exchange Rates for Automatic Purchasing using News Article Data (2018) - IEEE 7th Global Conference on Consumer Electronics (GCCE)

[18] Sungjick Lee, Han-joon Kim - News Keyword Extraction for Topic Tracking - Fourth International Conference on Networked Computing and Advanced Information Management

[19] Taishi Saito, Osamu Uchida - Automatic Labeling for News Article Classification Based on Paragraph Vector (2017) - 9th International Conference on Information Technology and Electrical Engineering (ICITEE), Phuket, Thailand

[20] Naoya OKUMURA Takao MIURA - Generating Headline Candidates for News Articles - (2016) - IEEE 17th International Conference on Information Reuse and Integration

[21] Maryam Bahojb Imani, Swarup Chandra, Samuel Ma, Latifur Khan, Bhavani Thuraisingham - Focus Location Extraction from Political News Reports with Bias Correction (2017) - IEEE International Conference on Big Data (BIGDATA)

[22] Gisel Bastidas Guacho, Sara Abdali, Neil Shah, Evangelos E. Papalexakis - Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings (2018) - IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

[23] Yu Shuqi,Wu Bin - Exploiting Structured News Information to Improve Event Detection via Dual-level Clustering (2018) - IEEE Third International Conference on Data Science in Cyberspace

[24] Souneil Park, Jungil Kim, Kyung Soon Lee, and Junehwa Song - Disputant Relation-Based Classification for Contrasting Opposing Views of Contentious News Issues (2013) - IEEE Transactions on Knowledge and Data Engineering ( Volume: 25 , Issue: 12 , Dec. 2013 )

[25] Zhenzhong Li, Wenqian Shang, Menghan Yan - News text classification model based on topic model (2016) - IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)

[26] Chenbin Li, Guohua Zhan, Zhihua Li - News Text Classification Based on Improved Bi-LSTM-CNN (2018) - 9th International Conference on Information Technology in Medicine and Education

[27] Priya P. Raut, Nitin N. Patil - Classification of controversial news article based on disputant relation by SVM classifier (2015) - 23nd Signal Processing and Communications Applications Conference (SIU)

[28] Syafruddin Syarifm, Anwar, Dewiani - Trending Topic Prediction by Optimizing K-Nearest Neighbor Algorithm (2017) - 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)

[29] Ning Gu, Duo-yong Sun, Bo Li, Ze Li - Sentiment Analysis for Topics based on Interaction Chain Model (2015) - European Intelligence and Security Informatics Conference

[30] Shilpa P, Madhu Kumar S D - Feature Oriented Sentiment Analysis in Social Networking Sites to Track Malicious Campaigners (2015) - IEEE Recent Advances in Intelligent Computational Systems (RAICS)

[31] Chiyu Cai, Linjing Li, Daniel Zeng - New Words Enlightened Sentiment Analysis in Social Media -(2016) - IEEE International Journal

[32] S. Foroozan, M.A. Azmi Murad, and N.M. Sharef, A.R. Abdul Latiff - Improving Sentiment Classification Accuracy of Financial News using N-gram Approach and Feature Weighting Methods (2015) - 2nd International Conference on Information Science and Security (ICISS)

[33] G. Preethi and P. Venkata Krishna, Mohammad S. Obaidat, V. Saritha, Sumanth Yenduri - Application of Deep Learning to Sentiment Analysis for Recommender System on Cloud (2017) - IEEE Journal

[34] Eissa M.Alshari, Azreen Azman, Shyamala Doraisamy - Improvement of Sentiment Analysis based on Clustering of Word2Vec Features (2017 28th International Workshop on Database and Expert Systems Applications

[35] Shahnawaz, Parmanand Astya - Sentiment Analysis: Approaches and Open Issues (2017) - International Conference on Computing, Communication and Automation (ICCCA)

[36] Vishal S. Shirsat, Rajkumar S. Jagdale, S. N. Deshmukh - Document Level Sentiment Analysis from News Articles (2017) - IEEE Journal

[37] Eissa M.Alshari, Azreen Azman, Shyamala Doraisamy, Norwati Mustapha and Mostafa Alkeshr - Effective Method for Sentiment Lexical Dictionary Enrichment based on Word2Vec for Sentiment Analysis (2018) - Fourth International Conference on Information Retrieval and Knowledge Management

[38] Saurabh Dorle, Dr.Nitin Pise - Political Sentiment Analysis through Social Media (2018) - Proceedings of the Second International Conference on Computing Methodologies and Communication (ICCMC 2018)

[39] Mohammad Kamel, Neda Keyvani, Hadi Sadoghi-Yazi - Sentimental Content Analysis and Knowledge Extraction from News Articles (2018) - IEEE Journal

[40] Lu Ye, Rui-Feng Xu, Jun Xu - Emotion Prediction of News Articles From Reader's Perspective Based on Multi-label Classification (2012) - Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian

[41] Nurulhuda Zainuddin, Ali Selamat - Sentiment Analysis Using Support Vector Machine (2014) - IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014)

[42] Xueying Zhang, Xianghan Zheng - Comparison of Text Sentiment Analysis based on Machine Learning (2016) - 15th International Symposium on Parallel and Distributed Computing

[43] Mazhar Iqbal Rana x, Shehzad Khalid y, Muhammad Usman Akbar z - News Classification Based On Their Headlines: A Review (2014) - 17th IEEE International Multi Topic Conference 2014

[44] Pingping Lin, Rong Xiao and Yan Zhang - News Event Summarization Complemented by Micropoints (2015) - 31st IEEE International Conference on Data Engineering Workshops

[45] Leo Breiman - Random Forests (2001) - Machine Learning, 45, 532, 2001, Kluwer Academic Publishers

[46] Mary L. McHugh, Interrater reliability: the kappa statistic (2012) - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

[47] https://www.oreilly.com/ideas/evaluating-machine-learning-models/page/3/evaluation-metrics

[48] https://medium.com/@jonathan_hui/build-a-deep-learning-dataset-part-2-a6837ffa2d9e

[49] http://mlg.ucd.ie/files/datasets/bbc-fulltext.zip - BBC News Dataset

[50] https://docs.aws.amazon.com/comprehend/latest/dg/how-entities.html

[51] https://docs.aws.amazon.com/comprehend/latest/dg/how-key-phrases.html

[52] https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/

[53] https://www.geeksforgeeks.org/clustering-in-machine-learning/

[54] https://en.wikipedia.org/wiki/Tf-idf

[55] https://www.quora.com/What-is-the-difference-between-regression-classification-and-clustering-in-machine-learning

[56] https://www.datascience.com/blog/k-means-clustering

[57] https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd

[58] LDA - https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[59] https://www.geeksforgeeks.org/ml-linear-regression/

[60] https://medium.com/simple-ai/logistic-regression-intro-to-machine-learning-7-ba18ab305b24/

[61] https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/

[62] https://en.wikipedia.org/wiki/Limited-memory_BFGS

[63] https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

[64] https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

[65] https://towardsdatascience.com/why-linear-regression-is-not-suitable-for-binary-classification-c64457be8e28