https://doi.org/10.52131/jcsit.2021.0201.0010

*i*RASD Journal of Computer Science and Information Technology

Volume 2, Number 1, 2021, Pages 52 - 62

Journal Homepage:
https://journals.internationalrasd.org/index.php/jcsit

News Location Classification

Malik Shahzad Ahmad[1], Muhammad Azhar Bhatti[2]

[1] Principal Software Engineer, Addo, Lahore, Pakistan. Email: malikshahzad@addo.ai
[2] Visiting Lecturer, Department of Economics, The Islamia University of Bahawalpur, Pakistan. Email: Azhar.bhatti219@gmail.com

ARTICLE INFO

Article History:
Received: September 05, 2021
Revised: October 29, 2021
Accepted: December 27, 2021
Available Online: December 31, 2021

Keywords:
News Paper
Information
Digitalization
Hidden Feature
Convolutional Neural Network (CNN)
F11 score

ABSTRACT

Every day there are a lot of things that happen around the world. There are various ways to record every event that is occurring around the world, such as news, blogs, and articles. Over the past few years, there are multiple news available on every event that has occurred. It adds to the size of information that is available for human beings to consume. People are, moving from paper-based newspapers to digital newspapers to get their daily feed of news and digitization has a role to play in this behaviour. These days every person is preoccupied with a lot of work, online and offline, as mentioned earlier the amount of information is being increased with every passing day. For this reason, people are only interested in news that match their interests. A large amount of data in the form of text is available online, hence its classification based on its hidden features can lead to the better recommendation of news to individuals. In this research work, we have used focus area and temporal features to classify news using a Convolutional Neural Network (CNN). The results of the proposed methodology in the form of precision, accuracy, recall, and F1-Score show that these features indeed can be used for recommender systems.

Corresponding Author's Email: malikshahzad@addo.ai

1. Introduction

There is a variety of data that makes up a news, including local and international areas. One factor of news authenticity is that all the topics from particular area are being discussed in it. With every passing moment there is an event happening around the world which is being documented in the form of articles, news and blogs. Since there are multiple sources to report a single event the amount of information available for even a single event is increasing with every passing day, which leads to individuals only being interested in the news that they like. News types, news related to specific area can be the factor of interest of an individual (Mitchell, Gottfried, Barthel, & Shearer, 2016; Zaila & Montesi, 2015). The news is mostly available in the form of textual data, which makes it difficult to retrieve information that to which interest a specific news document belongs to, classification of available textual data with respect to their location and their temporal features can lead to better classification into sub types of news and news can be recommended in a better way to readers.

There is different method being used by Information Retrieval (IR) system to extract information required by user from textual data within acceptable amount of time. IR systems are used to understand user query and provide as much as possible results quickly. For better news classification more aspects are needed to be added for news retrieval. There is more geographical information available in news documents than any other document. It can be a good feature to classify documents because it can be an effective way to present news with respect to location (Zaila & Montesi, 2015). News documents have another feature known as temporal information associated to it. There can be different type of temporal information

available with each document, such as when a particular event that is being reported in that news occurred and when that news was reported. Mainly there is two types of information available in textual data, one is explicit and other one is implicit. Explicit temporal information states a time, date, month or year explicitly on the other hand implicit information is usually not that clearly presented in text and there is always a reference to that information, such as "three days from now" has a reference date which is today. Temporal information can be important while classifying news documents for users (Dilrukshi, De Zoysa, & Caldera, 2013).

The inverted pyramid paradigm is followed by news documents, figure 1. The most important information is presented by headlines of the news which always at the top and then sections with less information tend to be at the bottom of news document. It is a style of writing news document that is most used (Keith, Horning, & Mitra, 2020). It is represented as inverted triangle and the top of the triangle represents the agenda of the news following the important information about the event. The base region of the news pyramid addresses general data or subtleties of the news. To arrange the news with the part of the area, an upset pyramid worldview structure is useful.

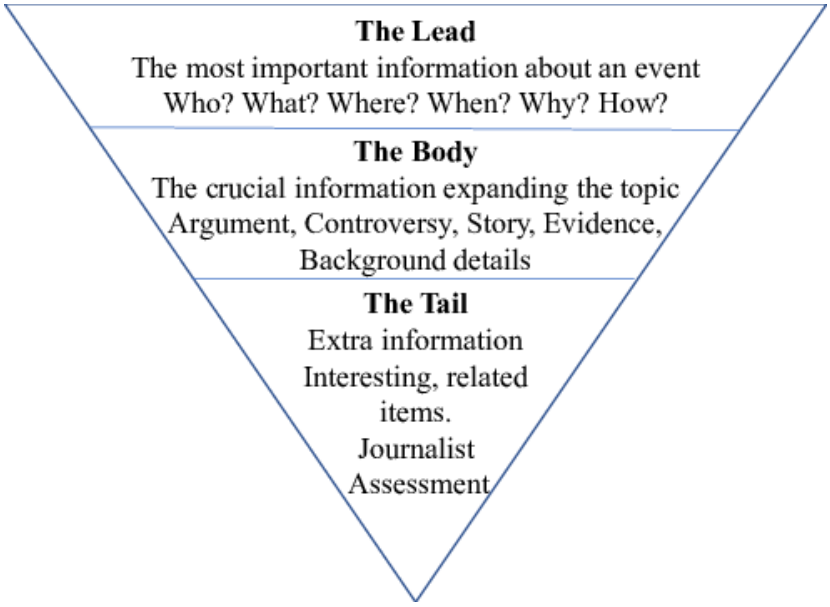

Figure 1: Inverted Pyramid Paradigm

Named Entity Recognition (NER) groups the data got from the text of a report. It is utilized to label elements with their comparing classes. For instance, when area data "Lal Masjid" is separated, it very well may be determined that the news is about the assault on Lal Masjid. It is the main substance recognizable proof/extraction procedure utilized in IR as it extricates designated data from a lot of information (Yadav & Bethard, 2019). In this examination, NER was utilized to extricate topographical data from text-based information for the grouping of information records in light of the area referenced in them. The goal was to extricate significant data that is essential for the order of information.

In this research work we used NER technique to identify locations in news documents. The accompanying advances depict the system of NER:

- As a matter of first importance, NER peruses the report.
- Then, at that point, substances (area, name, individual association, time) are separated.
- Finally, it groups the substances in the predefined classifications.

You can likewise characterize your classifications according to need and prerequisite and you can utilize this arrangement to name information and train the informational index. Classes in NER are shortened as LOC (area), (individual), Organization (Association), (Date), GPE (International Substance) NORP (Identities or Strict or Political Gathering), and so on.

This framework has two significant techniques for element extraction:

- In the principal strategy, just those elements are characterized that have a place with some classification from the predefined list.

- In the subsequent technique, the extractor recognizes the elements in view of typical statements, i.e., when "@" comes in the middle of a string it distinguishes it as an email address.

Information is preprocessed prior to accepting it as a contribution to NER. Unique markup is converted into literary substance and information is clarified. Explanation has metadata, for example, date, creator name, data of work, and so forth in the purposed framework, the NER approach was utilized to extricate topographical data from the dataset of information reports. Area data assisted with ordering the news in a superior and more viable manner.

2. Literature Review

Arrangement of information is essentially the characterization of message considering various situations. In past examinations, different grouping calculations are utilized to sort the text. Classifiers like calculated relapse, Support Vector Machine (SVM), k Nearest Neighbors (KNN), and Credulous Bayes are utilized for classification. Li, Shang, and Yan (2016) utilized a Bidirectional Long Short Term Memory Convolutional Neural Organization (Bi-LSTM-CNN) model to group news text into ten classes. The order was done based on setting data utilizing semantics. In another review Dilrukshi et al. (2013)**,** short messages from Twitter were ordered into twelve gatherings by utilizing SVM. The motivation behind grouping them was to recognize the most well-known news class at a given time in a given spot. A component vector made utilizing the sack of-words approach was given as contribution to the SVM. This highlight vector contained expressions of each short message as elements. J. Ahmed and Ahmed (2021) utilized a profound learning CNN to consequently characterize news archives into classes. These classes included 'Tech', 'Media', 'Business', 'Sports', 'Legislative issues', 'World News', 'Amusement', and 'Wrongdoing'. The reports were first named utilizing managed text marking, and afterward a vector of records and names was taken care of as contribution to the CNN for preparing.

Li et al. (2016) additionally classify news reports in view of classes of information. The model that they utilized depended on Latent Dirichlet Allotment (LDA). The elements of the text were diminished and includes were removed utilizing the point model. Softmax relapse calculation was utilized to arrange news into various classifications including legislative issues, sports, diversion, and so forth Ordinarily, a strategic relapse calculation is utilized for grouping yet it gives the outcome in twofold structure. As different classes of information were taken, so softmax relapse calculation was utilized all things being equal. In another review Krishnalal, Rengarajan, and Srinivasagan (2010), Stowed away Markov Model and SVM were utilized to characterize news reports in light of classifications of information. News archives were gotten utilizing JFileChooser Programming interface and parsed utilizing Java StringTokenizer. They were then preprocessed and named by their classifications (legislative issues, money, and sports). Computerized highlights were removed utilizing Gee lastly, characterization was finished utilizing SVM.

In a concentrate by Lai, Xu, Liu, and Zhao (2015), order of information records was done alongside feeling investigation of client audits. For the order of information records as per classifications of information, the highlights were separated utilizing character-level CNN, and grouping was finished utilizing LSTM. The classifications included sci/tech, business, sports, innovation, auto, amusement, money, and world.

Time gives a structure to analyze the coherent movement of activities and occasions. Kim, Pethe, and Skiena (2020) zeroed in on this viewpoint and separated implied and unequivocal worldly articulations from text and afterward grouped the text as per the specific time it compared to. For this reason, they utilized Bidirectional Encoder Representation from Transformers (BERT). Order of information text as per transient articulations assists the client with understanding the class of information in a superior manner. A concentrate by Strötgen and Gertz (2010) proposes a standard based framework that removes worldly articulations from text and standardizes the record. These fleeting articulations, phonetic pieces of information, and information assets are predominantly removed from message utilizing normal articulation designs. Extraction of transient articulations was incorporated into an Unstructured Data the Executives Design as a part. Blend models in light of language

displaying and metric space models are additionally used to dissect inquiries as per the verifiable fleeting articulation of a client.

J. Wang and Wu (2017) proposed an IR framework that examinations implied fleeting questions in these habits for example language demonstrating and metric space models. In the proposed technique, the transient aims of the question were investigated utilizing either DBpedia or the actual archive. In another review Campos, Dias, Jorge, and Nunes (2017), Nonexclusive Fleeting Assessment which is a comparability measure, and an order model (GTE-Class) is utilized to recognize the period in the text. Web indexes generally can't confine the consequences of a pursuit to a specific timeframe. As a rule, the latest outcomes are shown. This study attempted to defeat this inadequacy by understanding the understood time spans of the question. Worldly articulations were identified in the inquiries as well as a bunch of important dates was additionally recovered from sources on the web.

In IR frameworks, mathematical qualities called loads are related with terms in a report to address the significance of each term. The higher the heaviness of a term is, the higher is its significance. Izzah and Girsang (2021) utilized the weighting plan with a SVM classifier and 10 overlay cross-approval procedure to group text, in which the recurrence of a word was partitioned by archive length, and afterward the heaviness of each word in the record was determined. The weighting plan utilized in this study was changed Term Recurrence Converse Report Recurrence (TF-IDF-Assoc). TF-IDF alongside SVM was likewise utilized in Dadgar, Araghi, and Farahani (2016) to characterize news. TF-IDF was utilized for include extraction and SVM was utilized for grouping. The review utilized 2 BBC and 20Newsgroup datasets to order news. H. Ahmed, Traore, and Saad (2018) utilized word-based n-gram portrayal of text and separated Term Recurrence (TF) and TF-IDF elements to group archives into genuine and counterfeit news. For grouping purposes, they utilized Choice Tree, LR, KNN, Straight Help Vector Machines SVM, and Stochastic Inclination Plummet (SGD) calculations.

Bandhakavi, Wiratunga, Padmanabhan, and Massie (2017) extricated feelings from text utilizing a Space Explicit Feeling Dictionary (DSEL) generator considering the Unigram Blend Model (UMM). The DSEL generator learned inclination highlights for pitifully named (tweets) and marked (episode reports, news features, and websites) feeling text. A multiclass SVM classifier was then used to arrange the archives in light of feelings. Kirange and Deshmukh (2012) grouped news reports as per the feelings of the peruses separated from the features of the news. Feelings of shock, distress, satisfaction, dread, revulsion, and outrage were incorporated. Features were taken from the 'Full of feeling Text' dataset and were named and clarified physically by six annotators. The word reference for feelings words was taken from the 'Word Net Influence' information base. The SVM classifier was then used to group the features in view of feelings. One feature could be arranged into more than one inclination.

In a Chinese report Jia, Chen, and Yu (2009) grouped news features in the Chinese language as per the pursuer’s feelings. The eight feelings they characterized were cruise by, blissful, miserable, interesting, furious, exhausting, sympathy, and contacting. For these feelings, they separated the accompanying five classes of elements:

- News Area: DM
- Grammatical feature: POS
- Word/Grammatical feature: Word/Pos Trigram (WPT), Word/Pos Bigram (WPB), and Word/Pos Unigram (WPU)
- Words: Word Trigram (WT), Word Bigram (WB), and Word Unigram (WU)
- Chinese characters: Character Trigram (CT), Character Bigram (CB), and Character Unigram (CU)

Utilizing blends of these elements, features were grouped considering feelings through SVM. In another review Al Masum, Prendinger, and Ishizuka (2007a), two ways to deal with group news as per feelings had been taken on. The main methodology was to involve default feelings for a piece of information e.g., Italy soccer group commending their achievement in FIFA 2006 is a cheerful information naturally. The subsequent methodology was to set feelings as per the client's inclinations e.g., if the client has chosen a pessimistic inclination

for Italy, for him the above news is miserable. To accomplish this, first, they separated etymological parts and their connections from the sentences in the message. Then, they doled out context-oriented valence utilizing SenseNet (Al Masum, Prendinger, & Ishizuka, 2007b) to the etymological parts and made a rundown of named substances. They then, at that point, doled out preset and client inclination feelings to the named elements. At last, they performed two sorts of groupings as indicated by the two methodologies referenced above utilizing the OCC model.

The topographical data of an archive can be a helpful component for recovering successful outcomes for client questions. Geological Data in news has become significant data in literary records. In a concentrate by Zaila and Montesi (2015), geological data concealed in text archives was recovered and examined. In this exploration, the geological philosophy was made from the assets of Wikipedia, WordNet in the wake of recognizing the topographical data the outcomes were positioned and converged with the standard printed positioning rundown to create end-product. In one more review Strötgen and Gertz (2010), a model was proposed in which both worldly data and topographical data were separated and joined that is more significant in archives investigations i.e grouping, representation, and so on UIMA pipeline was utilized to get report sources, extricate worldly and geological data, and store the outcome in the data set. The co-event approach was utilized to check whether the two sorts of data are connected with each other.

In a new exploration Yaşar and Tekir (2020), the fleeting and spatial focal point of literary reports were determined by utilizing worldly entropy with Pointwise Shared Data (PMI). The association among word and spot was likewise determined with the assistance of PMI. This technique was utilized because it utilizes likelihood values to standardize the scores as opposed to utilizing just crude frequencies. The viability and achievement of this proposed approach were assessed on more than one dataset of reports. Datasets from Wikipedia, Twitter, and Flickr were utilized to assess the framework. Watanabe (2018) ordered news archives based on topographical data utilizing three unique methodologies. The first was a basic catchphrase matching strategy. The subsequent one was using NER and spot name disambiguation systems (Geoparser.io and Open Calais). The last one was a semi-directed methodology utilizing a little physically assembled word reference. This classifier perceived names of spots, individuals, and associations. In view of the score related with each term, the most pertinent nations were likewise distinguished.

Imani, Khan, and Thuraisingham (2019) separated areas from news reports at the nation level by utilizing NER that distinguishes sentences that contain geographic data. They then, at that point, distinguished the essential center area for each record among every one of the areas separated from that archive. At last, they applied an administered characterization method to arrange the news reports. Their emphasis was on political information. English, Arabic, and Spanish news stories were essential for the dataset in this exploration. Rao and Sachdev (2017) proposed a model for the characterization of the news in view of city-level area, giving clients an assortment of city-explicit news because of questions. The substance of news stories was separated from HTML pages utilizing their own created web crawler. Their web crawler played out this undertaking in three stages:

- Parsing Truly Basic Partnership (RSS) channels to get Uniform Asset Finders (URLs)
- Gathering URLs in a record
- Separating articles utilizing these URLs

For characterization purposes, SVM, Gullible Bayes, and Irregular Timberland classifiers were utilized among which Arbitrary Woodland classifier played out awesome.

One more review Hassan and Rahman (2017) involved various information digging procedures for examining on the web wrongdoing news stories. Online news stories aren't organized. They extricated data from these unstructured news stories. In the first place, they investigated whether the articles were connected with wrongdoing. Then, they utilized NER and sentence order to separate areas. These areas were then distinguished as 'wrongdoing area' or 'not wrongdoing area'. Wrongdoing areas included areas referenced in sentences containing wrongdoing related terms or the sentences close to them. Areas of police headquarters were additionally viewed as wrongdoing areas. 'Not wrongdoing' areas incorporated the area of emergency clinics or addresses of the people in question/crooks.

Considering wrongdoing areas, comparative or pertinent wrongdoing stories were gathered together.

Mukherjee and Sarkar (2020) concentrated on the crime percentages in various regions by extricating areas referenced in wrongdoing news reports. The undertaking was isolated into two sections; isolating wrongdoing news archives from different reports and removing areas from them. The wrongdoing include per locale in a state was imagined and horror inclined regions were perceived. Multinomial Credulous Bayes classifier classifies wrongdoing and non-wrongdoing news records. Conditional Random Fields (CRF) were utilized for separating areas from wrongdoing news records.

A concentrate by Zhen-fang, Pei-yu, and Ran (2008) classified news stories considering geological data, as well as fleeting articulations related with them. Word vector and word scattering parallelly communicated the proposed model for this review. Word vector was framed utilizing the word2vec model and the text scattering vector was determined. Spatial elements were gotten by contributing word vectors to Multiple CNN (MCNN). Then, at that point, these spatial highlights were inputted to LSTM to learn transient elements. For classification, a mix of deep learning procedures including Multilayer Perceptron (MLP), LSTM, and CNN models.

The system of endurance and genetic advancement of the fittest in normal determination is utilized to find ideal boundaries of attributes in a genetic calculation (X.-P. Wang & Cao, 2002). This calculation works on the exactness and eliminates the deceptive decisions in the classification of archives. The genetic calculation manages global optimization and adaptive probability. These methods are straightforward and solid in this way, they are utilized generally. They are additionally imitated in an indigenous habitat of genetic and natural advancement. This calculation has been utilized by a few scientists to further develop the text classification process. Zhen-fang et al. (2008) involved genetic calculation in text order interestingly. They fabricated and advanced the client format. The deficiencies of the genetic calculation were improved by presenting simulated annealing. Test examination showed that a genetic calculation is a compelling and attainable technique for the classification of information.

If the information is fragmented, or the data is fuzzy, it very well may be managed fuzzy correlation. For a very long-time grouping, it changes the property estimation over to fuzzy sets. T.-Y. Wang and Chiang (2009) analyzed the one-against-one fuzzy support vector machine and one-against-one support vector machine on their Reuter's news information to investigate the difficulties. They discovered that the previous method performed better compared to the last one. In one more exploration [38], the choice rule was improved by planning another calculation that utilized fuzzy correlation and KNN. This new calculation further developed lopsided class arrangement and in general adequacy. To this end numerous other late investigates additionally utilize fuzzy identicalness relations for working on the accuracy of information arrangement. They utilize fuzzy rationale or fuzzy correlation with other machine learning algorithms to further develop accuracy.

3. Methodology

The proposed methodology is shown in figure 2. The initial step included the assortment of information from Google News. The dataset contained a Comma-Isolated Qualities (CSV) record including date and season of distributing, source, connect, title, text, and rundown of 14870 news reports. The following stage was preprocessing of this dataset. In this progression information overt repetitiveness, unique characters, non-English words, and stop words were taken out to clean the information. Overt repetitiveness was taken out in the URL. On the off chance that there was more than one record that had similar URL, just one of those records was kept in the CSV document. Unique characters and non-English words were eliminated utilizing customary articulations. Also, Normal Language Tool stash (NLTK) (Bird, Klein, & Loper, 2009) was utilized to eliminate stop words. When the information had been cleaned, it was given to the subsequent stage. In the following stage, names of areas were separated utilizing NER. These included names of states, urban areas, and nations. They were added to the CSV document. Manual comment was then done to isolate the names of urban communities from nations. Then, fleeting articulations were removed and changed over

to a standard structure utilizing SUTime (Chang & Manning, 2012). They were added to the CSV record also. At long last, this CSV document was given as contribution to the classifier, and the news was grouped based on the center region. The dataset for this study was gathered from Google News. It is a news perusing administration given by Google and was at first evolved by Krishna Bharat. News stories are coordinated on Google News and a continuous progression of their connections is introduced. Various magazines and distributers distribute these news stories on the web.

Google News is the world's biggest news accumulating administration. 14870 news stories were scratched for this dataset utilizing python. The news scrapper was carried out utilizing Pandas and Boa constrictor libraries. News stories were delivered in four different timeframes; that day, that week, that month, and that year were scratched. Articles were gathered from various sources on Google Web including TechCrunch, CNET, The Skirt, Apple Insider, PEOPLE.com, CNBC, Times Envoy Record, Aljazeera.com, Anadolu Organization, Reuters, CNN, and so forth They were connected with Coronavirus, the illustrious wedding, the world's first iPad, Osama container Loaded, Gangnam style, Nelson Mandela, Malala Yousafzai, Robin Williams' self-destruction, ISIS, worldwide environmental change, ruling English ruler, Donald Trump, the biggest ladies' walk, MeToo pandemic, the main photograph of a dark opening, Virginia school football trainer, Imprint Cuban, most extravagant artist. The date and season of distributing, source, interface, the title of the article, text, and its outline were put away in a CSV record.

3.1. Data Preprocessing

Textual dataset is preprocessed to clean it and convert it to a standard organization prior to passing it to the model. Four kinds of preprocessing were done on the dataset in this study.

1. Data Redundancy Removal
2. Special Characters Removal
3. Non-English Words Removal
4. Stop Words Removal

There can be more than one duplicate of similar information in the dataset. Any sort of duplication in information is named information overt repetitiveness. A repetitive information is one that whenever erased, doesn't cause loss of any data. There was some overt repetitiveness in the dataset gathered for this research.

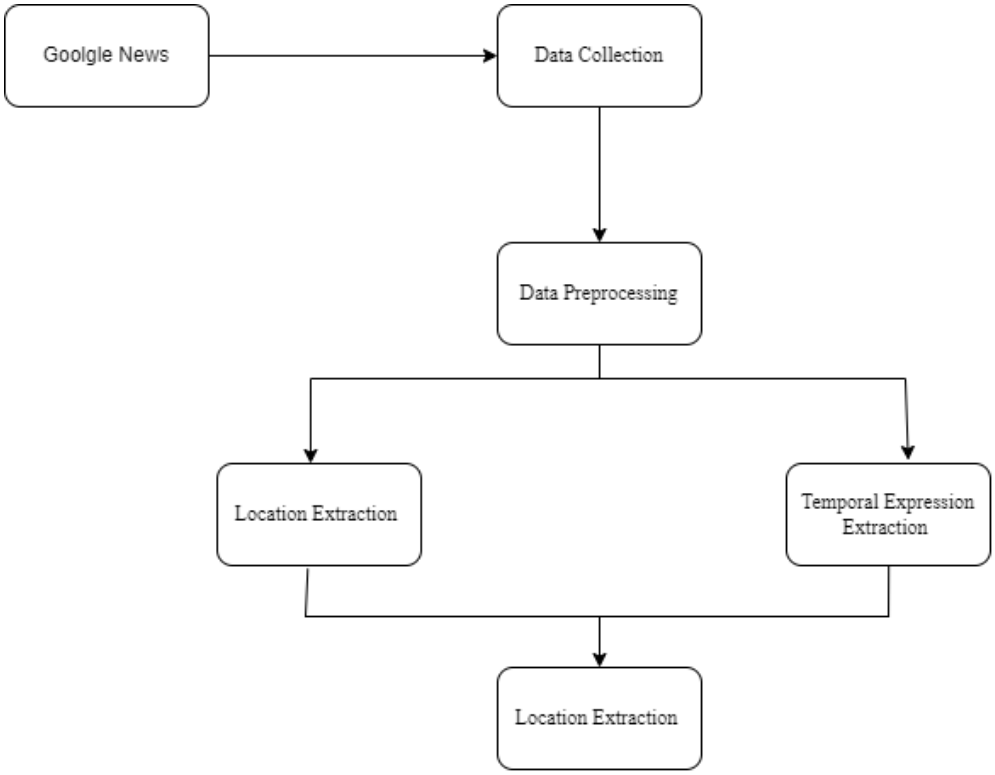

Figure 2: Methodology

This overt repetitiveness was eliminated as the initial step of pre-processing. It was taken out in the connection field. On the off chance that there was more than one record that had similar URL, just one of those records was kept in the CSV document. Albeit the outline of both the records is unique, they have a similar connection, so one of them was erased. The equivalent was finished any remaining repetitive information records. Eliminating repetitive information is vital to sum up the model. In any case, the model sees the copy information as more significant. Remembering copies for information brings about non-random testing, subsequently prompting model overfitting and a one-sided model.

All non-alphanumeric characters are viewed as exceptional characters. These incorporate accentuations and images. In normal language handling, it is critical to eliminate them prior to passing the information to the model. Exceptional characters are most usually found in money numbers, references, remarks, and so on These characters acquire the clamor the calculation and have no worth in understanding the text. The customary articulation had been utilized in this review to eliminate exceptional characters from the text. This customary articulation eliminated all characters that were either highlight (_) or NOT alphanumeric characters (\\w) and blank areas (\\s). Condition for highlight was incorporated independently on the grounds that alphanumeric characters incorporate digits, letters in order, and highlights. Barring alphanumeric characters from expulsion likewise prohibited highlights. Thus, they were taken out independently. Stop words incorporate conjunctions, pronouns, relational words, articles, and so on They are the most well-known words utilized in the English language, yet they don't add a lot of data. The web index overlooks these words while looking as well as while recovering information. These words take up significant capacity and handling time, so they are taken out prior to handling a characteristic language. NLTK offers a library that can distinguish and eliminate prevent words from the text. This study involved NLTK for stop words removal.

3.2. Features Extraction

NER was utilized for area extraction in this review. It is a characteristic language handling procedure that distinguishes every one of the named elements in a flood of text and indicates their sorts. These substances incorporate appropriate names like numbers, organizations, nations, items, rates, money related qualities, time, and so forth NER can distinguish and recognize one token and two symbolic elements. This study utilized spaCy's (Honnibal & Montani, 2017) underlying NER model for area extraction. It involves

capitalization as a prompt to distinguish named substances and afterward group them. The 'GPE' elements were separated that incorporate states, urban communities, and nations. Next the countries and cities were separated.

SUTime was utilized for worldly articulation extraction in this review. It is a characteristic language handling library that perceives and standardizes time articulations utilizing a deterministic rule-based framework. To start with, it removes text that makes reference to some kind of worldly articulation e.g., 'next Wednesday at 3 pm'. It then, at that point, takes a current reference time and converts the text to a standard format.

3.3. Classification

Four different combinations of features were used to classify location, which are as following.

1. News Text
2. News Text and cities
3. News Text and Temporal Expressions
4. News Text, Temporal Expressions, and Cities

For this research work a sequential CNN was designed, having 30 hidden layers with ReLU activation function in first 27 layers and for last three dense layers softmax function was used. Dataset was divided into three parts with 70% training data, 15% validation data and 15% test data. Adam optimizer [54] was used to train the model end to end for 100 epochs. The best model in terms of lowest validation loss was selected. Then batch size of 32 was used to train.

4. Results

To analyze performance of model seven different matrices were used naming micro precision, macro precision, average accuracy, micro F1-score, macro F1-score, micro recall and macro recall. The experimental results are presented in table 1. Experiments indicate the different combinations for features used for classification. As it can be observed in results that fourth experiment in which news text, temporal expressions and cities are used to train the model performed better than all other experiments. In first experiment only text of news documents was used for training, as it can be observed that in terms of accuracy this matric performed well but on all other measures it was lacking performance as compared to all other matrices, for second experiment the model was trained using news text and cities names. The results of this experiment show that the model performed well in terms of precision and recall on the other hand it didn't perform well in terms of average accuracy.

Table 1
Results and Discussion

Experiments	Average Accuracy	Micro Precision	Macro Precision	Micro Recall	Macro Recall	Micro F1-Score	Macro F1-Score
E1	97.35	92.28	92.28	94.97	89.98	93.61	91.12
E2	92.78	96.58	94.38	97.45	92.14	97.01	93.25
E3	98.10	92.46	91.52	93.36	96.36	92.91	93.88
E4	98.42	98.13	99.0	96.75	98.7	97.44	98.85

Model didn't perform well in terms of F1 score and precision in third experiment in which news text and temporal expression extracted from that text were used as input features to the model.

5. Conclusion

Due to the digitization in modern world, most of the individuals tend to get to the online sources to get their news feed. As there is a lot of information is available online people are only interested to read news that are aligned with their interests. There has been a lot of research in this area to deliver only relevant news to their reader. In this research work we

have proposed a framework using hidden features of a news document for classification. It utilizes the news text, cities names mentioned in the text and temporal information present in the document. A deep learning model named Convolutional Neural Network was trained using mentioned features. Three different variants of features were used to train the model. In first experiment only news text was used, in second experiment news text and cities mentioned in that text were used, in third experiments news text, cities names and temporal information present in text were used. Experimental results indicate that the model performed better on presented features as compared all other.

References

Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy, 1*(1), e9. doi:10.1002/spy2.9

Ahmed, J., & Ahmed, M. (2021). ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES. *IIUM Engineering Journal, 22*(2), 210-225. doi:10.31436/iiumej.v22i2.1662

Al Masum, S. M., Prendinger, H., & Ishizuka, M. (2007a). *Emotion sensitive news agent: An approach towards user centric emotion sensing from the news.* Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence (WI'07).

Al Masum, S. M., Prendinger, H., & Ishizuka, M. (2007b). *SenseNet: A linguistic tool to visualize numerical-valence based sentiment of textual data.* Paper presented at the Proceedings of the International Conference on Natural Language Processing (ICON).

Bandhakavi, A., Wiratunga, N., Padmanabhan, D., & Massie, S. (2017). Lexicon based feature extraction for emotion text classification. *Pattern recognition letters, 93*, 133-142. doi:10.1016/j.patrec.2016.12.009

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*: " O'Reilly Media, Inc.".

Campos, R., Dias, G., Jorge, A. M., & Nunes, C. (2017). Identifying top relevant dates for implicit time sensitive queries. *Information Retrieval Journal, 20*(4), 363-398. doi:10.1007/s10791-017-9302-1

Chang, A. X., & Manning, C. D. (2012). *Sutime: A library for recognizing and normalizing time expressions.* Paper presented at the Lrec.

Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016). *A novel text mining approach based on TF-IDF and Support Vector Machine for news classification.* Paper presented at the 2016 IEEE International Conference on Engineering and Technology (ICETECH).

Dilrukshi, I., De Zoysa, K., & Caldera, A. (2013). *Twitter news classification using SVM.* Paper presented at the 2013 8th International Conference on Computer Science & Education.

Hassan, M., & Rahman, M. Z. (2017). *Crime news analysis: Location and story detection.* Paper presented at the 2017 20th International Conference of Computer and Information Technology (ICCIT).

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear, 7*(1), 411-420.

Imani, M. B., Khan, L., & Thuraisingham, B. (2019). *Where did the political news event happen? primary focus location extraction in different languages.* Paper presented at the 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC).

Izzah, I. K., & Girsang, A. S. (2021). Modified TF-Assoc Term Weighting Method for Text Classification on News Dataset from Twitter. *IAENG International Journal of Computer Science, 48*(1).

Jia, Y., Chen, Z., & Yu, S. (2009). *Reader emotion classification of news headlines.* Paper presented at the 2009 International Conference on Natural Language Processing and Knowledge Engineering.

Keith, B., Horning, M., & Mitra, T. (2020). Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. *Computational Journalism C+ J*.

Kim, A., Pethe, C., & Skiena, S. (2020). What time is it? Temporal Analysis of Novels. *arXiv preprint arXiv:2011.04124*. doi:10.48550/arXiv.2011.04124

Kirange, D., & Deshmukh, R. (2012). Emotion classification of news headlines using SVM. *Asian Journal of Computer Science and Information Technology, 5*(2), 104-106.

Krishnalal, G., Rengarajan, S. B., & Srinivasagan, K. (2010). A new text mining approach based on HMM-SVM for web news classification. *International Journal of Computer Applications, 1*(19), 98-104.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). *Recurrent convolutional neural networks for text classification.* Paper presented at the Twenty-ninth AAAI conference on artificial intelligence.

Li, Z., Shang, W., & Yan, M. (2016). *News text classification model based on topic model.* Paper presented at the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).

Mitchell, A., Gottfried, J., Barthel, M., & Shearer, E. (2016). The modern news consumer: News attitudes and practices in the digital era.

Mukherjee, S., & Sarkar, K. (2020). *Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas.* Paper presented at the 2020 IEEE Calcutta Conference (CALCON).

Rao, V., & Sachdev, J. (2017). *A machine learning approach to classify news articles based on location.* Paper presented at the 2017 International Conference on Intelligent Sustainable Systems (ICISS).

Strötgen, J., & Gertz, M. (2010). *Heideltime: High quality rule-based extraction and normalization of temporal expressions.* Paper presented at the Proceedings of the 5th international workshop on semantic evaluation.

Wang, J., & Wu, S. (2017). *Information retrieval with implicitly temporal queries.* Paper presented at the International Conference on Intelligent Data Engineering and Automated Learning.

Wang, T.-Y., & Chiang, H.-M. (2009). One-against-one fuzzy support vector machine classifier: An approach to text categorization. *Expert Systems with Applications, 36*(6), 10030-10034. doi:10.1016/j.eswa.2009.01.025

Wang, X.-P., & Cao, L.-M. (2002). Genetic algorithm: theory, application and software implementation. *Xi'an Jiaotong University Press, Xi'an*, 68-69.

Watanabe, K. (2018). Newsmap: A semi-supervised approach to geographical news classification. *Digital Journalism, 6*(3), 294-309. doi:10.1080/21670811.2017.1293487

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*. doi:10.48550/arXiv.1910.11470

Yaşar, D., & Tekir, S. (2020). Estimating spatiotemporal focus of documents using entropy with PMI. *Turkish Journal of Electrical Engineering & Computer Sciences, 28*(2), 1070-1085. doi:10.3906/elk-1907-10

Zaila, Y. L., & Montesi, D. (2015). *Geographic information extraction, disambiguation and ranking techniques.* Paper presented at the Proceedings of the 9th Workshop on Geographic Information Retrieval.

Zhen-fang, Z., Pei-yu, L., & Ran, L. (2008). *Research of text classification technology based on genetic annealing algorithm.* Paper presented at the 2008 International Symposium on Computational Intelligence and Design.