iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46181

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com

Named Entity Recognition in Social Media Data

Archisha Sharma[1], Shruti Shreya[2], Shrishail Terni[3]
[1, 2, 3]*Vellore Institute of Technology*

Abstract: ***In recent years, a lot of research is being carried out in the field of Named Entity Recognition of Social Media data. It is now easier for anyone to convey their opinions and information without any kind of authentication, rumours have multiplied as social media is now one of the commonly used media in the entire world. The fields in which Named Entity Recognition is used for analysis of data provided from various social media sites, include rumour detection, controversy detection, sentiment analysis, medical field (such as visualisation of the spread of COVID-19 or for various kinds of dietary concerns), topic detection and event detection. The purpose here is to present a summary of the present state of social media research and the impact created by information extracted from it. Through this paper, different methods proposed for the purpose of social media data extraction using Named Entity Recognition, have been studied in detail and a comparison has been provided for the same. Most of these papers use the most common metrics for evaluation of their performance, which includes precision, recall and accuracy. The proposed models have been tested on certain datasets extracted from social media networking sites such as twitter, facebook, etc. and their evaluated performance has been compared to the models proposed by several other similar approaches.***
Keywords: HSN (Hierarchical Self-adaptation Network), Bidirectional LSTM, Adaptive Co-Attention Network, Doc2vec, Demographics, Wordnet, Dynamic Time Warping, Memory Graph, Fuzzy Sentiment analysis, BERT, Intra word Code-Switching, MHA (Multi-head Hierarchical Attention), Named Entity Recognition, NLP, Social Media data.

I. INTRODUCTION

People's communication and collaboration are altering as a result of social media. It helps in providing businesses new ways for communicating with their customers. There is rise in efforts made by academics to understand the consequences. The purpose here is to present a summary of the current state of social media research and the impact created by information extracted from it. Various fields have been studied which use social media as a reservoir of user generated information. With social media being one of the most widely utilised media in the world, it is now easy for anyone to communicate their thoughts and knowledge without worrying about any consequences. Many user interactions on social media platforms are contentious, especially in polarised cultures. Instead of creating a space for discourse, these environments encourage the formation of users who discredit others' positions. Such interactions are common on news websites in the form of news comments. In the information era, social networks allowed us to establish a new paradigm, allowing not just unfettered access to social media but also give the freedom to publish and distribute new contents. Unsurprisingly, a lot of research has been conducted on demographics on online social media platforms such as Facebook and Twitter. The constant rise of social networks has made them one of the most important information sources for researchers and businesses, but pre-processing and analysis remain a significant problem. Named Entity Recognition is a means to recognise particular mentions of named entities from text belonging to predefined semantic types such as location, person, organisation etc. [108] In this paper, several papers using Named Entity Recognition for the purpose of extracting named entities from social media data, have been studied and analysed to a greater extent. The different models and methods proposed by the authors have been evaluated and a comparison of the final results obtained from these papers has been provided.

II. BACKGROUND

A named entity can be a word or a phrase that identifies one item from a set of other items that have the same attributes. [107] Some examples of it would be names of people, organisations, or even location names. When talking about a particular domain, examples for named entities are names of diseases, drugs or genes in the biomedical domain, key terms of emotions in sentiment analysis domain. The process of finding and classification of named entities in general text into predefined categories is known as Named Entity Recognition. In NER, if a list of tokens s = <w_1, w_2 , ..., w_N > is given, the output would be a list of tuple <I_s , I_e, t>, each of which has to be a named entity in s. The start and end indices of a named entity is given by $I_s \epsilon$ [1,N] and $I_e \epsilon$ [1,N], and the entity type from a given category set is t. A variety of downstream applications such as question answering, information retrieval, machine translation has NER as an important pre-processing step.

At least one named entity is present in 71% of search queries. [109] Named Entities are also being utilised for enhancing user experience in query autocompletion, entity cards and query recommendation. A collection of documents that contain annotations of one or more than one entity type is called a tagged corpus. Earlier work on datasets to be used for NER (before 2005) included annotating news articles with a limited number of entity types. This was suitable for NER tasks. Presently, more datasets are being developed on user generated text sources like articles, conversation, social media posts and comments, etc. The total number of tag types in tagged corpuses have also significantly increased.

III.RELATED WORK

The issue falls under the category of text categorization, which is categorising a text as positive, negative, or neutral. Text classification includes the task of sentiment analysis. Several domains have attempted to investigate this tracking problem. Spatial-temporal analysis, spikes explanation, health trending, politics, tourism, and other topics are among them. The majority of these research, on the other hand, have been limited to the language, with only a minor contribution made to the Arabic language. One of the most difficult aspects of this topic is the absence of prepared datasets to use as a corpus. Due to its complicated structure and morphology, the Arabic language poses significant hurdles and issues for sentiment analysis. The vast majority of sentiment corpora are in English or European languages. [1] Two forms of study that can be undertaken on social-media text are news-analytics and opinion-mining. The analysis is an attempt to make use of a significant volume of user-generated news content which are available online. [2] Using machine learning or lexical-based methodologies, managers monitor their clients on certain social media sites by tracking the opinions and extracting information from a large amount of user-generated content. [3] The majority of sentiment analysis text corpora come from websites and social media platforms. The definition of formal rules is required for the annotation of the acquired data. The standard strategy in sentiment analysis is to label the polarity as positive, negative or neutral and it can be done manually, semi-automatically or automatically.[4] Sentiment analysis has piqued the interest of researchers in a variety of applications. One of the applications that has recently been investigated in several domains while focusing mostly on the English language is tracking sentiments through time. [5] The domain of Named Entity Recognition encompasses this issue (NER). The goal of NER is to find entities in a text that have a specific meaning. Social media data can represent drug users' reactions to medication in real time and change swiftly. However, because there is such a little amount of annotated data, with less research done on it. Furthermore, this data has issues with colloquialism and informal vocabulary expression which makes ADR named entity detection difficult (NER). Because of the limited amount of data available from Twitter the proposed approach may be able to successfully handle the problem of not generating compelling findings.[6] This problem comes under the domain of Named Entity Recognition (NER). The aim of NER is to identify entities with special meanings in the input text. However, because there is such a little amount of annotated social media data, there is less research on it. Furthermore, social media data has issues with colloquialism and informal vocabulary expression, which makes ADR named entity detection difficult (NER). Because of the limited quantity of data available from Twitter, the proposed approach may be able to successfully handle the problem of not producing compelling findings. [12] ADR identification can be stated as a NER task. The ultimate goal of the NER job is to identify entities and types in text, which is required for numerous natural language processing (NLP) tasks such as relation extraction (RE).[13] Transfer learning is a useful model for low resource tasks in the high resource domain because it can acquire generic features or generalised knowledge. The feature or information is then transferred to a domain with limited resources. [16] Many strategies have used adversarial networks to transfer features from source to target resources in order to improve the extraction of domain invariant information.[17]

The domain of Named Entity Recognition encompasses this issue (NER). The goal of named entity recognition (NER) is to find entities in a text that have a specific meaning. Various methods for incorporating visual information from images into NER on social media are being investigated with remarkable success. Irrelevant photos have the potential to introduce noise into existing models. Because social media post messages are unstructured, large-scale data mining on social media is incredibly difficult.[18] The authors set out to anticipate the correct Named Entity sequence of multimedia posts in this paper. This problem falls within this category since it necessitates the use of natural language processing to address it. Multimodal representation, Bilinear Attention Network, and Adversarial Gated Bilinear Network are some of the solutions to this challenge. Only English is used in this paper to describe NER. One of the most significant problems in developing this system was a lack of data. Adversarial learning and a Gated Bilinear Attention Network were used by the authors. T-NER, CNN + BiLSTM + CRF, VAM, and others have done work in this particular area.[21] To solve some of the challenges discovered while utilising existing models, the study introduces a novel HSN (Hierarchical Self-Adaptation Network) model. It's an algorithm that isn't supervised. The fundamental goal of the multimodal NER job is to identify named entities in user-generated postings that include both text and images.

There are cases where images are misaligned or absent from the text, where existing models employ single attention and disregard numerous entity objects in both texts and images of postings, and therefore are unable to deliver outstanding performance. The proposed algorithm solves this issue. [28] Zhang et al proposed the Co-attention model in a similar problem using the simple operation for concatenation for fusing textual features and multimodal features. It ignored plain text information. [29] The plain text information has been neglected in this work by Lu et al. It captures the relevance between the image and the text via an attention method. It calculates the matching between both features with a single attention, resulting in incorrect or incomplete attention in multimodal interactivity. [30]

In Arshad et al's work, plain text information was ignored. It uses an attention strategy to capture the relationship between the image and the text. In multimodal interactivity, it calculates the matching between two features with a single attention, resulting in incorrect or incomplete attention. [31] Image captioning is a technique for creating a word from a visual area that is most closely related to the most recently generated term. With the purpose of obtaining richer semantic connections across multiple modalities, the proposed study employs an attention implementation similar to those employed in VQA. [33] The extraction of named entities from tweet-based texts is paired with an examination of hand-crafted features taken from other modalities such as hyperlinks and images in this work. To finish the challenge, the features collected using a bi-dierctional LSTM, a hybrid deep neural model, and a CNN are mixed, followed by a conditional random field. It focuses on the usage of hyperlinks and hand-crafted features, as well as multimodal deep learning-based models, to handle the problem of multimodal NER from Twitter data. [34] To improve named entity recognition, Wikipedia is leveraged as a source of external knowledge. On the basis of each candidate word sequence's Wikipedia entry, category labels are extracted and employed as features in a CRF-based NE tagger. [35]

The researchers combined neural network models with spelling features to improve NER performance in this study. BiLSTM provides these features in conjunction with word features, which are used as input features for CRF decoder.[37] Word embeddings were used in this research with the purpose of utilising relevant lexical information. This ensures that the representations and their use for labelling tasks improve. Bidirectional LSTMs were utilised to solve this problem. It has shown that there have been significant increases in performance when using lexicons.[38] Because it is easier for anyone to communicate thoughts and knowledge without any kind of validation, rumours have exploded as a result of social media being one of the most widely utilised media in the world. The ESODE framework is used to present a novel strategy for rumour detection on social media, which can aid in the integration of sentence reconfiguration, entity recognition, and ordinary differential equation networks. [39] A task-specific character-based bidirectional language model was combined with stacked LSTM networks to capture social-temporal settings and textual contents. Multi-layered attention models were utilised to jointly train attentive context embeddings. [42] Human emotional reactions to various rumour events are recorded in order to divide the posts, during which the variation of sentimental and contextual information of each event is recorded over time in order to detect rumour events.[44] The Corona virus has recently become a popular topic of debate in online forums, resulting in a vast volume of social media data that might be used to strengthen crisis management in a variety of ways. A new framework for monitoring virus spread in Italy has been proposed, which includes geo-tagging tweets based on the locations mentioned in the text, using a face detection algorithm to estimate the number of people appearing in posted images, and using a community detection approach to identify Twitter user communities. [45]

This paper presents a new methodology for processing, categorising, visualising, and analysing big data knowledge offered by the sociome on social media platforms. A methodology is described that includes natural language processing methodologies, machine learning algorithms, ontology-based named entity recognition methods, and graph mining techniques. [51] This concept investigated the use of ontology-based procedures for analysing radicalization indications in online messages, as well as the benefits of text mining over other methods.[52] Finding the most influential people in the twitter diabetic groups using a combination of non-grouping methodologies from network science, ethnography, and information retrieval, as well as health-related implications for public health professionals and policymakers. [53] An unsupervised mining visualisation technique is utilised to analyse the content of Reddit-published user messages, which can be used to compare mental health communities and to design and guide new patient education programmes. [54] A full summary of numerous location prediction systems is available on Twitter. The suggested approach is based on a combined analysis of three essential user profile registers: semantic processing of claimed location, declared time zone, and GPS coordinates. [55] An exploratory investigation of 24,634 tweets relating to human bowel disease has been published, with several user characterisation methodologies and text mining proposed for uncovering important health outcomes to enhance decision-making among the various user roles.[56]In social media engineering, demographics are critical. Yahoo Answers is an online community-based question-and-answer platform that allows people from all over the world to participate in question-and-answer forums.

Diverse age groups have different interests, needs (e.g., some medications and assisted care), values, incomes, and shopping habits, as is well documented. In addition to analysing these clusters, it is simple to examine how these discrepancies alter over time, such as how personal expenditures change as we become older. [57]

They noticed that open questions got less updates than those that are answered, leading them to conclude that these changes are made to promote readability and engagement. Chua and Banerjee examined why some enquiries elicit responses while others go unanswered prior to their study. Asker traits (such as popularity and participation) and content elements (such as degree of details, specificity, clarity, and socio-emotional value) were found to be good predictors in this study. Extrinsic factors that influence the chance of replying to inquiries on Sina Weibo were investigated by Liu and Jansen [30]. [58] There has been a lot of research done on analysing several aspects of cQA friends, as well as many other social networking platforms. [64] Among other things, they looked into how members react to shifting norms at different stages of their group's life. Individuals discovered that they become increasingly attentive to community norms until they reach maximum synchrony with the community language about one-third of their projected lifespan; after that, a gap between their language and the community language forms and widens until they abandon it. [59] Prior to their investigation, Chua and Banerjee investigated why some inquiries elicit responses while others go unanswered. Asker traits (such as popularity and participation) and content elements (such as degree of details, specificity, clarity, and socio-emotional value) were found to be good predictors in this study. [61] The responses in this tab are sorted using machine learning approaches that predict the usefulness of responses. This tab only displays content that their system deems to be of better quality. [62] One of the most important tasks in Natural Language Processing is sentiment analysis, sometimes known as opinion mining. This project allows you to examine user opinions, feelings, emotions, and product or service assessments. This is why it is critical that research institutes and enterprises meet their obligations. Plan for future improvements in the quality of their products or products and services. The primary source of sentiment data is social media. Users are more likely to comment on items and services in this type of analysis. As a result, numerous works on sentiment analysis have been published. [63] Fuzzy logic is used in a variety of fields to replicate notion opposites and the uncertainty that comes with them. The resources WordNet and SenticNet are used to generate a knowledge graph. The graph is then used to transfer sentiment information gathered from labelled data sets using a graph-propagation technique. The graph is divided into two levels: one shows semantic connections between concepts, while the other shows connections between concept membership functions and domains. [64]

A new lexicon-based study has been published. Using the SentiWordNet and fuzzy linguistic hedges, the authors describe an unsupervised method for creating sentences and analysing their emotion scores and polarity. Using a fuzzy entropy filter and k-means clustering, the method may also extract essential keywords for sentiment analysis. [65] For sentiment analysis of movie reviews, the author presents a fuzzy rule-based technique. The authors claim that fuzzy logic is better adapted to the intrinsic ambiguity of natural language, and that rule-based learning techniques provide more interpretable results. In terms of accuracy, the method produces somewhat better results than baseline methods and adds interpretability to the sentiment categorization process. [66] The authors proposed a new way for encoding feelings that is based on the Type-1 Ordered Weighted Averaging operator and uses linguistic term sets rather than numerical values (T1OWA). The related fuzzy set representations of linguistic concepts produced from each user's most significant opinions are also aggregated using this method. [67]

A fuzzy-based multi-domain sentiment analysis technique was proposed by Dragoni et al. [30]. The method uses any conceptual domain overlaps to build wide models for computing sentiment polarity in texts from any domain. The learned polarities are represented using fuzzy logic, and then merged with SenticNet and General Inquirer vocabulary-derived conceptual knowledge. The findings support the feasibility of the concept. (68) A social media event has five stages: beginning, developing, climaxing, descending, and disappearing. By detecting the stages of evolution of social media events like Twitter and Sina Weibo, businesses and governments can take action before emergent phenomena like Twitter and Sina Weibo become unmanageable. The foundation paper's main concern is predicting the evolution of social media events. [69] NLP, Text mining, IR, and other areas have seen fast growth in knowledge graphs (in this study, we just use the term "graph"). A recent survey on knowledge graphs was published in the literature, and it covered representation, acquisition, and applications. [70] A graph kernel is a technique for calculating the similarity of two graphs. In the recent decade, graph kernel computing has gotten a lot of attention, and it can be employed in a variety of ways. The Convolution kernel is the most often used approach, which decomposes each graph into a set of subgraphs and compares them pairwise. [71]

Cai et al. used four operations to depict the dynamics of events in order to discover the evolutionary relationship between them: generate, absorb, divide, and merge. These event operations are essential for tracking the progress of events. They also developed a multi-layer structure to increase the performance of a typical inverted file and speed up event evolution monitoring. [72] A new method was proposed for recognising and tracking real-time occurrences.

They started by making a graph out of a list of entities, using the items as nodes and the similarities as edge weights. A community detection algorithm was then used to extract the community from the graph, and the community was given the name event. Finally, they use the found communities' shared entities to link to them. Even while this research includes event evolution tracking, they are more concerned with topic recognition than with evolutionary stages. The focus of this research, on the other hand, is on determining the evolutionary stages of occurrences. [73]

Because social data has an inherent visual element, some knowledge-graph research focused on social media data analysis. Despite the fact that knowledge graphs have gotten a lot of attention in recent years, this study doesn't dig into the theoretical aspects of them. Our fundamental goal is to make knowledge graphs a powerful tool for representing events and interactions. The event's progression stages are then discovered using graph-based text similarity. [74] There are signs that individuals' ideological views on social media are so evident that it would be simple to foresee the start of a conflict by studying how people interact in news debates. If this measurement was collected during the dialogue, it might be possible to anticipate the outcome of future polarisation and argument. [75] For decades, researchers have studied the causes and consequences of polarisation in human society. In a landmark study, Zachary (1977) revealed that the homogenization of viewpoints in human groups aids the long-term stability of these organisations. He also discovered that information sharing between organisations causes conflict. [76] During the 2004 presidential election in the United States, Adamic and Glance (2005) revealed that users of conservative and liberal blogs had different connection behaviours. Users that shared the same political inclination interacted more and talked more of each other's material, according to the survey. [77]

Using topic models, Choi, Jung, and Myaeng (2010) studied which specific themes created the most significant disagreement in networks. The authors employed sentiment analysis to identify texts that had a high emotional impact, and then used topic models to create searches which would yield such documents. According to the findings, the questions accurately described the opposing subjects [78]. Popescu and Pennacchiotti (2010) developed subject classifiers for Twitter which distinguished between controversial and non-controversial events. They accomplished this by incorporating lexicon-based linguistic aspects such as an OpinionFinder emotive lexicon, a lexicon of disputed Wikipedia words, and a lexicon of horrible words. They investigated the effects of tweet volume, spatiotemporal location, and linguistic features, and discovered that all of these variables were useful for the task. [79] Popescu and Pennacchiotti (2010) developed subject classifiers for Twitter that distinguished between controversial and non-controversial events. They accomplished this by incorporating lexicon-based linguistic aspects such as an OpinionFinder emotive lexicon, a lexicon of disputed Wikipedia words, and a lexicon of horrible words. They investigated the effects of tweet volume, spatiotemporal location, and linguistic features, and discovered that all of these variables were useful for the task. [80]

The challenge of mapping free text to relevant entities in a structured knowledge base (KB), such as Wikipedia, is referred to as entity linkage. When natural language material is linked to huge knowledge networks, applications can take advantage of rich semantic relationships that are implicit in natural language but clearly stated in the knowledge graph. The problem of mapping free text to relevant entities in a structured knowledge base (KB), such as a database, is referred to as entity linkage. Wikipedia. Natural language content can be linked to massive knowledge networks that are implicit in plain English but express themselves directly in the knowledge graph, allowing applications to exploit rich semantic links. As a result, recent research has discovered that many entity linking models perform poorly when applied to social media material. [81] To address this issue, the concept of "zero-shot" learning was developed (Logeswaran et al., 2019; Shi & Weninger, 2018). One of the biggest benefits of zero-shot learning is that the model can accurately map things even if it hasn't been trained on any. Others have used Wikipedia's type systems to better comprehend the multiple contexts for a specific mention in order to improve model disambiguation. [83] Yamada, Asai, Shindo, Takeda, and Matsumoto (2020) used a new masked entity prediction task to train BERT. They were able to introduce contextualised entities to BERT using a novel pre-training task, allowing for enhanced performance on a range of entity-based tasks. [84] To extract information from tweets, Fang and Chang (2014) develop an end-to-end spatiotemporal model. Unfortunately, the data utilised in these three articles isn't available to the general public.[85]

A paucity of social media training and testing data, as well as a focus on only specific aspects of the problem, such as news headlines and Wikipedia, are some of the challenges that researchers in this field face. We evaluate some of the most recent entity linking strategies in this paper, and we urge that future research focus on social discussion boards like Reddit, with holistic end-to-end models rather than fragmented approaches to entity linking. [86] In recent years, a plethora of academic and commercial NER systems have emerged, the majority of which are characterised by generic categorization schemas, or schemas targeted at capturing broad knowledge about the world by providing basic conceptions and concepts for things. The ability to swiftly access and exploit these complex NER systems via APIs or pre-trained models is required for data integration, question answering, privacy protection, and knowledge base creation. [88]

Because tasks in Natural Language Processing (NLP) are intended to evaluate texts published in a single language, mixed words are frequently disregarded. In the instance of Language Identification, for example, CS processing has altered this assumption (LID). LID is the process of determining the language type of a text. It is one of the most significant pre-processing methods in natural language processing. The majority of this field's study has been on the document level. In CS, the focus has shifted to the word level. Despite this, few scholars have concentrated on subword-level language recognition, which involves segmenting mixed words and assigning a language ID to each fragment. In most LID systems, CS intra-words are labelled as mixed words, and their internal information is lost as a result. [91] The limitations are especially important for morphologically rich languages like Arabic. Indeed, Arabic social media are often characterised by a large variation of unstandardised dialectal Arabic, where speakers use Arabic letters as well as arabizi (Yaghan, 2008), an informal Arabic chat alphabet in which words are written in their transliterated form using Latin characters and numbers that replace some letters/syllables. [95]

Thanks to its API, researchers were able to explore its social big data in a streaming or query-based data retrieval mode. The Twitter developer page has documentation for the API. This API allows for direct messaging, search, advertisement management, and account activity control. It has a number of restrictions in place to prevent developers from abusing the service. For example, it has a rate limit for users or programmes. In recent years, more spatiotemporal models in social networks have been developed to help with a range of issues, and they are classified as follows based on how data is collected (across time or space): (1) link based - models that use link analysis (e.g., Page Rank) to find users with experience and unique locations. (2) content-based - models that combine data from a user's profile with geographic information. (3) collaborative filtering - models that impact users' decisions (e.g., location history). (4) time-progressive - models that consider the likelihood of an impact (i.e., the immediate, near, and far future). [100] Sentiment analysis has a long history, dating back to the 1950s when it was largely used on written texts. This is the first Arabic study to use social media to track the mood of news organisations over time. Sentiment analysis is a technique for extracting subjective information from online content such as texts, tweets, blogs, social media, news items, reviews, and comments. Natural language processing (NLP), statistics, and other methods like machine learning are used to do this. In this subject, which is dominated by English and other European languages, the Arabic language has gotten minimal attention. [101]

IV. DISCUSSION

In order to overcome some of the challenges discovered while using existing models, a novel HSN (Hierarchical Self-Adaptation Network) model was developed. The multimodal NER (Named Entity Recognition) job's main purpose is to recognise named entities in user-generated postings that include both text and graphics. Visual features with appropriate image and text alignment have been found to improve all previous multimodal NER approaches. In social media data, however, this isn't always the case. There are times when the photos are mismatched or absent from the text, and these models are unable to deliver the high-quality results that they typically do. It's also worth noting that earlier models largely neglect the existing multiples, they just use single attention while capturing the semantic interactions across distinct modalities in both texts and photos of posts. The unique model described in this research addresses all of the challenges that have been found. This HSN model includes a cross-modal interaction module for fostering the semantic interactions of many entity objects across multiple modalities. Self-adaptive multimodal integration module for dealing with difficulties such as missing or mismatched images with texts. It also aids in the restraint of fusion feature noises and the distribution of greater weight on them. Multimodal attention has previously been successfully applied to both language and vision-related tasks. It allows models to concentrate on both vital aspects of a work, such as visuals and text. This approach has already been used by VQA to locate sections in photos that are the most closely connected to the text given. It's also been employed in picture captioning to produce a phrase based on the visual area that's most similar to the last one. This paper's implementation strategy is discovered to be similar to that of VQA, which strives to increase semantic interaction between many modalities. The following are the changes in the methods used: Multi-head Hierarchical Attention is used to capture important aspects, which are then fused to provide multimodal features. Using bi-directional integration, noise in fused features is effectively reduced. This also enables for the full usage of both fused and plain textual features in the allocation. The HSN model's anti-interference capacity and flexibility were tested in real-world scenarios using a Real-World dataset for NER. This collection includes both plain-text and multimodal Twitter postings. This model provided state-of-the-art results on both the regular Twitter multimodal NER dataset and the Real-world multimodal NER dataset after extensive testing. The findings acquired from the tests in both datasets aid in confirming the proposed model's efficiency. The suggested model's efficiency is then validated by comparison experiments on various datasets. The results of the qualitative analysis are then presented to show that the attention modules used are interpretable. Recall, precision, and F1 score were the evaluation measures employed in the trials.

Only if both the type and the boundaries match ground truth is a named entity's recognition regarded correct. The proposed approach will be expanded in the future to handle more multimodal tasks such as fine-grained name-tagging and entity-linking. [28]

This paper tries to solve the problem of multimodal NER from Twitter data. Named Entity Recognition from tweets is found to be very challenging because of limited length of tweets, presence of noisy texts, presence of hashtags (#) and association of tweets with images and hyperlinks. Existing research on twitter Named Entity Recognition excludes hyperlinks and hand-crafted features, instead focusing on multimodal deep learning-based models. However, in addition to extracting named entities from tweet-based texts, this study looks into the hand-crafted features extracted from various modalities including hyperlinks and graphics. To finish the challenge, the features retrieved using a bidirectional LSTM, a hybrid deep neural model, and a CNN are integrated, followed by a conditional random field. CNN+BiLSTM+CRF, BiLSTM+CRF, and the Adaptive Co-Attention Network are all state-of-the-art models used in the model. Experiments with multimodal Twitter data, such as URLs, photos, and text, have shown that using these character-level hand-crafted characteristics significantly improves system performance. The proposed models' results were displayed using the standard NER dataset (CoNLL 2003 dataset). This Twitter-based dataset was utilised as a Word limit, and its simplistic design makes it simple for people to use. Data scraping for research purposes is also possible with Twitter app development (which isn't possible with other social media platforms like Instagram and Facebook). It is a better choice for operations like information extraction, named entity recognition, and text mining than tweets because of the enormous number of tweets. In the current environment, tweets are the most commonly used in various fields linked to social media data analysis, including hate-speech detection, rumour detection, cyber-bully detection, opinion mining, event identification, emotion and sentiment recognition and analyses. The performance of several models on the dataset was assessed using metrics such as f-measure, accuracy, precision, recall, and f1-score. Illustration of the reason for the picture elements not working would be supplied as part of future work. Deeper neural network designs would be used to extract visual features like EfficientNet, ResNet, and MantraNetre.Different features from the photos would be extracted to support text based on recent CNN models. Work will also be done to create a large-scale multimodal tweet corpus with URLs, text, and images. The construction of a few shared representations between the text and image modalities is also part of the future effort. Some of the most recent captioning models for photos will be used to generate captions for images, which will help improve NER accuracy by boosting context once they are applied to the text. On the finer version of the dataset, ideas for supplementary information extraction using various attention methods (such as weighted attention) will be applied. [34]

Because it is now easier for anyone to convey their opinions and information without any kind of authentication, rumours have multiplied as social media has become one of the most commonly used media in the world. Previous research in the topic of rumour detection has been done. However, these were primarily concerned with hand-extracted characteristics and hence spent very little time representing text. In this study, a unique approach for rumour detection on social media is proposed, which can aid in the integration of sentence reconfiguration, entity recognition, and ordinary differential equation networks under the ESODE framework. The entity recognition algorithm employed is intended to aid in the semantic comprehension of these rumour texts. The sentence reconfiguration, on the other hand, was created with the goal of increasing the frequency of significant terms. Additional statistical features from the three primary aspects are collected to form the overall feature map. The rumour content's linguistic qualities are disseminated because of user characteristics. Structures of propagation networks Finally, the ODEnet (Ordinary Differential Equation Network) is employed to detect rumours. In this rumour detection test, both sentence reconfiguration and entity recognition are found to play a substantial role. The framework of the proposed rumour detection model contains the following steps: The ER (Entity Recognition) block extracts explanation sentences (ES) from existing knowledge bases extracted from Weibo or Twitter texts. The SR (Sentence Reconfiguration) block is used to reconfigure the combined sentences using the Stanford Dependency Parser (SDP) and the proposed sorting mechanism. In the embedding block, the statistical features from the statistical block are combined with the document embeddings from Doc2vec. With the use of uniform sampling, a feature map is constructed. Finally, a detection result is created when this feature map is fed into the Classifier. The ODE-net was inspired by the CNN model, which consists of four MaxPooling layers, four Conv2D layers, and two Dense layers. As a result, this CNN model was utilised as a baseline in this study to demonstrate how the suggested model has improved.

The experimental results obtained on these datasets from Weibo and Twitter show that the proposed technique outperforms its predecessors significantly. To produce results in the setup that are comparable to relevant works, a 10-fold cross-validation is used to calculate Recall, Precision, and F-Score for each of the analysed methods. Three text classification models, LSTM-vote, GRU-2, and SAtt-BLSTM convNet, as well as three rumour detection models, RPDNN, GLO-PGNN, and STS-NN, were compared to the results of the proposed method. [40] The COVID-19 has recently become a popular topic of discussion in internet forums, resulting in a vast volume of social media data that might be used to strengthen crisis management in a variety of ways.

In this research, a novel framework for evaluating, collecting, and displaying Twitter postings is given, which will be used to track the spread of the virus that has affected numerous persons in Italy. The techniques presented and evaluated include a deep learning localisation technique for geotagging posts based on locations mentioned in the text, a face-detection algorithm for estimating the number of people who appear in posted images, and a community detection approach for identifying communities of various Twitter users. In addition, the collected posts are examined for the purpose of predicting their dependability as well as detecting trending events and subjects. Lastly, an online platform is exhibited, which includes an interactive map for filtering and presenting analysed posts, a visual analytics dashboard for proper presentation of the findings of the themes, communities, and approaches for event identification, and a localization technique's consequence. The evaluation measures used to compare the proposed system's performance to that of the baseline system and state-of-the-art techniques are precision, recall, f1-score, and runtime. Face counting models can be evaluated using the metrics Mean Absolute Error (MAE) and Mean Squared Error (MSE) (MSE). On a multi-level level, an integrated system is given that aids in the real-time collecting, analysis, and visualisation of Twitter posts. For automatic geotagging, a biLSTM-based model that is specifically trained for the Italian and English languages is used. The Evalita2009 dataset was used to test Italian, while the CoNLL2003 dataset was used to test English. A mix of runtime and F1-scores was used to choose the most effective models. In terms of face identification, the Tinyfaces methodology, as well as the other available SoA detectors, have been implemented and assessed for the specific job of counting faces. The results show that in extremely congested scenarios on small scale faces, the performance is considerably better. Finally, comparing the different community detection strategies in terms of execution time and modularity revealed that the Louvain algorithm, which was used in this paper, was superior. With the steady growth of the collection, various procedures could be adopted in future work for this system, such as developing more advanced Interactive Map filters, maintaining a relatively lower response time of displaying the results, and also extending the Visual Analytics Dashboard with additional methodologies that can help end-users' situational awareness and response to the current pandemic crisis. Consolidating qualitative and quantitative data, such as statistical and demographic data, related to the current COVID-19 pandemic crisis in the specific region of interest, could be very useful in this direction for improving assessment and situational awareness among relevant stakeholders, as well as assisting them properly in the effective decision-making process.[46]

This paper presents a new methodology for processing, categorising, displaying, and analysing the sociome's big data knowledge on social media platforms. A methodology is described that includes natural language processing methodologies, machine learning algorithms, ontology-based named entity recognition methods, and graph mining techniques. The goal is to Using domain ontologies to reduce the lexical noise caused by varied methods of users' expressions. Reduction of extraneous communications by guaranteeing accurate identification and focusing on the patient and other individual's experiences from public discussion Individual demographic data is inferred by a combination examination of geographical, linguistic, and visual profile information. Obtaining information on various shared resources, including social media data as well as semantic analysis of web content and Using knowledge graph representation techniques and semantic processing of public speech to do a community detection and evaluation of the health issue study.

In this paper, a thorough examination of the least-studied area of public health on Twitter (immunology disorders and allergies) is carried out. As a result, a large number of health-related conclusions are available. It's mostly about gluten-free foods. Natural language processing, machine learning techniques, and named entity recognition are all utilised in the proposed method to classify unique user profiles and the messages they compose. Multiple complementary methods, such as clustering and knowledge graph reconstruction, have also enabled a multi-layered, holistic analysis for acquiring new information and investigating many degrees of viewpoint and details. On a broad level, the proposed research is a synthesis of the several methodologies used in relevant research in this field. The proposed model serves as a whole new and comprehensive approach for analysing social media knowledge from a biomedical perspective that is primarily focused on patient and individual experiences as a result of combining all of these methodologies. The rates of agreement for the various characterization techniques were calculated using common measures such as recall, F-score, and accuracy. In light of the results acquired from the various methodologies used, the methodology employed has proven to be effective. This proposed method has proved to be of significant relevance by studying almost 1.1 million unique tweets/messages which were composed by almost 4 lakh distinct users related to this field of gluten-free food. Finally, future work will focus on the integration of other language ontologies, despite the fact that a portion of the techniques created in this study is language independent in order to process more messages and evaluate a comparatively bigger volume of the world population. [52]

Social media has evolved into the most comprehensive, broadest, and dynamic collection of data on human activity. It aids in the comprehension of organisations, individuals, and cultures. Annotating datasets from social media news, as well as tracking public attitude toward news entities, is a critical topic. Tracking the evolution of sentiment in different locations, as well as keeping cyber concerns under control, are becoming more important. Several domains have attempted to investigate this tracking problem.

Spatial-temporal analysis, spikes explanation, health trending, politics, tourism, and other topics are among them. The vast majority of these studies, on the other hand, have focused on the English language, with only a modest contribution to Arabic. Two types of research on social media texts have been done: news analytics and opinion mining. Managers of a social media site can keep an eye on its users by controlling their opinions and extracting subjective data from a cornucopia of user-generated content. This is accomplished through the use of machine learning and lexical-based techniques. Technology like Thomson Reuters News Analytics (TRNA), which collects and analyses news data in real time and targets news feeds in English, is commonly used for news analytics. On manually annotated data, the annotation process is semi-automated using machine learning techniques. The suggested system collects data from Twitter's API (Twitter API) streaming service. Then, using a sentiment analysis model that is a semi-supervised neural network technique, a sentiment class is automatically assigned to each tweet. Individuals, places, organisations, and events are then identified using the named entity recognition paradigm. Another tool, a third-party tool, is used to extract the detected items from each tweet. The suggested work intends to develop and annotate a massive, brand new Arabic news sentiment corpus from Twitter due to the unavailability of an acceptable Arabic news sentiment corpus. A novel mechanism for tracking sentiment toward news entities has also been developed. The final models were evaluated on a dataset that had never been seen before. This would help to ensure that the datasets generated are of high quality. The system was evaluated using F1-score measurements. It considers both Precision(P) and Recall(R) of the data while calculating the score of the test set. Existing Arabic datasets were either short in size, lacked temporal markers, or employed entirely automated ways to annotate information such as user ratings. A semi-supervised algorithm was used to annotate the data in the proposed corpus. Designing a system to crawl, annotate, and analyse data in real time, both sentiment and category, is one of the future projects. [1]

A serious public health risk is ADRs. Drug safety issues, such as adverse drug reactions (ADRs), have become increasingly prevalent now that a wide range of medications are mass produced. As a result, text-based detection of drug and adverse reaction entities has become increasingly important in pharmaceutical safety. Because of the rapid development of social media, people have been encouraged to share information and debate issues online, including their concerns about their own hazardous drug responses. This data is considered real-world and more up-to-date than data from other sources, making it particularly relevant for ADR research. Conditional random fields, hidden markov models, support vector machines, and rules-based techniques were employed in previous research on named entity recognition (NER). A shared layer or an adversarial discriminator were used to achieve feature-sharing. Some examples of its application include a common Bi-LSTM and a private Bi-LSTM for Chinese NER, an adversarial discriminator for POS tagging, and a dual adversarial transfer network (DATNet) with a shared Bi-LSTM layer for conducting NER. Extracting entities from diverse sources has also been done using shared layers with adversarial discriminators. The proposed study intends to incorporate shared biomedical information from PubMed into a social media dataset using an adversarial transfer learning algorithm. Because of the limited quantity of data available from Twitter, the proposed method may be able to overcome the difficulty of not producing compelling results. charCNN is used to tackle spelling problems because it is better at capturing the common characteristics of linked structures. To better learn characteristics from various sources, PubMed and Twitter both use the charCNN structure. The importance of each job was balanced, and a scale parameter was imposed to integrate the loss functions of the various tasks. By adding biological domain knowledge from PubMed, the suggested model exhibits significant improvement in performance on Twitter data (target resource). Biomedical data was exchanged, but Twitter data's unique characteristics were kept. To obtain the desired outcomes, the model mixes Bi-LSTM and CRF with additional methods. The tests employed two different PubMed datasets and one Twitter dataset. The source databases are ADE and TwiMed-PubMed, respectively. The target dataset was the TwiMed-Twitter3 dataset, which was retrieved using the Twitter ID. The basic study's target dataset consists of three types of entities: drugs, symptoms, and diseases. The system was evaluated using F1 scores. The top F1 scores on the suggested technique were from TwiMed-PubMed (68.58 percent) and ADE (67.36 percent). The results reveal that the adversarial transfer learning framework outperformed the other models on TwiMed-Twitter. When the results of the suggested approach were compared to those of the KB-embedding RNN, it was obvious that adding domain information improved the results significantly. The proposed model could obtain the set target biomedical domain information by employing the method of mapping data-shared feature space. The target specific features were determined to be more effective in the task presented. The proposed model does not annotate the unlabelled data, unlike the co-training technique.[7]

Nowadays, the issues relating to adverse drug reactions are being considered to be of greater importance than ever before. Also, with internet usage increasing in the 21st century, everyone has got a platform where they can voice their opinions or grievances so as to gain support from other people. This is why they post stuff on the internet and it is these posts that make up resources for ADR recognition due to their timelines. At the same time, the fact that such datasets are very rare cannot be ignored.

Apart from these standard issues faced in the identification of ADR, it is the informal nature of the social media platforms which makes the identification of ADR all the more difficult. There are two models or efforts to handle the same problem in the earlier works. In the first, ADR identification was supposed to be formulated as a NER job. The NER task's main goal is to detect the components and types in the text, which is a prerequisite for various common language handling (NLP) tasks like connection extraction (RE). It prohibited ADRs from using online media posts that contained useful material that could be extracted. As a result, a few experts have recommended antagonistic exchange networks for move learners and ill-disposed organisations to collaborate. Ill-disposed exchange organisations, such as NER and text grouping, have been examined in various NLP projects to improve execution. To advance the semantic representations of low asset dialects, they proposed an antagonistic exchange network model to analyse cross-lingual information. They focus on internet media ADR recognised proof errands in this piece. They subsequently went on to obtain a bilinear consideration instrument with choose elements that are superior for the objective task from the common part. They also presented a one-of-a-kind disaster that may distribute loads to diverse data in order to regulate the preparation incline. Experiments on two separate web-based media ADR datasets, TwiMed-Twitter and CADEC, reveal that our model can alleviate worries about online media datasets. They employed four datasets in the article, including TwiMed-PubMed, which was created from 1000 PubMed phrases. They then used the ADE dataset, which was created from data taken from 2471 sentences. The TwiMed-Twitter dataset followed, with extraction findings from 1000 tweets, only 625 of which were publicly available. Finally, the CADEC dataset, which was derived from medical forum messages, was employed. There were 1248 sentences in total. All of these datasets were chosen because they came from a range of sources, ensuring that the research was diverse and that the best potential results were obtained. The TwiMed-Twitter corpus had the greatest P value of 0.6833, R of 0.7373, and F1 score of 0.6998, while the TwiMed-Twitter corpus had the lowest P value of 0.6748, R of 0.6703, and F1 score of 0.6710 for the dataset CADEC Future examination bearings will include determining the distinction between "Finding" and "ADR" substances, considering language attributes and linguistic data, investigating the distinctions between the two types of elements, and further developing the recognition execution of element types that are excluded from the source area or complex element types. [13]

Because messages are typically short and contexts are sparse, named entity recognition (NER) in social media posts is difficult. Recent research has shown that visual information can improve NER performance by providing more contextual information for texts. The image-level features, on the other hand, neglect the mapping relationships between fine-grained visual objects and textual entities, leading to mistake detection in entities of various types. A modality attention that focuses on image, word, and character level representation has been proposed for multimodal NER (MNER) tasks. Their strategy solely considers the attention spans of text and single visuals. An adaptive co-attention model was also presented, in which text and visual attention are collected simultaneously. An adversarial gated bilinear attention neural network is also proposed in the paper (AGBAN). The model uses adversarial training to translate two separate representations into a single representation by extracting entity-related characteristics from both visual objects and textual. The dataset used in the research is a collection of tweets. The proposed model was tested on a Twitter multimodal social media dataset. The dataset has four different sorts of entities: person, location, organisation, and miscellaneous. It contains 8257 tweets posted by 2116 people. The suggested model is compared to models that just use text data or models that combine image-level visual elements.

The best model was chosen based on its performance on development datasets, and the test dataset performance of the chosen model was provided. In terms of F1 value, the object-AGBAN model outperforms the other models significantly. The model can surpass the BERT-NER in Precision and F1 values thanks to the Gated Bilinear Attention module and Adversarial Learning. Adversarial training can efficiently leverage commonalities across heterogeneous data sources, according to extensive testing. The paper plans to use knowledge-based methods in its multimodal representation in the future to create a more robust and effective NER model.[19]

The subject tackled in this foundation article is entity recognition on social media posts using a multi modal uncertainty aware network to create candidate labels for a social media post for social media with constrained fields in order to label posts that are not as contextual as they can be. With the advancement of deep learning and representation learning, neural network-based multimodal NER approaches have been developed to predict identified entities in social media using both image and text input. Despite their significant improvements over text-based approaches, these systems have two drawbacks: They neglect the mapping relationships between visible objects and named entities, which is the first notable drawback. Another flaw is that past research has ignored the discrepancy in distribution of image and text features. To predict entity labels, they simply concatenate the representations of words and an associated image. The gated bilinear attention network is a potential solution to the difficulty stated above for capturing the interactions of multimodal characteristics visual objects and textual words. It was accomplished using a Bayesian neural network. Some of the earlier works investigate various methods for incorporating visual information from images into NER on social media, with considerable success.

Existing solutions, on the other hand, overlook a regular occurrence on social media: the image does not always match the words uploaded. As a result, the irrelevant photos could inject noise into current models. Two public datasets, Twitter-2015 and Twitter-2017, were used to test the suggested approach. Zhang et al. were the first to create Twitter-2015, which gathered multimodal tweets from 2014 to 2015. Similarly, Lu et al. built Twitter-2017, a database that includes multimodal tweets from 2016 to 2017. Precision, recall, and F1-score are some of the evaluation methods used to assess the data obtained.[22]

In social media engineering, demographics are critical. Yahoo Answers is an online community-based question-answering platform that allows people from all over the world to participate in question-answering forums. As is well known, different age groups have varied interests, needs (e.g., some medications and assisted care), values, incomes, and shopping patterns. It is simple to see how these discrepancies vary over time, such as how personal expenditures change as we become older, in addition to analysing these clusters. Age demographics are crucial for diversifying and improving the dynamicity of cQA platforms when they are integrated into question routing, expert finding, personalization, and customised displays. The purpose of offering a diversity of results is to ignite community members' interest in learning new things by studying new topics. For cQA analysis, none of the prior studies have taken demographic factors like age into account. Some of these were algorithm ranking, question and answer legitimacy, and linguistic approach to questioning. There was only one study on sentiment analysis on cQA platforms that focused on the age demographic briefly. The datasets utilised in the election were fragments of text authored by people aged 10 to 99 years old between 1918 and 2010. The text was separated into sentences after it was collected. CoreNLP was used to eliminate all stop words and words that communicated salutations once they were broken into sentences. They then deleted all numerical entities that had no symbolic meaning, such as mistyped ones. Overall, this approach resulted in 1,141,553 sentences being produced by 697,630 community members. The following is the distribution: Only one age reference was obtained by 69.30 percent of the members, while 17.47 percent obtained two references and 13.23 percent obtained more than two references. The filtered data was then separated into five groups: Matures, Baby Boomers, Generation Xers, Generation Yers, and Generation Zers. Various patterns emerged from the categorisation, such as Gen Zers talking more about sports, education, and politics, whilst Gen Yers leaned more toward physical characteristics of the human body, and Gen Xers asked more questions about childbearing, myspace, and financing. On each piece of text in each category, the following linguistic characteristics were calculated: Bag of words, HPSG parser, Dependency Tree, Predicate Analysis, Acronyms, Explicit Semantic Analysis, Wordne, Rhetorical Structure Theory, Misc, and URLs are some of the terms used. Naive Bayes Classification, Support Vector Machine, Online Learning, Maximum Entropy Models, Fast Text, and Deep Neural Networks were among the machine learning algorithms used. The dataset was separated into three sets: the training set, which contained 60% of the data, and the test and validation sets, which shared the remaining 40%. After testing all of the models, it was discovered that the maximum Entropy model produced the best accuracy, recall precision, and F1 score. The researchers hope to apply multi-view learning to maximise the synergy of many input sources and modalities. Ensembles may be an effective, and in many cases cost-effective, way to integrate the results of many models built on top of various classes of inputs and modalities. [102]

The fundamental goal of this research study is to do fuzzy sentiment analysis in social networks. Machine learning algorithms are used to solve sentiment analysis from social media networks, which is a fairly old issue. It categorises emotions into positive, neutral, and negative categories; however, the documents must first be classified, and the annotation process can only employ the classes or labels. It's impossible to say how many positive, slightly positive, neutral, negative, and somewhat negative comments there are. Previous studies on sentiment analysis have relied mostly on machine learning techniques and the utilisation of lexical resources. In this topic, there are other research that combine the use of ontologies and language patterns. Using words extracted from social networks to generate a Fuzzy Sentiment Dimension that facilitates multidimensional sentiment analysis in social networks is the recommended technique employed in this paper. Several academics [5,8–13] have looked at the difficulty of constructing a multidimensional model from unstructured data such as texts, but none have approached it from a fuzzy perspective. Our contribution is to combine classical dimensions with fuzzy analysis of user opinions in social network writings. Developing an automated document clustering method that considers the sentiments expressed in the texts. Then, using linguistic labels, we assign a sentiment rating to each document automatically. Finally, we construct a hierarchical structure that allows us to investigate attitudes at various levels of detail. Developing an adaptive (totally automated) system for picking linguistic labels and determining their membership functions. We combine many clusters created from the document clustering process that are relatively near to each other using this strategy. We used a consistent approach to define membership functions, taking into account the distribution of documents in the dataset. Defining the storage and query extensions that support the FSD. The first allows us to create structures like cubes, dimensions, hierarchies, and levels to give fuzzy multidimensional user opinion analysis. The second lets you query the FSD with multidimensional model actions like roll-up and drill-down.

Two datasets were taken from Twitter, and the other two were taken from Movie Reviews. The first batch of data consists of 1,600,000 generic tweets with emoticons removed and polarity sorted into six categories. The second dataset contains 25000 tweets from a Twitter account in the United States. The third one comes from the IMDB movie database, while the fourth one may be downloaded from the Kaggle website. Textblob and Vader are used to calculate the sentiment score. The K Means method proved to be the most accurate of all the machine learning algorithms. The text in the IMDB dataset has been categorised as positive or negative, with a sentiment score ranging from [-0.35,0.45]. This study paper's next work will include the integration of the FSD and multidimensional model.[103]

Detecting the stages of evolution of social media events such as Twitter and Sina Weibo allows businesses and governments to intervene before they become unmanageable. The key issue of the foundation paper is anticipating the evolution of social media events. Beginning, Developing, Climax, Descending, and Disappearing are the five stages of a social media event. Previous studies used counting the amount of tweets about the event to determine which stage the process was in. Previous research has mostly focused on event extraction, with little attention paid to the stages of event evolution. The following characteristics distinguish this study from past research: Consider the following characteristics to understand the evolutionary stage of events. Microblogs could be used to detect the longevity of events, according to the study. Second, the bag-of-words paradigm has been proved to be inadequate in expressing events. 500,000 tweets from Twitter and Sina Weibo were used in the analysis. A graph-of-words model is used to represent events. The graph kernel has been shown to be an effective method for calculating graph similarity. This research paper's recommended solution is: By proposing a KPIG graph, we can combine textual, statistical, and connecting information about events. The KPIG graph can provide more detailed information about events than previous keyword or microblog-based models. The foundational research paper outlines a graph kernel-based approach for identifying the stages of an event's evolution using KPIG graphs. Use a shortest-path-based graph kernel in particular to measure the similarity and changes between KPIG graphs. This research work, in particular, presents a novel way for characterising occurrences using a new graph model that combines textual and statistical data. Three innovative techniques are used: (1) to generate a KPIG graph; (2) to quantify the change in textual and statistical content of two sub events; and (3) to determine evolutionary stages. The researchers asked ten students who knew the result of an event to manually annotate each occurrence in each time slice to determine the event's stage. We see a time series curve since the algorithm created acquires a change in the sub event (Si) at any moment I The algorithm's performance is measured by how comparable the time series are. Euclidean distance and Dynamic Time Warping are the two methodologies used to compare time similarities. The suggested unique algorithm definitely beats all other known methods such as SPGK, PMGK, and WLGK, as shown by the pictorial representations of Euclidean distance and Dynamic Time. To evaluate their idea, the researchers propose testing their method on a wider dataset. They also use machine learning to train the best parameters for their system. Researchers also propose to collaborate with the fellow researchers of Facebook and Twitter to expand their database. [104]

The constraints of entity linking models on social media datasets are the focus of the research article. As a result, the researchers chose to create a publicly available entity linking dataset from Reddit, which has 17316 connected entities annotated by three human annotators and categorised into Gold, Silver, and Bronze to demonstrate inter annotator agreement. Entity linking is separated into two subtasks in layman's terms: (1) Mention detection, which identifies entities that need to be linked to the database; (2) Entity Disambiguation, which identifies a database record that matches the entity. Only a little amount of study has been done in this field because it has existed in the shadows of natural language processing. Zero shot Learning, which maps entities to the Wikipedia knowledge base, is one of the previous works. To pick entities for disambiguation, one successful strategy involves utilising Reinforcement Learning and a global sequence perspective. Researchers have also developed entity linking models that combine the procedures of mention detection and entity disambiguation on Twitter Knowledge Base to create an end-to-end entity linking mechanism. BERT is a model that was built to obtain good results on some knowledge bases with only modest entity prediction tasks. The database's information was gathered from a variety of Reddit subreddits. The most popular remark (upvotes - downvotes) on each post was used as the criterion for selection. Because the comments were written in English and were primarily focused on American events, three American annotators were chosen for the annotation task. All of the annotators' annotations were recorded and categorised as Gold, Silver, and Bronze Annotations. All of the annotators agreed on gold annotations, two agreed on silver annotations, and one agreed on bronze annotations. The annotations were tidied up by deleting redirecting links to eliminate the disagreements. Other forms of arguments were also resolved in the end. The second purpose was to assess how well entity linking methods performed on the new dataset. Because the Bronze Annotations were so ambiguous, the annotations were only carried on the Gold and Silver Annotations. Because the Bronze Annotations were so ambiguous, the annotations were only carried on the Gold and Silver Annotations.

The entity disambiguation model was then constructed using six methods: two high-quality baselines and four well-known entity disambiguation models. Prior, deep-ed, muller-nel, wnel, and End to End were the six models created. Two significant conclusions can be derived from these observations. On the Reddit annotation dataset, the simple Query baseline model performs excellently, but as Silver annotations are added to the dataset, performance falls. Second, deep neural network models do not outperform baseline models much on our Reddit dataset. On the pooled dataset, the Query baseline model achieved the greatest F1 score, with all techniques scoring higher than the top neural model. The precision of the neural End-2-End model reflects this: it not only achieves the highest precision of all approaches, but also the lowest level of recall. The researchers' goal is to develop a more comprehensive entity linkage model for social media conversation. [105]

Social media has a significant impact on users. Many social media conversations are confrontational, especially in polarised cultures. Instead of encouraging the formation of users who denigrate others' positions, these environments support the formation of users who do so. News comments are a common form of such interaction on news websites. This is detrimental to the formation of a democratic and deliberative climate, emphasising the need for autonomous systems that can detect polarisation and disagreement early on. The following are the work's main contributions: (1) The researchers propose GENE, a method for building user networks based on the combined depiction of persons, entities, and the users' propensity toward those entities (positive, neutral or negative). (2) Extensive research shows that for the objective of dispute identification, GENE provides a rich representation of user networks conditioned on polarised entities that outperforms both lexicon-based techniques and graph representations that neglect entities and polarisation contexts. (3) In an early controversy detection setting, GENE outperforms other approaches, favouring the early prediction of polarisation in a user network and the description of a scenario for dispute genesis. (4) A fresh news and chat thread dataset has been released. The primary goal of this research is to create a GENE graph. There are three basic phases in GENE. The first step is to develop a user-entity model, which allows us to quantify each user's bias toward the corpus' most prevalent entities. In a second stage, the user-entity model is employed to create a multi-relational graph. Using embeddings derived from the graph, GENE constructs a polarised network of intervening users in a discussion using defined entities. Finally, a third stage analyses the interactions in the generated graph, resulting in the detection of disagreement. To implement these stages, GENE examines a number of data processing modules. The research was based on information gathered from Emol, a Chilean internet news portal. 1 Emol makes the data publicly available in JSON format. The dataset contains 143,340 news items retrieved between April 1, 2016, and April 20, 2019. There are 122,778 comments on this collection of news. (106).

The problem mentioned in the base paper is Domain Adaptation, which occurs when we try to learn a well performing model on a distinct (but similar) target data distribution from a source data distribution. The goal of named entity recognition (NER) is to quickly recognise essential aspects in a document, such as people's names, places, brands, and monetary values. When dealing with enormous datasets, extracting the major entities in a text can help sort unstructured data and uncover relevant information. Although the contrast and combination of distinct entity types among various NER systems is a very hard topic for specialists, there are numerous NER systems in the globe that employ generic entity type classification schemas.

Therefore, in this base paper we suggest an approach named L2AWE (Learning to Adapt with word embeddings) which is aiming at adjusting a NER system focused on a source generic classified schema to a given goal by making use of a rich semantic input text. Such high-level input representation is adopted to motivate the intuition that the word embeddings which will be implicitly captured. Although, the investigation on microblog posts were used in the proposed approach which was mapping entity types from a source to a target classified schema, whereas it can be applied to a variety of different textual formats. It is a machine learning problem of adapting the types of entity mentions from a source to a target classified schema. NER model identifies a set of given entities according to a source schema. The main goal of it is to learn how to map the source type probability distribution to the target one. By jointly considering word embeddings of the entity we get the best adaptation abilities, and the probability distribution over the source entity types as the input space, and by using Support vector Machines as the machine learning classifier. The datasets used to perform an experimental analysis of the proposed approach are the three benchmarks of microblog posts on social media. The two datasets used by the NER and the linking challenges for the #Microposts2015 (Rizzo et al., 2015) and #Microposts2016 (Rizzo, van Erp, Plu, & Troncy, 2016) challenges. The third dataset has been published in the context of the shared task ‘‘Novel and Emerging Entity Recognition’’ at the 3rd Workshop on Noisy User-generated Text (W-NUT) (Derczynski, Nichols, van Erp, & Limsopatham, 2017). The evaluation is completed through different model configurations and by the comparison amongst the aforementioned baselines (Accuracy, Precision, Recall and F-measure have been estimated). Other researchers have proposed the use of machine learning techniques to semi-automatically create semantic mappings between entity types, these approaches have been based on collecting feedback on class-to-class mappings in order to improve ontology alignments but in this paper, we extended the former

investigations by introducing word embeddings for adapting a pre-trained NER system to novel generic classification schemas. The future work highlights the possibilities to investigate some additional and promising models. [88]

Using multiple languages in a similar context is called Code-Switching. Many people mix different languages in text and speech. In Arabic countries Arab people tend to use English words in Arabic. In addition to this by adding Arabic prefix or suffix to English words they use code switching. Most languages like Arabic need language identification also for Intra Word CS. In social media people use informal texts which contain a handful of various types of CS data. The data we get from social media needs to be analysed and investigated for different linguistic tasks. An overview of relevant work concerning collecting CS data by transcribing speech data from social media. Tackling Intra-Word CS data is one of the most crucial tasks of language identification. This problem is solved by tagging corresponding language ID and segmenting mixed words in LID. This creates the first annotated Arabic-English corpus for the CS Intra-Word LID tasks. There are two baseline models presented in this paper, Naïve Bayes and Character BiLSTM for Ar-En text. The Naïve Bayes algorithm is a machine learning classification algorithm based on Bayes theorem of probability that predicts the class of unknown datasets. The main model of this was constructed using segmental recurrent neural networks (SegRNN) and it is of three layers Character Embedding, BiLSTM and time distributed layer. The Ar-En CS data were collected by us and then they annotated the tokens with their corresponding language tag. This is the first annotated Ar-En dataset for Intra Word code switching language identification task. The data was collected first from Twitter using Tweepy API, secondly from Facebook and thirdly from WhatsApp. A new web-based application which helps to annotate data with their corresponding language ID. There is no existing open-source application for the mixed data annotation process. In the application there can be two types of users, the first is the owner or researcher who uploads the unlabelled data and the second is the annotator, who annotates the available data. In the evaluation process the data used was composed of 23,428 tokens for training and 6893 tokens for testing. For future work the same experiment will be done for different pairs of languages. [91]

Embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning. It allows computing semantic and syntactic similarities between linguistic units. It became one of the most popular representations in NLP. The problem discussed in this paper is multi-level embeddings combining representations from different units for processing Arabic social media contents. The representations are pre trained or learned. Outperforming word-level baselines in different NLP tasks such as machine translation, NER and part of speech tagging. The aim of this research is to propose for the first time in Arabic NLP a depth study of various sub-words configurations ranging from character-to-character n-grams for social media text classification. There are many neural architectures proposed through which we can learn different characters, sub-word and word embeddings. A combination of three layers that investigates various forms of composition functions in order to arrive at a final representation of a text. Extrinsic evaluations on these three text tasks can be used to assess the effectiveness of the representations. We test common dialect-agnostic and dialect-specific datasets that include recently utilised shared data tasks. Because different dialects demand varied composition configurations, our methodology is the most productive in dialects. The two phases were generally as follows: 1) segment a word into basic units, and 2) compose the representation of these units to generate the final representation of a word. We used multiple CNN architectures to learn these embeddings while altering the input vector representations in this paper. The data comes from Facebook comments, Twitter reviews, and blogs. The finding reveals multi-level dialect embeddings. We think that this research will be extremely valuable in the development of future multi-level language models for Arabic dialects. [95]

With the widespread use of the internet and its applications, our social networks are bombarded with brief glimpse messages every day, resulting in a big data scale issue that is disseminated in a number of ways and therefore becoming an attractive topic for many scholars. Many of them try to focus on detecting and forming current topics on social media platforms, but this leads to a slew of unsolved questions. In order to keep appropriate methodology, research concerns generate a problem of maintaining distinct challenges of adverse languages. We employ transformers in conjunction with an incremental community detection logarithm in this study. Transformer establishes semantic relationships between distinct words in diverse settings. Named Entity Recognition (NER) is a technique for extracting and tuning themes from multimodal data. For topic detection and tracking from Twitter, a variety of strategies have been proposed. Memory graph, a unique technology that employs cognitive memorization in the human brain, is the result of this combination. We can make this system function with any stream size and any computer by using hyper-parameters like the rate of forgetting. The modularity of our work makes it more useful in real-world and corporate settings when working with enormous social data. The three foundations of our proposed system are transformer semantic similarity, word community identification, and multimodal named entity recognizer (TopicBERT). Data from social media streams such as Twitter is fed into MongoDB, which feeds multimodal data into other system features. In this case, MongoDB acts as a cache for the entire system in case of long delays or system failures.

Delays occur as a result of the different sections' complexity and speed metrics. The database, on the other hand, is a memory graph-based storage facility for word graphs. We use two criteria to assess our performance: keyword precision and subject recall. Keyword precision is defined as the number of correctly recognised keywords divided by the number of ground-truth keywords. The results show that our proposed technique outperforms other current best practises.[100]

These days, data modelling from a spatial-temporal perspective is becoming more prevalent. In order to comprehend spatiotemporal patterns, a wide range of applications, such as social network data, must be examined. However, analysing or classifying these patterns as the conventional context in various types of event data takes time because they are often sophisticated and difficult to categorise as the conventional context. In order to investigate the traffic viral within the text from the aspect of impressive bad implications, we should spatially-temporally locate the event and geographical locations and provide a semantically interpreting of what transpired. We give an evaluation of the best models and methodologies utilised for social media data processing in order to formalise a new theory of action and time. The goal of this research is to establish a baseline of data from which researchers can deduce the recurrence of events in texts, as well as the characters who appear in them, and their link to time and place. Assembling and organising a social media collection, as well as examining its metadata. In order to decode in texts, the occurrence of events, coupled with their related characters, and their relationship with time and space, we explain a machine learning strategy to develop a novel system. In general, tiny documents such as tweets lack sufficient information to reveal their semantic context. Novels and other longer documents, on the other hand, have far too much diversity. Several tests were carried out in order to improve RTCoViD. Accuracy is a typical metric for assessing the performance of suggested models. As a result, for this investigation, we employed the Multinomial Naive Bayes classifier. In fact, the total number of correct classifications divided by the total number of classifications made at a given point in time is the formula. As shown, the community has responded to the coronavirus pandemic by developing datasets that will aid in the development of novel cures or forecasting models that will help us better predict behaviour and even warn us about future disasters. Previous research has shown that spatio-temporal-semantic models have a lot of potential for accurately forecasting infectious diseases in time and space. This discovery could aid in the development of smart vaccine technology by providing new ways of detecting potential viruses and their consequences on human health. Governments can be notified based on our findings so that appropriate constraints can be imposed swiftly to help stop the pandemic from spreading.[101]

V. EXPERIMENTS

The suggested model's efficiency is validated through comparative experiments on the datasets. The qualitative analysis is then presented to demonstrate the attention modules' interpretability. Recall, precision, and F1 score were the evaluation measures employed in the trials. Only if both the type and the boundaries match ground truth is a named entity's recognition regarded as correct. When compared to other similar approaches, the suggested model has the best F1-scores in real-world datasets. Experiments on these datasets show improved adaptability and anti-interference capacity when applied to real-world data. The comparative results show that the model can adapt to real-life scenarios with great results. The entities' F1 scores. F1-scores, namely - Person, Organisation, Location, Miscellaneous and Overall are calculated.[28] The performance of different models on the different datasets has been evaluated using metrics such as F-measure, accuracy, Precision, recall and F1-score. The suggested method outperformed the F1-scores obtained by a number of earlier studies, including (Ma & Hovy, 2016), (Zhang et al., 2018), and others. As a result of these tests, it is possible to infer that the hand-crafted features performed admirably. The entities' F1-scores, which include Person, Organisation, Location, Miscellaneous, and Overall, are calculated. [34] A 10-fold cross-validation is utilised to calculate Recall, Precision, and F-Score for each of the examined algorithms in order to provide results in the setting that are comparable to relevant works. The suggested method's results were compared to the results of three text classification models, LSTM-vote, GRU-2, and SAtt-BLSTM convNet, as well as three rumour detection models, RPDNN, GLO-PGNN, and STS-NN. A 10-fold cross-validation is utilised to calculate Recall, Precision, and F-Score for each of the examined algorithms in order to provide results in the setting that are comparable to relevant works.[40]

Precision, recall, f1-score, and runtime are the evaluation measures used to compare the proposed system's performance to that of the baseline system and state-of-the-art techniques. The metrics Mean Absolute Error (MAE) and Mean Squared Error (MSE) can be used to assess the performance of face counting models (MSE). [46] The rates of agreement for the several characterization algorithms were calculated using conventional measures. In light of the results acquired from the various methodologies used, the methodology employed has proven to be effective. F-score, Precision, and Recall are the measures used. For testing, F-measure-micro, Recall-micro, Precision-micro, and F-measure-macro were employed. L2AWE can attain a level of accuracy greater than 80% while maintaining a constant standard deviation. [88] For the purposes of evaluation, accuracy and macro-averaged F1-scores have also been produced for some situations. [95]

The approach's multimodal nature improves performance in the presence of noise and allows it to do recognition without an image. The model's great performance makes it ideal for the work at hand. [100] F1 scores, as well as P and R values, were used to assess the model. The model worked effectively for both types of datasets, with the TwiMed-Twitter corpus having the highest P value of 0.6833, R of 0.7373, and F1 score of 0.6998, and the CADEC dataset having P of 0.6748, R of 0.6703, and F1 score of 0.6710. [13] The object-AGBAN model outperforms the other models in every way.[22]

TABLE I
COMPARISON OF BASE PAPERS

S. No.	Base Paper	Approach	Evaluation metrics	Datasets	Results obtained
1	[28]	HSN (Hierarchical Self-adaptation Network) model	Recall Precision F1 Score	NER multimodal Twitter and NER dataset	Precision 74.92% Recall 73.45% F1 74.18%
2	[34]	Hybrid-DNL+CNN+BiLSTM+CRF	f-measure accuracy precision recall f1-score	Multimodal dataset	Precision 71.43% Recall 72.63% F1 72.07%
3	[13]	Adversarial transfer model with bilinear attention	Recall Precision F1 Score	ADR datasets TwiMed-Twitter CADEC TwiMed-PubMed	Precision 68.33% Recall 73.73% F1 score 69.98%
4	[19]	Uncertainty aware multimodal NER	Recall Precision F1 Score	Twitter-2015 Twitter-2017	Precision 73.02% Recall 74.75% F1 score 73.87%
5	[22]	Adversarial Gated Bilinear Attention Network	Recall Precision F1 Score	A multimodal social media dataset from Twitter containing 4 entities: Person, Location, Organization, Misc.	Precision 75.42% Recall 72.39% F1 score 73.25%
6	[1]	Semi-supervised neural network.	Recall Precision F1 Score	Manually annotated datasets in Arabic were used, for which, raw data had been extracted from Twitter.	F1 68.00%
7	[7]	Adversarial transfer learning model	Recall Precision F1 Score	ADE TwiMed-PubMed TwiMed-Twitter3 dataset	Precision 68.02% Recall 73.84% F1 score 68.58%
8	[40]	ESODE	Recall Precision	Datasets from Twitter and Weibo	Precision 92.31% Recall 92.44%

			F1 Score		F1 92.37%
9	[46]	biLSTM+Tf+Lou	Recall Precision F1 Score Mean Absolute Error Mean Squared Error	CoNNL2003 dataset EVALITA 2009 track dataset WIDER FACE dataset	F1 score 90.59%/8 secs (English dataset) F1 score 78.75%/14 secs (Italian dataset)
10	[52]	NLP-ML-GM-ONER	Recall Precision F1 Score	Tweets from Twitter API	F score individuals 84.00% patients 89.00% user gender 86.00% user location 95.00%
11	[102]	Sentiment analysis based on demographics	Accuracy, Precision, Recall, F-1 score	A self-collected dataset containing posts from Yahoo	Highest Precision and accuracy for maximum Entropy model Accuracy: 78.84 Precision: 74.03
12	[103]	Using Fuzzy Sentiment analysis techniques in Social media networking platforms.	Accuracy Precision Recall F-1 score	2 datasets from Twitter and 2 datasets from Movie Reviews.	Accuracy- 81.76% Precision - 82.03% Recall- 74.89% F-1 score
13	[104]	Graph-based approach for detecting events on social media	Accuracy Precision Recall F-1 score DynamicTime Warping	Dataset consists of 500,000 tweets recorded from Twitter and Seina Weiboo.	Accuracy - 81.06% Precision - 80.87%
14	[105]	GENE-based approach for detecting controversy and polarised environment	Accuracy Precision Recall F-1 score	Emol, a Chilean internet news portal. The dataset comprises a total of 143,340 news retrieved from April 1, 2016, to April 20, 2019.	Accuracy - 79.8% Precision - 82.4% Recall - 70%
15	[106]	Entity linking using Entity Disambiguation models.	Micro Precision and Micro F-1 scores	Reddit Entity Linking Dataset	Micro- Precision - 91.2% Micro -Recall- 92.4%
16	[88]	Adapt with Word Embeddings	*Precision-micro,*	Three microblog post	*Precision*- 89%

			Recall-micro, *F-measure-micro* and *F-measure-macro*	benchmark datasets, two provided by the Named Entity Recognition and Linking Challenges.	*Recall*- 89% *F-measure* micro-89% *F-measure macro*-79%
17	[91]	Language Identification of Intra-Word Code-Switching.	Accuracy F-1 score	Twitter, Facebook, and WhatsApp data, Tokens in Arabic, English, Arabizi, and Engari are included in the corpus.	Accuracy- 89% F-1 score- 94.84%
18	[95]	Learning sub word embeddings.	Accuracy Macro-averaged F1-scores	Twifilemo Emotion Classification (E-C) task data24	Accuracy -89% Macro-averaged-86%F1-scores-88%
19	[100]	Tokenization and stemming of words	Accuracy Recall Precision	Datasets part of social sensor project	Accuracy-90% Recall-84% Precision-85.5%
20	[101]	Correlate syntax with semantics using Treeops	Accuracy	Dataset from John Hopkins University (2019-nCoV Data Repository)	Accuracy-90%

VI. CONCLUSION

Social media analysis using Named Entity Recognition is a broad field with a lot of unexplored territory. The results of the poll revealed a wide range of information. The fact that prediction errors are more likely to occur around generational divides is fascinating. When we compare the findings of the end-2-end model to the NER and entity disambiguation baselines, we see that the mention detection sub-task is to blame for the majority of the errors. That is, it appears to be rather difficult to determine which word or words constitute an entity addressed in social media writing. However, once these events have been identified, linking them to the appropriate entity becomes much easier. Another important factor to consider is demographics as they play a vital role in sentiment analysis, evolutionary process of social media events etc. A variety of machine learning methods were used like Random Forest, Decision Tree, Support Vector Machine, Multi-Layer Perceptron and some deep learning techniques like LSTM, Bidirectional LSTM etc. Different models proved to be useful in different applications. Some novel methods such as GENE (A graph-based detection on social media analysis) also can be used for further research work in this domain. Some of the areas in which Named Entity Recognition in social media played a vital role, were found to be sentiment analysis, rumour detection and prevention, biomedical analysis in dietary concerns (e.g., gluten-related tweets and posts), visualisation of the spread of Covid-19 in certain areas, topic detection, controversy detection and event detection. Precision, recall, f1-score, and runtime are the major evaluation measures used to compare the proposed system's performance to baseline systems and state-of-the-art methodologies. The performance of the models utilised in specific areas was further assessed using the metrics Mean Absolute Error (MAE) and Mean Squared Error (MSE) (MSE). When compared to previous approaches proposed with comparable intents and objectives, most of these studies achieve adequate performance with greater accuracy and f1-scores.

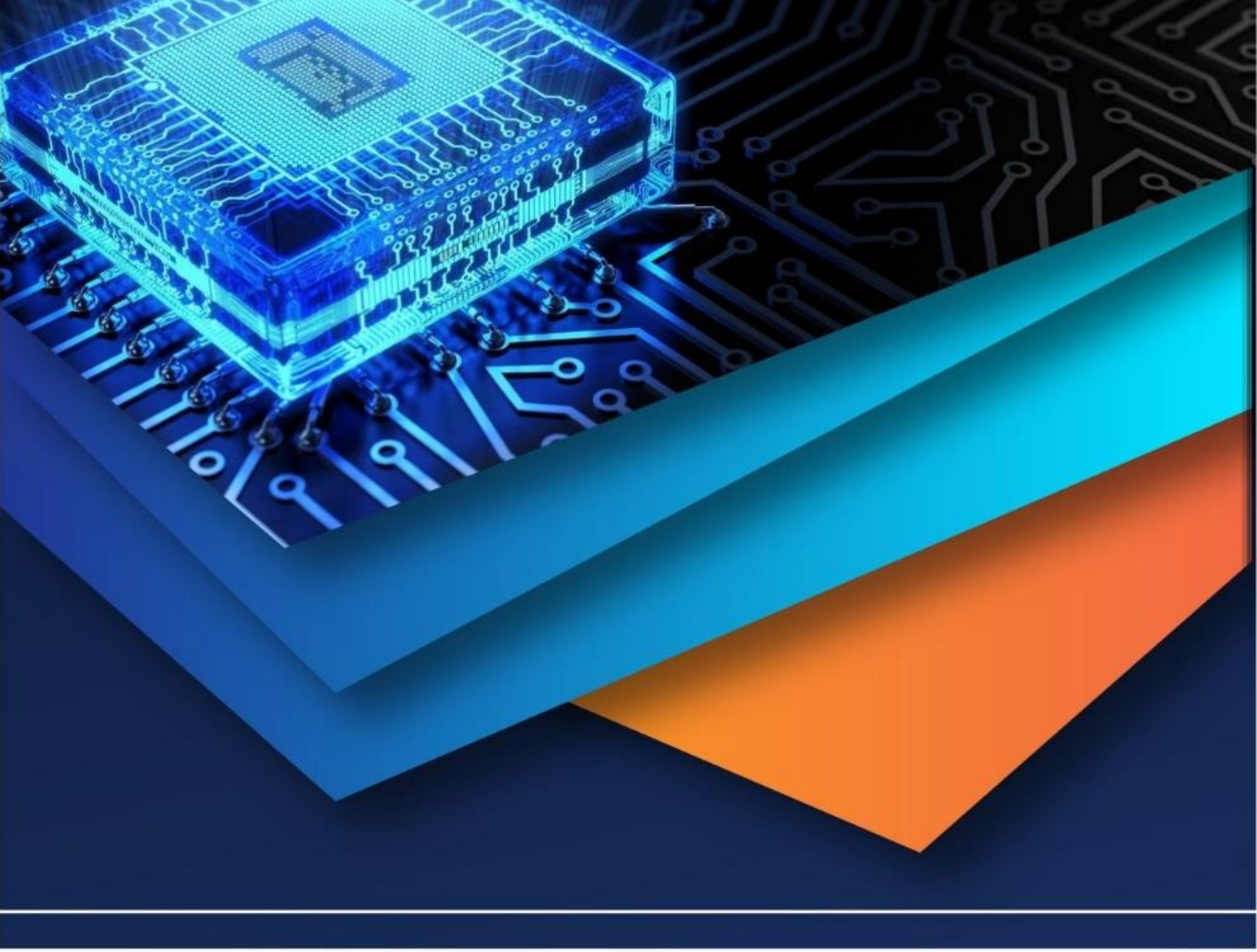
REFERENCES

[1] Al-Laith, A., & Shahbaz, M. (2021). Tracking sentiment towards news entities from Arabic news on social media. Future Generation Computer Systems, 118, 467-484.
[2] Bogdan Batrinca, Philip C. Treleaven, Social media analytics: A survey of techniques, tools and platforms, AI Soc. 30 (1) (2015) 89–116.
[3] Sotiris B. Kotsiantis, Ioannis D. Zaharakis, Panayiotis E. Pintelas, Machine learning: A review of classification and combining techniques, Artif. Intell. Rev. 26 (3) (2006) 159–190.

[4] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, Asa: A framework for arabic sentiment analysis, J. Inf. Sci. 46 (4) (2020) 544–559
[5] Brahim Ait Ben Ali, Soukaina Mihi, Ismail El Bazi, Nabil Laachfoubi, A recent survey of arabic named entity recognition on social media, Rev. Intell. Artif. 34 (2) (2020) 125–135.
[6] Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (4) (2014) 1093–1113.
[7] Zhang, T., Lin, H., Ren, Y., Yang, Z., Wang, J., Duan, X., & Xu, B. (2021). Identifying adverse drug reaction entities from social media with adversarial transfer learning model. Neurocomputing, 453, 254-262.
[8] Ma X, Hovy E. End-to-end sequence labelling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354, 2016.
[9] B.Y. Lin, W. Lu, in: Neural adaptation layers for cross-domain named entity recognition, Association for Computational Linguistics, 2018, pp. 2012–2022.
[10] Y. Lin, S. Yang, V. Stoyanov, et al., A multi-lingual multi-task architecture for low-resource sequence labelling[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1, Long Papers) (2018:) 799–809.
[11] Yang Y S, Zhang M, Chen W, et al. Adversarial learning for chinese NER from crowd annotations[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[12] J.T. Zhou, H. Zhang, D. Jin, et al., Dual Adversarial Neural Transfer for LowResource Named Entity Recognition[C]//Proceedings of the 57th Annual Meeting of the Association for, Computat. Linguist. (2019:) 3461–3471.
[13] Zhang, T., Lin, H., Ren, Y., Yang, Z., Wang, J., Zhang, S., ... & Duan, X. (2021). Adversarial transfer network with bilinear attention for the detection of adverse drug reactions from social media. Applied Soft Computing, 106, 107358
[14] Y. Chen, C. Zhou, T. Li, et al., Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training, J. Biomed. Inform. 96 (2019) 103252.
[15] N. Kang, B. Singh, C. Bui, et al., Knowledge-based extraction of adverse drug events from biomedical text, BMC Bioinformatics 15 (1) (2014) 64.
[16] E. Aramaki, Y. Miura, M. Tonoike, et al., Extraction of adverse drug effects from clinical records, MedInfo 160 (2010) 739–743.
[17] J.T. Zhou, H. Zhang, D. Jin, et al., Dual adversarial neural transfer for low-resource named entity recognition, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3461–3471.
[18] K. Bousmalis, G. Trigeorgis, N. Silberman, et al., Domain separation networks, 2016, arXiv preprint arXiv:1608.06019.
[19] Liu, L., Wang, M., Zhang, M., Qing, L., & He, X. (2021). UAMNer: uncertainty-aware multimodal named entity recognition in social media posts. Applied Intelligence, 1-17.
[20] Xiong W, Yu M, Chang S, Guo X, Wang WY (2019) Improving question answering over incomplete KBs with knowledgeware reader. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp 4258–4264. Association for Computational Linguistics
[21] Zhang Q, Fu J, Liu X, Huang X (2018) Adaptive co-attention network for named entity recognition in tweets. In: AAAI, pp 5674–5681
[22] C. Zheng, Z. Wu, T. Wang, Y. Cai and Q. Li, "Object-Aware Multimodal Named Entity Recognition in Social Media Posts With Adversarial Learning," in IEEE Transactions on Multimedia, vol. 23, pp. 2520-2532, 2021, doi: 10.1109/TMM.2020.3013398.
[23] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, “Joint extraction of entities and relations based on a novel tagging scheme,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1227–1236.
[24] J. Wang, L. Yu, W. Zhang, Y. Gong, Y. Xu, B. Wang, P. Zhang, and D. Zhang, “Irgan: A minimax game for unifying generative and discriminative information retrieval models,” in Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, 2017, pp. 515–524.
[25] S. Moon, L. Neves, and V. Carvalho, “Multimodal named entity recognition for short social media posts,” arXiv preprint arXiv:1802.07862, 2018.
[26] A. Ritter, O. Etzioni, and S. Clark, “Open domain event extraction from twitter,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 1104–1112.
[27] T. Sakaki, M. Okazaki, and Y. Matsuo, “Tweet analysis for real-time event detection and earthquake reporting system development,” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 919–931, 2012.
[28] Tian, Y., Sun, X., Yu, H., Li, Y., & Fu, K. (2021). Hierarchical self-adaptation network for multimodal named entity recognition in social media. Neurocomputing, 439, 12-21.
[29] Q. Zhang, J. Fu, X. Liu, X. Huang, Adaptive co-attention network for named entity recognition in tweets, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[30] D. Lu, L. Neves, V. Carvalho, N. Zhang, H. Ji, Visual attention model for name tagging in multimodal social media, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1990–1999.
[31] O. Arshad, I. Gallo, S. Nawaz, A. Calefati, Aiding intra-text representations with visual context for multimodal named entity recognition, arXiv preprint arXiv:1904.01356.
[32] R. Cadène, H. Ben-younes, M. Cord, N. Thome, MUREL: multimodal relational reasoning for visual question answering, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/ IEEE, 2019, pp. 1989–1998. doi:10.1109/CVPR.2019.00209.
[33] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.
[34] Suman, C., Reddy, S. M., Saha, S., & Bhattacharyya, P. (2021). Why pay more? A simple and efficient named entity recognition system for tweets. Expert Systems with Applications, 167, 114101.
[35] Torisawa, K. et al. (2007). Exploiting wikipedia as external knowledge for named entity recognition. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) (pp. 698–707).
[36] Chieu, H. L., & Ng, H. T. (2002). Named entity recognition: a maximum entropy approach using global information. In Proceedings of the 19th international conference on Computational linguistics-Volume 1 (pp. 1–7). Association for Computational Linguistics.
[37] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.
[38] Passos, A., Kumar, V., & McCallum, A. (2014). Lexicon infused phrase embeddings for named entity resolution. arXiv preprint arXiv:1404.5367.

[39] Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. Transactions of the Association for Computational Linguistics, 4 , 357–370.
[40] Ma, T., Zhou, H., Tian, Y., & Al-Nabhan, N. (2021). A novel rumour detection algorithm based on entity recognition, sentence reconfiguration, and ordinary differential equation network. Neurocomputing, 447, 224-234.
[41] S.A. Alkhudair, S.H. Ding, B.C. Fung, J. Liu, Detecting breaking news rumours of emerging topics in social media, Inf. Process. Manage. 57 (2) (2020) 102018.
[42] J.P. Singh, A. Kumar, N.P. Rana, Y.K. Dwivedi, Attention-based lstm network for rumour veracity estimation of tweets, Inf. Syst. Front. (2020) 1–16.
[43] J. Gao, S. Han, X. Song, F. Ciravegna, Rp-dnn: a tweet level propagation context based deep neural networks for early rumour detection in social media, in: LREC 2020 Proceedings: The International Conference on Language Resources and Evaluation, European Language Resources Association, 2020.
[44] F. Xu, V.S. Sheng, M. Wang, Near real-time topic-driven rumor detection in source microblogs, Knowledge-Based Syst. 207 (2020) 106391.
[45] Z. Wang, Y. Guo, Rumour events detection enhanced by encoding sentimental information into time series division and word representations, Neurocomputing 397 (2020) 224–243.
[46] Andreadis, S., Antzoulatos, G., Mavropoulos, T., Giannakeris, P., Tzionis, G., Pantelidis, N., ... & Kompatsiaris, I. (2021). A social media analytics platform visualising the spread of COVID-19 in Italy via exploitation of automatically geotagged tweets. Online Social Networks and Media, 23, 100134.
[47] R. Kouzy, J. Abi Jaoude, A. Kraitem, M.B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E.W. Akl, K. Baddour, Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter, Cureus 12 (3) (2020).
[48] J.S. Brennen, F.M. Simon, P.N. Howard, R.K. Nielsen, Types, sources, and claims of Covid-19 misinformation, Reuters Institute, 2020.
[49] K.-C. Yang, C. Torres-Lugo, F. Menczer, Prevalence of low-credibility information on twitter during the covid-19 outbreak, 2020, arXiv preprint arXiv:2004.14484.
[50] W. Ahmed, J. Vidal-Alaball, J. Downing, F.L. Seguí, COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data, J. Med. Internet Res. 22 (5) (2020) e19458.
[51] C. Li, L.J. Chen, X. Chen, M. Zhang, C.P. Pang, H. Chen, Retrospective analysis of the possibility of predicting the COVID-19 outbreak from internet searches and social media data, China, 2020, Eurosurveillance 25 (10) (2020) 2000199.
[52] Pérez-Pérez, M., Igrejas, G., Fdez-Riverola, F., & Lourenço, A. (2021). A framework to extract biomedical knowledge from gluten-related tweets: The case of dietary concerns in digital era. Artificial Intelligence in Medicine, 118, 102131.
[53] Masmoudi A, Barhamgi M, Faci N, Saoud Z, Belhajjame K, Benslimane D, et al. An ontology-based approach for mining radicalization indicators from online messages. In: Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA. vol. 2018- May. Institute of Electrical and Electronics Engineers Inc; 2018. p. 609–16.
[54] Beguerisse-Díaz M, McLennan AK, Gardu˜no-Hern´andez G, Barahona M, Ulijaszek SJ. The 'who' and 'what' of #diabetes on Twitter. Digit Health 2017;3.
[55] Park A, Conway M. Tracking health related discussions on Reddit for public health applications. In: AMIA. Annu Symp Proceedings AMIA Symp 2017; 2017. p. 1362–71.
[56] Zheng X, Han J, Sun A. A survey of location prediction on Twitter. IEEE Trans Knowl Data Eng 2018;30:1652–71.
[57] P´erez-P´erez M, P´erez-Rodríguez G, Fdez-Riverola F, Lourenço A. Using twitter to understand the human bowel disease community: exploratory analysis of key topics. J Med Internet Res 2019:21.
[58] Momtazi, S. (2018). Unsupervised Latent Dirichlet Allocation for supervised question classification. Information Processing & Management, 54(3), 380-393.
[59] Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013, May). No country for old members: User lifecycle and linguistic change in online communities. In Proceedings of the 22nd international conference on World Wide Web (pp. 307-318).
[60] Lyu, S., Ouyang, W., Wang, Y., Shen, H., & Cheng, X. (2019, May). What we vote for? Answer selection from user expertise view in community question answering. In The World Wide Web Conference (pp. 1198-1209).
[61] Maity, S. K., Kharb, A., & Mukherjee, A. (2018). Analyzing the linguistic structure of question texts to characterize answerability in quora. IEEE Transactions on Computational Social Systems, 5(3), 816-828
[62] Sun, Z., Sun, Y., Chang, X., Wang, Q., Yan, X., Pan, Z., & Li, Z. P. (2020). Community detection based on the Matthew effect. Knowledge-Based Systems, 205, 106256.
[63] A. Figueroa, Male or female: What traits characterize questions prompted by each gender in community question answering? Expert Syst. Appl. 90 (2017) 405–413
[64] A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E. Seligman, et al., Personality, gender, and age in the language of social media: The open-vocabulary approach, PLoS One 8 (9) (2013) e73791
[65] Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W. Daelemans, Overview of the 2nd author profiling task at pan 2014, in: CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014, pp. 1–30
[66] Kucuktunc, B.B. Cambazoglu, I. Weber, H. Ferhatosmanoglu, A largescale sentiment analysis for yahoo! answers, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, ACM, New York, NY, USA, 2012, pp. 633–642
[67] Dragoni, M., & Petrucci, G. (2018). A fuzzy-based strategy for multi-domain sentiment analysis. International Journal of Approximate Reasoning, 93, 59-73.
[68] Pérez, J. M., Berlanga, R., Aramburu, M. J., & Pedersen, T. B. (2008, October). Towards a data warehouse contextualized with web opinions. In 2008 IEEE International Conference on e-Business Engineering (pp. 697-702). IEEE.
[69] Moalla, I., Nabli, A., Bouzguenda, L., & Hammami, M. (2016, November). Data warehouse design from social media for opinion analysis: The case of Facebook and Twitter. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.
[70] Walha, A., Ghozzi, F., & Gargouri, F. (2016, November). A Lexicon approach to multidimensional analysis of tweets opinion. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.
[71] Moalla, I., Nabli, A., Bouzguenda, L., & Hammami, M. (2016, November). Data warehouse design from social media for opinion analysis: The case of Facebook and Twitter. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.
[72] Moalla, I., Nabli, A., Bouzguenda, L., & Hammami, M. (2016, November). Data warehouse design from social media for opinion analysis: The case of Facebook and Twitter. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.

[73] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).
[74] C. Deng, Z. Zhao, Y. Wang, Z. Zhang, Z. Feng, GraphZoom: A multilevel spectral approach for accurate and scalable graph embedding, in: ICLR'2020, 2020
[75] Jin, P., Cui, T., Wang, Q., & Jensen, C. S. (2016, April). Effective similarity search on indoor moving-object trajectories. In International conference on database systems for advanced applications (pp. 181-197). Springer, Cham.
[76] . Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'04, 2004, pp. 404–411.
[77] Wang, C., Song, Y., Li, H., Zhang, M., & Han, J. (2015, November). Knowsim: A document similarity measure on structured heterogeneous information networks. In 2015 IEEE International Conference on Data Mining (pp. 1015-1020). IEEE.
[78] Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., ... & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. Information Processing & Management, 51(2), 32-49.
[79] Enes, K. B., Brum, P. P. V., Cunha, T. O., Murai, F., da Silva, A. P. C., & Pappa, G. L. (2018, December). Reddit weight loss communities: do they have what it takes for effective health interventions?. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 508-513). IEEE.
[80] Dutta, S., Das, D., & Chakraborty, T. (2020). Changing views: Persuasion modeling and argument extraction from online discussions. Information Processing & Management, 57(2), 102085.
[81] Witten, I. H., & Milne, D. N. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links.
[82] Gaumont, N., Panahi, M., & Chavalarias, D. (2018). Reconstruction of the socio-semantic dynamics of political activist Twitter networks—Method and application to the 2017 French presidential election. PloS one, 13(9), e0201879.
[83] Ruan, Y., Fuhry, D., & Parthasarathy, S. (2013, May). Efficient community detection in large networks using content and links. In Proceedings of the 22nd international conference on World Wide Web (pp. 1089-1098).
[84] Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.
[85] Mejova, Y., Zhang, A. X., Diakopoulos, N., & Castillo, C. (2014). Controversy and sentiment in online news. arXiv preprint arXiv:1409.8152.
[86] A. Matakos and A. Gionis, "Tell me Something My Friends do not Know: Diversity Maximization in Social Networks," 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 327-336, doi: 10.1109/ICDM.2018.00048.
[87] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
[88] Nozza, D., Manchanda, P., Fersini, E., Palmonari, M., & Messina, E. (2021). LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems. Information Processing & Management, 58(3), 102537.
[89] Bhatia, P., Celikkaya, B., Khalilia, M., & Senthivel, S. (2019, December). Comprehend medical: a named entity recognition and relationship extraction web service. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 1844-1851). IEEE.
[90] Bommasani, R., Davis, K., & Cardie, C. (2020, July). Interpreting pretrained contextualized representations via reductions to static embeddings. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4758-4781).
[91] Sabty, C., Mesabah, I., Çetinoğlu, Ö., & Abdennadher, S. (2021). Language Identification of Intra-Word Code-Switching for Arabic–English. Array, 12, 100104.
[92] Newman, D. (2002). The phonetic status of Arabic within the world's languages: the uniqueness of the lughat al-daad. Antwerp papers in linguistics., 100, 65-75.
[93] Egger, P. H., & Toubal, F. (2016). Common spoken languages and international trade. In The Palgrave handbook of economics and language (pp. 263-289). Palgrave Macmillan, London.
[94] Bullock, B. E., & Toribio, A. J. E. (2009). The Cambridge handbook of linguistic code-switching. Cambridge University Press.
[95] Moudjari, L., Benamara, F., & Akli-Astouati, K. (2021). Multi-level embeddings for processing Arabic social media contents. Computer Speech & Language
[96] Agarwal, O., Durupınar, F., Badler, N. I., & Nenkova, A. (2019). Word embeddings (also) encode human personality stereotypes. In Proceedings of the Joint Conference on Lexical and Computational Semantics (SEM 2019) (pp. 205–211)
[97] Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy. Transactions of the Association of Computational Linguistics, 6, 483–495
[98] Bagheri, E., Ensan, F., & Al-Obeidat, F. N. (2018). Neural word and entity embeddings for ad hoc retrieval. Information Processing & Management, 54, 657–673
[99] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146
[100] Asgari-Chenaghlu, M., Feizi-Derakhshi, M.-R., farzinvash, L., Balafar, M.-A., & Motamed, C. (2021). TopicBERT: A cognitive approach for topic detection from multimodal post stream using BERT and memory–graph. Chaos, Solitons & Fractals
[101] Şerban Boghiu, Daniela Gîfu,A Spatial-Temporal Model for Event Detection in Social Media,Procedia Computer Science,Volume 176,2020,Pages 541-550
[102] Figueroa, A., & Timilsina, M. (2021). What identifies different age cohorts in Yahoo! Answers?. Knowledge-Based Systems, 228, 107278.
[103] Gutiérrez-Batista, K., Vila, M. A., & Martin-Bautista, M. J. (2021). Building a fuzzy sentiment dimension for multidimensional analysis in social networks. Applied Soft Computing, 108, 107390.
[104] Mu, L., Jin, P., Zhao, J., & Chen, E. (2021). Detecting evolutionary stages of events on social media: A graph-kernel-based approach. Future Generation Computer Systems, 123, 219-232.
[105] Mendoza, M., Parra, D., & Soto, Á. (2020). GENE: graph generation conditioned on named entities for polarity and controversy detection in social media. Information Processing & Management, 57(6), 102366.
[106] Botzer, N., Ding, Y., & Weninger, T. (2021). Reddit entity linking dataset. Information Processing & Management, 58(3), 102479.
[107] K. Balog, Entity-Oriented Search. Springer, 2018.
[108] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvist. Investig., vol. 30, no. 1, pp. 3–26, 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY