Visualization of Arabic Entities in Online Social Media using Machine Learning

Khowla Mohammed Alyamani[1], Abdul Khader Jilani Saudagar[2]
Information Systems Department, College of Computer and Information Sciences
Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

***Abstract*—In recent years, the use of social media and the amount of exchangeable data have increased considerably. The increase in exchangeable data makes data mining, analysis and visualization of relevant information a challenging task. This research work assesses, categorizes, and analyzes Arabic entities on social media selected by users at certain time intervals. To accomplish this aim, the authors built a highly efficient classification model to classify entities according to three categories: person, location, and organization. The developed model captures an entity and specific time, collects all the posts on tweeter that refer to the entity at this specific time, and then classifies, visualize the entity through three methods. It first starts with classifying the entity through a corpus model that depends on customized corpus. If the entity is not classified through that model, it will be send to an indicators model which uses the pre-indicators or post-indicators for classing. Finally, the entity is passed to a gazetteer model which searches for the entity in three gazetteers (person, location, and organization), and accordingly determines the number of times the entity reference is repeated. This work allows scholars and researchers in different fields to visualize the frequency of entities referenced by a community. It also compares how references to entities change over time. The experimental results show that accuracy of the developed model in classifying the tweets is nearly 90%.**

***Keywords*—Machine learning; classification; visualization; Arabic entities; social media**

I. Introduction

People use social media platforms such as Twitter, Instagram, Snapchat, and Facebook to share details about their daily activities, review products, share comments with their online and offline friends, discuss events, and advertise products. Social media has caused many changes: it has generated new business opportunities, changed advertising methods, provided opportunities to earn a living online, made it possible to make virtual friends, and has changed how news spreads. Web 2.0 allows Internet users to communicate and share information 24/7, with one of its most important facets being social media. Social media platforms include forum, micro blogging, social bookmarking, social administration, and wiki applications and websites. Twitter, started in March 2006 and is considered as one of the well-known social media platforms. It provides services of online news and social networking to its users, with users posting and interacting through messages known as tweets. One tweet is limited to 280 characters, and users can include tags, or hash tags, to label or categorize their messages according to contextually appropriate tags for every named entity in a text.

When it comes to social media on Arabic platforms using Arabic language which contains more than 12 million words, is one of the most widely spoken languages in the world, with more than 450 million people speaking Arabic as a native language. Moreover, 60 countries use Arabic as the official language or second language, and more than 60 Arabic dialects exist in the Arab countries.

Researchers in the past have identified three types of Arabic language [8, 12, 16]:

1) Classical Arabic (CA) is the formal version that has been used incessantly for more than 1,500 years. It is the language of the Holy Quran and most Arabic religious books.

2) Modern Standard Arabic (MSA) is the Arabic language of magazines, and education. MSA mirrors the needs of contemporary expression, and CA reflects the needs of older styles. Most Arabic name entity recognition (ANER) systems support MSA.

3) Colloquial Arabic dialects are the language used by Arabs in their informal everyday communications, and dialects differ from one region to another.

One of the many challenges in Arabic name entity tagging stems from the prevalence of homographs. For example, أشرف means Ashraf (name) or he supervised (verb). Some homographs have different meanings depending on their vowelization, which refers to the practice of supplying the vowels in a word, that usually are not written. For example, عقد can mean different meaning under the vowelization as shown in Table I.

TABLE I. EXAMPLE OF A DIACRITICIZED ARABIC WORD REPRESENTING 5 VOWELIZATION

Word in Arabic	Transliteration	Meaning
عِقِد	Egad	Necklace
عَقَد	Agad	Contract
عَقَد	Eaqad	Decade
عُقِد	Oqead	Convention
عُقَد	Oqad	Knot

One Arabic word may contain one or more prefixes and one or more suffixes in different combinations. For example, “و” means and is pronounced “wa”, “ل” means for and is pronounced “le”. Sometimes و or ل or both are added at the beginning of the word, while “ه”, pronounced as “ha”, is added to the end of the word, for example, ولكتابه (wa le ktabieh) means and for his book. Another problem is that a non-Arabic word can be written in Arabic in many different ways, for example, the word Google can be written in three different ways, as “غوغل ,جوجل and قوقل”. Additionally, Arabic letters can be written using different shapes according to their position in the word, for example, the letter alif has four forms, namely آ ,إ ,أ, and ا. Finally, because users typically informally write messages and posts, data scantiness is prevalent on social media [15].

Arabic entity recognition systems still face problems, including a lack of knowledge related to analyzing data, especially data written in colloquial Arabic dialects. Consequently, users are suffering in analyzing references to entities over a selected time interval due to lack of developed systems. Displaying the frequency of references to a particular entity in a selected period demonstrates society's interest in a special event and confirms how that interest changes over time. Besides, showing the frequency of an entity is repeated in different categories (person’s name, company, and location) is important, especially for developers and researchers in particular fields such as marketing, advertising, and economics.

The contribution of this research work is to develop a solution to deal with text written in classical Arabic, modern standard Arabic, and colloquial Arabic dialects by classifying the entities referenced into categories, mainly person’s name, location, and organization. This research work also helps to understand the use of a machine learning approach in building a system for recognizing entities referenced in Arabic on social media, especially on Twitter, which follows different writing rules than those used in news or formal texts. Furthermore, this work has significant importance in visualizing the frequency of references to entities in a given time, as it is important for researchers in different fields to understand how society’s interest in some subjects can change over time.

This paper is organized into five sections. Section 2 provides an overview of several previous systems related to the current study and recent advances in ANER. Section 3 explains how the system used in the current study was built and how it works, while Section 4 contains the results of the research and a discussion of these results. Finally, Section 5 presents the conclusion and a discussion of possible future work.

II. Literature Review

ANER researchers have been developing systems that depend on two approaches: the rule-based approach and the machine learning (ML) approach. Rule-based systems depend on linguistic rules for recognizing named entities (NEs). The advantage of using a rule-based approach for design NER systems is that approach depends on the core of linguistic knowledge. Additionally, the ruled-based approach is not portable.

Due to these problems, researchers turned their attention to machine learning-based approaches, which involve learning NE tagging decisions from annotated texts. Machine learning-based approaches can be classified into three categories: supervised, unsupervised and semi-supervised. Supervised methods employ tagged corpus to learn the model [4]. Here, sets of features are extracted for each word. The tags of words act as supervisors to tune the model parameters. Unsupervised methods can perform NER in the absence of labeled data. They try to learn representations from the data. Furthermore, the most common approach used in ML for NER is supervised learning (SL). Support vector machines (SVM) and conditional random fields have identified as the most suitable techniques for ANER descend by using the ML approach.

Hybrid technologies use a combination of rule-based and machine learning-based algorithms [2, 6]. The empirical results indicate that the hybrid approach outperforms both the rule-based and the ML-based approaches when they are processed independently because it combines two technologies.

A. Name Entity Recognition Systems

The NERA system developed in [23, 24] generalizes the findings from the Person Name Entity Recognition for Arabic (PERA) system. NERA addresses major challenges posed by NER in the Arabic language. With an architecture optimized for rule-based systems, the NERA system has three components—gazetteers, local grammar in the form of regular expressions, and a filtering mechanism—that uses indicators to formulate recognition rules to solve the lack of capitalization for proper nouns in their system. The researchers built their corpora due to the unavailability of free Arabic corpora for research purposes. Moreover, commercially available Arabic corpora are oriented toward the newswire domain, so the researchers found unequal coverage of the types of named entities involved in their research.

A hybrid approach [22] using decision trees supported by support vector machines and logistic regression classifiers for an ANER system to evaluate the system performance was developed to recognize eleven types of Arabic named entities: person, location, organization, date, time, price, measurement, percent, phone number, ISBN, and file name. Moreover, the authors in [18] developed NERA 2.0 to enhance the coverage and performance of the rule-based NER system. They present a novel methodology for improving the rule-based component (NERA), using gazetteers and Arabic linguistics rules, and also using indicators or keywords for NEs, which assist the detection process via the linguistic rules. The mechanism uses the recognition decisions made by the hybrid NER system to identify the weaknesses of the rule-based component and derive new linguistic rules aiming to enhance the rule base, which helps to achieve more reliable and accurate results.

Furthermore, authors in [25] presented an Arabic information extraction (IE) system used to analyze large volumes of news text every day to extract several NE types: person, organization, location, date, number, and quotations by and about individuals. Similarly, in [11] the authors developed RenA, an NER for the Arabic language. RenA extracts entities from news articles collected from online resources, and a chunker tokenizes the text based on whitespace. The authors in [3] proposed a rule-based NER system that can be used in Web applications. The system was evaluated using ANERcorp. The researchers conducted two experiments to study the effect of Arabic prefixes and suffixes on the recognition results. The verification process improved the recognition results of NEs across all types, although these improvements were not symmetrical.

Alanazi et al in [4] used features such as corpus, gazetteers, and part-of-speech (POS) tags of words to develop an ANER system for text in the medical domain. The authors in [1] presented an integrated approach between two machine learning techniques, semi-supervised pattern recognition and conditional random fields (CRF) classifier as a supervised technique. Gazetteers and different types of indicators are used for classification. The authors developed an integrated semantic-based ML model [6] for enhancing ANER. The authors' idea was to combine several linguistic features and to use syntactic dependencies to infer semantic relationships between named entities. Accordingly, the model combines internal features that represent linguistic features such as part of speech, gazetteer, indicators, and corpus and external features that represent the semantics of relationships between the three named entities, such as classes, instance, and relations, to enhance the accuracy of recognizing them using external knowledge sources, such as Arabic WordNet (AWN) ontology. They introduced internal and external features to the CRF classifier, which is an effective strategy for ANER [7]. Zayed and El-Beltagy in 2015 [26] focused on the task of Arabic person names recognition. Their model is a combination of rule-based and statistical models, and it uses the unsupervised learning of patterns and clustered dictionaries as constraints to identify a person's name and resolve its ambiguity.

In [10] the author aimed to build a tool for POS tagging and NERA using a rule-based technique. The POS tagger contains two phases. In the first phase, words are pasted into a lexicon analyzer and in the second phase morphological information is used to presume the class of the word. The named-entity detector applies the rules to the text and assigns the correct label (tag) for each word. The authors in [2] built a Malayalam NER classifier and in [17] for Vietnamese by using a neural network. Liu and et al. in [14] provided a semi-supervised learning framework to recognize entities in English tweets by combining a linear conditional random fields (CRF) model and K-nearest neighbors (KNN) classifier. Likewise, Patawar and Potey in [19] implemented a semi-supervised algorithm, also by combining the CRF approach with the KNN classifier, to build a NER system for Indian tweets. Finally, in [9] the authors presented TwitterNEED, a hybrid approach for named entity extraction and named entity disambiguation for English tweets.

Researchers have been developing ANER systems [21] based on one of three approaches: the ML, rule-based, and hybrid approaches. The ML approach uses three technologies for classification: supervised, unsupervised, and semi-supervised. Moreover, common resources on which researchers can draw to build their classifiers and recognition systems include corpus (customized or off-the-shelf), gazetteers, indicators, decision tree, logistic regression, POS tags, and Arabic linguistics.

To the best of author's knowledge, no Arabic entity recognitions systems exist for use with colloquial Arabic dialects. Although some studies have involved Twitter content, but, they did not support Arabic content, thus they cannot be used for the Arabic language because the Arabic language follows different and specific rules, also some work use twitter content but not classify to different classes.

Furthermore, the systems examined do not allow users to search for information about specific entities during selected periods. Consequently, the current study aimed to build an ANER system to fill that gap. The system could be used for Arabic tweets, and it could provide users with the flexibility to search for information about an entity during a specific through tweets.

III. Research Methodology

The main goal of the research was to classify and visualize entities frequently referred to on Twitter over a given time. To achieve the research objectives, machine-learning techniques were used to build a system able to visualize Arabic social media content, focusing on Twitter. The system process involves four steps: identifying the entities and time periods, preprocessing or preparing the data, building the classifier model, and performing data visualization of frequency and classification. The steps are discussed in detail below. Fig. 1 shows the system flow chart.

Step 1: Data Collection: The first step was to clarify the difference between "entity" and "named entity" [5]. "Entity" is defined as something that exists by itself, such as locations, humans, or animals. When an entity is given a name, it becomes a "named entity". As an example, Al Riyadh city would be a named entity. In this research, an entity refers to any Arabic word, including adjectives, verbs, and nouns.

The user, in this case the researcher, sets the data collection parameters, the entity reference in Arabic script, and the time and date in a valid format ("DD/MM/YYYY"). The system then retrieves tweets written during the selected time containing a reference to the entity. Subsequently, the system sends the retrieved tweets to pre-processing before saving them to a file.

The code used for retrieving tweets from Twitter by date and/or keyword(s) was originally written by Simon Lindgren. The researcher edited the code to make it suitable for Arabic tweets.

Step 2: Preprocessing: The data (tweets that contain a reference to the given entity posted during the given time) undergo a cleaning process to remove the redundant spaces between words and all symbols such as +, @, $, and #, punctuation, numbers, non-Arabic letters, and links. White space characters are used to define the words in the text.

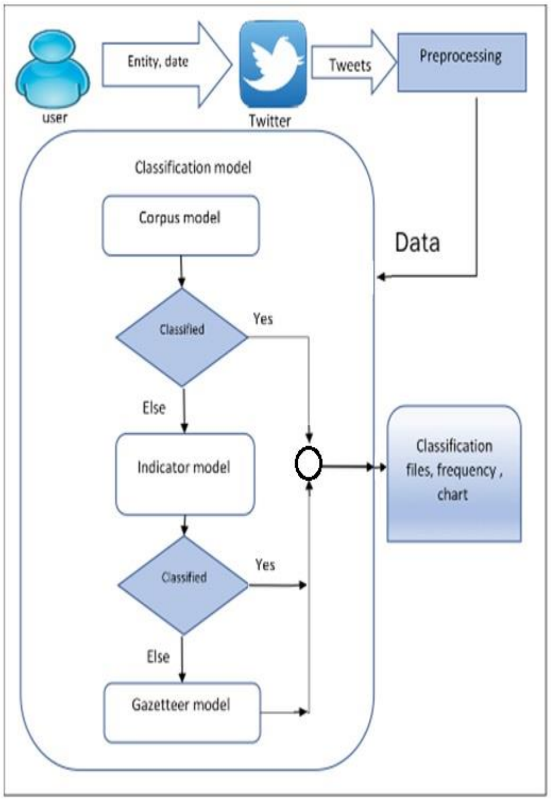

Fig. 1. System Flow Chart.

Step 3: Building the Classifier: Machine learning involves two steps [27, 28]: training, during which the classifier model is built, and testing, during which the classifier model is tested. As in previous works, this classifier does not depend on one feature; it is dependent upon four features and uses both internal and external features to build the model. Internal features include corpus, indicators, and gazetteers. The corpus is an annotated linguistic resource that shows the Arabic grammar, syntax, and morphology for each word. In this model, the corpus was used to show the POS tag for the entity. Indicators are a set of words used to recognize names (NEs). They include preceding indicators (e.g., Mr. "السيد”) and post-indicators word (e.g., almotahedah “المتحده”). Gazetteers can be defined as a list of names of locations, organizations, people, and so forth.

External features were extracted using ontology: “a formal naming” and definition of the types, properties, and interrelationships of the entities that exist for a particular domain of discourse. In this case, the classifier follows three steps to classify the entities: classification based on the corpus model, a model classified by indicators, and using gazetteers.

The works in [1, 3, 4, 23, 24] using a corpus to classifying and recognize in their works. Corpus is a strong method that gives reliable results. It depends on manual tagging by the developer and on context similarity; however, it cannot classify all possible entities because it can only classify the entities found on the corpus vocabulary list. This restriction led to the use of other classification methods. One such method was searching for indicators. Using indicators [1, 6] enhances the accuracy of NER. In the case of failure to classify through use of indicators, the classification model will refer to gazetteer lists. Many researchers [1, 4, 6, 18] use gazetteers to improve their works.

Ontology is known as a digital repository or a large database that finds connections based on an understanding of the meanings and relationships. Ontology has five components: entity or vocabulary, classes, properties that have attributes for each class, relationships (characterized by their descriptions, names, connotations), and language axioms (including clear rules that restrict the use of concepts, and include confirmed and previously known data). This ordering comes from using ontology rules, which detract from the corpus and indicators methods and depend on linguistic grammar for classification. A description of the three steps follows.

a) Corpus model: In the current research, the researcher rebuilt the corpus and made changes to adapt the model to the Arabic language. The first step in this process was building a customized corpus for the classifier, which required that the sentences contain both the entity reference and its POS tag, separated by a forward slash (/). The corpus encompasses samples from tweets during different periods—approximately 250 tweets and more than 2,000 entity references classified as person, location, organization, or non. These entity references could mean a person’s name or a location, for example, شام, or Sham. Sham could be a person’s name or the geographical area that includes Syria, Lebanon, Palestine, and Jordan. Another example is Hend, which can be a person’s name or the Arabic name of India. Additionally, the corpus includes entity references with different purposes, such as Mesk, which could be a person’s name or Mesk International Company. Some entity references could refer to all three categories. For example, Najd could be a person’s name, a location (Najd Plateau), or the name of a firm. Using these types of entity names allows the system to check the accuracy of classifications.

Using a customized corpus that depends on tweets yields more accurate results than using a corpus that depends on news content, as tweets are written in colloquial Arabic dialects that are used more commonly than formal Arabic [13, 20] found in a study involving the Turkish language that accuracy dropped from 91% to 19% when using NER from news rather than tweets.

The corpus method uses the KNN algorithm for classification. The KNN algorithm is a supervised learning approach; the model contains both entity names and POS tags and based on the similarity approach feature, uses context similarity. In this case K = 3. In the corpus for each entity referring to one class, there were about three or four sentences.

Consequently, the model first separates the sentence into smaller parts, with each small part containing an entity reference and its POS tags. It assigns each entity an ID with a normal random number and an ID for POS tags where each node contains the entity reference and the three preceding words. If fewer than three entities are involved, the entity value = unknown word, word ID = 0, POS tag = unknown tag, and POS ID = 0.

The classifications are applied only to entities that are listed in the corpus and that have IDs and POS tags. If an unannotated word appears, the model drops the whole sentence.

For example, in corpus:

جميل/per الذيابي/per الفارق/non بين/non التخاذل/non و/non المواجهة/non عكاظ/org مقالات/non

POS ID = [22 22 7 7 7 7 7 5 7]

جميل الذيابي الفارق بين التخاذل والمواجهة عكاظ مقالات

['جميل']

Word ID =12

[جميل, unknown word, unknown word, unknown word]

Tag= [per, unknown pos, unknown pos, unknown pos]

POS ID = [22,0,0,0]

POS tag=['per']

Node= [(12,22), (0,0), (0,0), (0,0)]

['جميل', 'الذيابي']

Word ID (691 = (الذيابى

[جميل, الذيابي, unknown word, unknown word]

Tag = [person, person, unknown pos, unknown pos]

POS ID = [22,22,0,0]

POS tag = ['per', 'per']

Node = [(12,22), (691,22), (0,0), (0,0)]

['جميل', 'الذيابي', 'الفارق']

Word ID ('692 = ('الفارق

[جميل, الذيابي, الفارق, unknown word]

Tag = [person, person, non, unknown pos]

POS ID = [22,22,7,0]

POS tag = ['per', 'per', 'non']

Node = [(12,22), (691,22), (692,7), (0,0)]

['جميل', 'الذيابي', 'الفارق', 'بين']

Word ID ('25 = (’بين

['جميل, الذيابي, الفارق, 'بين]

Tag = [person, person, non, non]

POS ID = [22,22,7,7]

POS tag = ['per', 'per', 'non', 'non']

Node = [(12,22), (691,22), (692,7), (25,7)]

Sending that sentence to the model for classifying entity (عكاظ):

جانب من تجهيزات جناح وزارة التعليم المشارك في الفعاليات واللى يفكر نفسه داخل سوق عكاظ

In corpus model:

[1416,302,41,10] set of word IDs for ['عكاظ' ,'سوق' ,'داخل' ,'نفسه']

[7,7,7,5] set of POS IDs for ['عكاظ' ,'سوق' ,'داخل' ,'نفسه']

Node for entity (10,5) ,(41,7) ,(302,7) ,(1416,7)] = (عكاظ)]

Classification and KNN chart shown in Fig. 2.

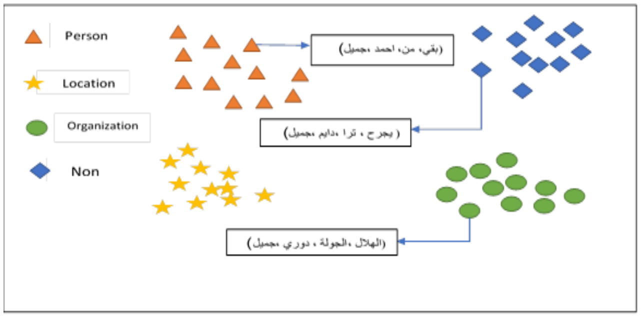

Fig. 2. K Nearest Neighbor Chart for Corpus Model.

b) Indicators model: The second priority for classification is using the indicators model. In this model, a deep neural network was used to classify the entity by using pre- or post-indicators. A neural network "is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates". Neural networks are powerful tools in the natural language processing of text. The original neural network algorithm used in this research was written by Source Dexter to classify the text into time, age, apology, greeting, or farewell. It was edited to be suited for classifying Arabic entities. This model depends on the unsupervised learning approach for classification. To ensure the reliability of the classification, the tweets that contained entity references were broken down into small parts containing a single entity and the following indicator for each word in the tweet. This approach is known as the n-gram technique. In this case (n = 2). Results were then classified by the order of entities in grams and indicators. The input layer accepts the gram that contains the entity reference, while the hidden layers that contain the indicators and classes classify the sentence. Subsequently, the output layer provides the classification. Fig. 3 shows the layer of the neural network.

Indicators are categorized as pre- and post-indicators for each class: the person, location, and organization classes. Arabic ontology axioms and grammar rules are followed in classifications by using the indicators, specifically by using a bigram, which is a model that looks for the entity and for the previous or next word and the group to which it belongs. It uses only indicators that come directly before or after the word.

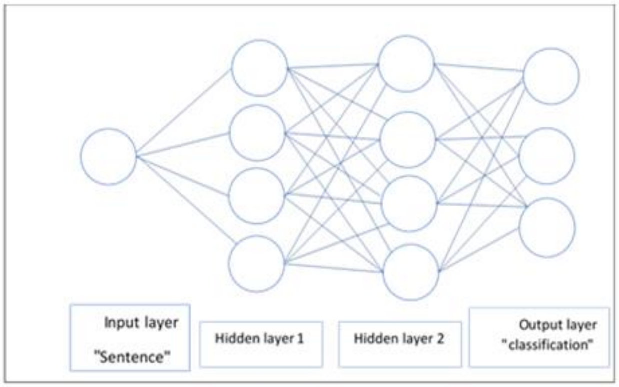

Fig. 3. Neural Network Chart for Indicator Model.

The indicators model attempts to list all possible formal and informal indicators that may contain the following.

- Name title, for example, السيد – mister and باشا – basha
- Job title, for example, ا لأستاذة – professor
- Relative adjective, for example, المتحدة – united and الأدبي – al Adabi
- Semantic fields, for example, شركة – company
- Interjection articles, for example, يا – oh
- Location indicators, for example, مطار – airport and شارع – street

c) Gazetteer model: Liu et al. in [14] and Patawar and Potey in [19] found that using gazetteers significantly improved the accuracy of NER. There are three classes of gazetteers: location, person, and organization. The gazetteers used in the current study include more than 2,000 names of people; more than 2,000 names of locations such as continents, countries, cities, states, political regions, towns, and villages; and approximately 300 names of organizations, including media companies, newspapers, construction firms, banks, insurance companies, airlines, telecommunication companies, and football teams. Entity names that might fall into more than one class were omitted from the gazetteers used in this research to improve the accuracy of the classification. For example, سعيد was omitted because it can be used as a person's name and as the adjective happy, and روما (Roma) was omitted because it is the name of a city and a football team.

The ontology was applied to ordering the gazetteers in the following manner: (1) person's name, (2) location's name, and (3) organization's name. If an entity is identified in one of the gazetteers, it is classified as being in that gazetteer's class.

Step 4: Data Visualization: The human brain tends to process visual information far more easily than written information. Visualization makes the results easier to read and encourage users to explore and even manipulate the data to uncover other factors. This creates a better attitude for use of analytics. Visualization is the act of interpreting visual terms or of putting into visible form.

In this step, the frequency of the entity returns to the user for each class. It displays the results in two ways. One visualizes the results during a given time interval, the other breaks the time intervals down and shows the frequency of each class during these smaller periods. Additionally, the tweets related to classified entities are saved in files depending on their class type and an additional file for tweets that content unclassified entity.

IV. Results and Discussion

A confusion matrix is a table that describes the performance of a classification model on a set of test data where the true values are known, with rows representing actual classes and columns representing predicted classes.

The accuracy of the model is measured by the overall success rate, calculated using (1).

$$Accuracy = \frac{total\ correct\ predicted}{total\ sample} \quad (1)$$

Misclassification, calculated using (2), is the rate of incorrect classifications.

$$Misclassification = 1 - accuracy \quad (2)$$

Precision, a positive predictive value, assesses what percentage of items identified as positive are positive. It can be calculated using (3).

$$precision = \frac{correct\ entities\ classefied}{total\ entities\ in\ acual\ class} \quad (3)$$

Recall (sensitivity) is the percentage of relevant instances that have been retrieved correctly from the total number of relevant instances, calculated as follows (4).

$$Recall = \frac{correct\ entities\ recognized}{total\ entties\ predeicted\ in\ same\ class} \quad (4)$$

The researcher used entity references that could refer to a name or a location, that could be the name of a person or an organization, and that could not be classified as any of the three classes (non) and randomly selected tweets posted at different times. Each individual model was measured before combining them.

A. Testing Models

a) Corpus model: The researchers used the entity Jamil to test the corpus model.

Entity 1: جميل – Jamil

Jamil can be the name of a person or an organization, or it can mean beautiful. Eighty-four tweets posted between 27/11/2014 and 30/11/2014 that contains a reference to "جميل" was randomly selected. The researcher measured the performance of the algorithm—specifically accuracy, recall, and precision—using the confusion matrix shown in Table II. Fig. 4 shows the frequency of each class in the whole period, while Fig. 5 shows the visualization of classification day by day.

Accuracy = 0.61

Misclassification = 0.39

Recall (organization) = 0.42

Recall (person) = 0.60

Recall (non) = 0.90

Precision (organization) = 0.87

Precision (person) = 0.60

Precision (non) = 0.50

TABLE II. CONFUSION MATRIX FOR JAMIL ENTITY USING THE CORPUS MODEL

N=84	Predicted: Organization	Predicted: Person	Predicted: Non	Total
Actual: organization	20	0	3	23
Actual: person	2	3	0	5
Actual:non	26	2	28	56
	48	5	31	

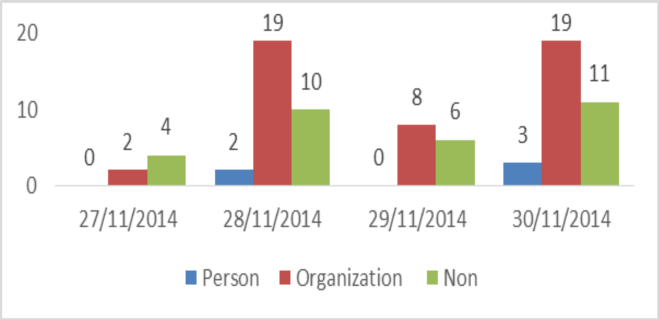

Fig. 4. Visualization of Classification of Jamil Entity Day by Day, using a Corpus Model.

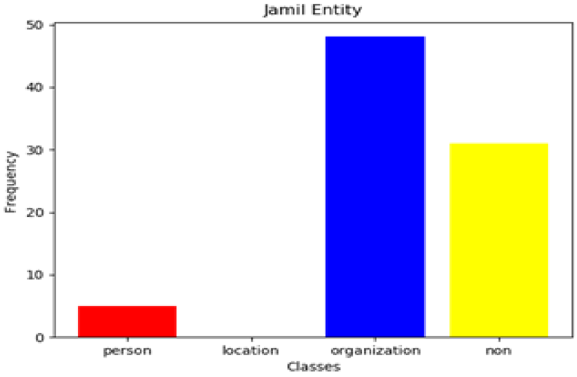

Fig. 5. Visualization of Classification of Jamil Entity as a Whole, Corpus Model.

b) Indicator model: In this model, different indicators were used in one testing sample measure. Some of the tweets contained more than one indicator.

Entity 1: خالد – Khalid.

The Khalid entity can be used with different indicators and classifying in different classes, for example, السيد خالد – Mr. Khalid, دكتور خالد – Dr. Khalid, مطار خالد – Khalid Airport, شارع خالد– Khalid Street, or may mean immortal. Ninety -nine tweets that contain the Khalid entity ware used to testing the indicator model, with some tweets containing more than one indicator, posted between 15/06/2016 and 19/06/2016. Table III shows the confusion matrix for this entity. Fig. 6 is visualized the frequency of each class, while Fig. 7 is the visualization day by day of the prediction classes for the testing sample.

TABLE III. CONFUSION MATRIX FOR KHALID ENTITY USING THE AN INDICATOR MODEL

N=99	Predicted : person	Predicted: location	Predicted: organization	Predicted: non	Total
Actual: person	69	1	0	0	70
Actual: location	12	16	0	0	28
Actual: organization	1	0	0	0	1
Actual: non	0	0	0	0	0
Total	82	17	0	0	

Accuracy = 0.80

Misclassification = 0.20

Recall (person) = 0.77

Recall (location) = 0.94

Recall (organization) = 0

Precision (person) = 0.90

Precision (location) = 0.47

Precision (organization) = 0

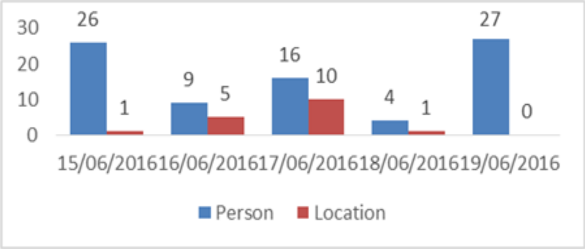

Fig. 6. Visualization of Classification of Khalid Entity Day by Day using Indicator Model.

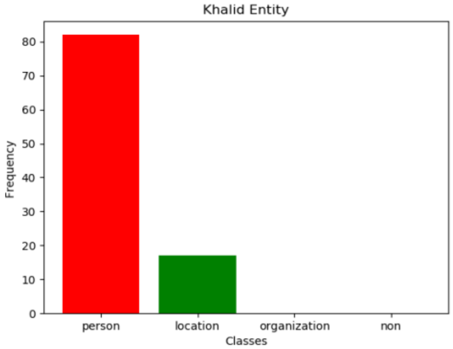

Fig. 7. Visualization of Classification of Khalid Entity as a Whole using an Indicator Model.

c) Gazetteer model: For testing the gazetteer model, the researcher chose entities with ambiguous meanings. For the name of an organization, فورد – Ford was chosen; for location, كندا – Canada; and for a person, العيسى– Al Easa was chosen.

Entity 1: فورد – Ford.

The researcher selected 57 tweets posted between 12/09/2014 and 16/09/2014 as a testing sample. The Ford entity can be a person or an organization, but the term is most often used in Arabic to refer to the Ford Company.

Table IV shows the confusion matrix for testing this sample. Fig. 8 visualizes the frequency of each class, while Fig. 9 is the daily visualization of the prediction classes for the testing sample.

Accuracy = 0.84

Misclassification = 0.16

Recall (organization) = 0.84

Precision (organization) = 1

TABLE IV. CONFUSION MATRIX FOR FORD ENTITY USING THE GAZETTEER MODEL

N=57	Predicted: person	Predicted: location	Predicted: organization	Predicted: non	Total
Actual: person	0	0	9	0	9
Actual: location	0	0	0	0	0
Actual: organization	0	0	48	0	48
Actual: non	0	0	0	0	0
Total	0	0	57	0	

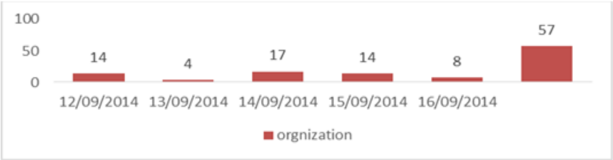

Fig. 8. Visualization of Classification of Ford Entity Day by Day using a Gazetteer Model.

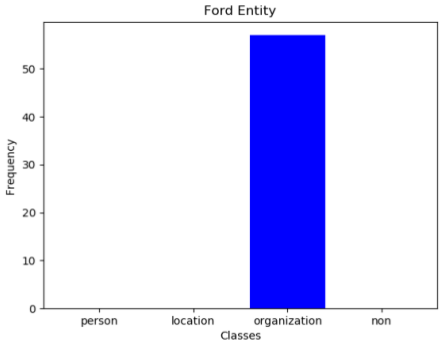

Fig. 9. Visualization of Classification of Ford Entity as a Whole using a Gazetteer Model.

B. Testing the Whole System

The same samples tested using individual models were retested using the proposed system. The tweets that remained unclassified after testing with the corpus model were tested using the indicator model. Those that still remained unclassified were then tested using the gazetteer model. Any tweets that remained unclassified after this were classified as non.

Entity 1: جميل – Jamil

The confusion matrix for the test conducted on this entity using the proposed system appears in Table V, with the confusion matrix measuring accuracy, recall, and precision.

Using the corpus method, the researcher classified 34 entities, the remaining 61 tweets being sent to the indicator method. One entity was classified as an organization through the indicator method, and the rest sent to the gazetteer method, where they were classified as non. Fig. 10, Fig. 11, and Fig. 12 visualize the results from each step, while Fig. 13 visualizes a daily classification, and Fig. 14 visualizes the frequency for each class as a whole.

Accuracy = 0.93

Misclassification = 0.07

Recall (organization) = 0.95

Recall (person) = 1

Recall (non) = 0.92

Precision (organization) = 0.87

Precision (person) = 0.60

Precision (non) = 0.98

TABLE V. CONFUSION MATRIX FOR JAMIL ENTITY USING THE WHOLE SYSTEM

N=84	Predicted: organization	Predicted:person	Predicted: non	Total
Actual: organization	20	0	3	23
Actual: person	0	3	2	5
Actual: non	1	0	55	56
Total	21	3	60	

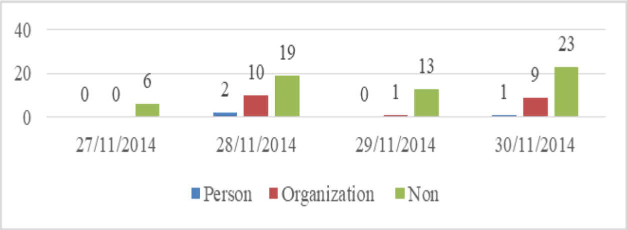

Fig. 10. Visualization of Results from First Step in Testing Jamil Entity through the whole Model.

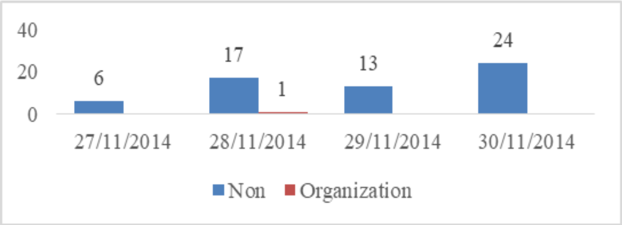

Fig. 11. Visualization of Results from Second Step in Testing Jamil Entity through the Whole Model.

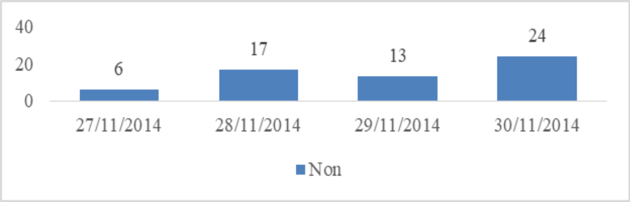

Fig. 12. Visualization of Results from Third set in Testing Jamil entity through the whole Model.

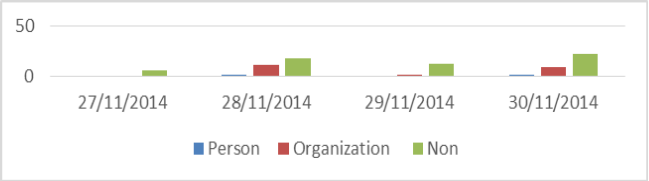

Fig. 13. Visualization of Classification for Jamil Entity Day by Day using the whole System.

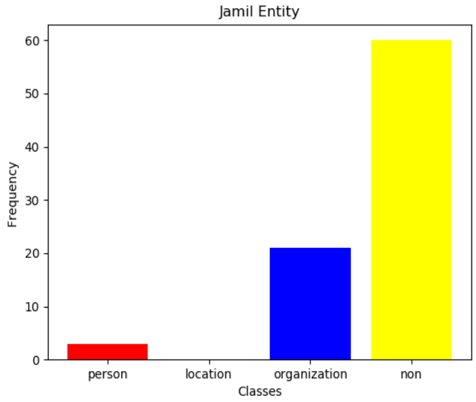

Fig. 14. Visualization of Classification for Jamil Entity as a whole using the whole System.

Entity 2: عكاظ – Aucadh

Out of total (114) as shown in Table VI, a total of 77 entities were classified through the corpus method. The remaining 37 tweets that had unclassified entities were sent to the indicator method. Of these, 16 entities were successfully classified, and the rest were classified using the gazetteer method. Fig. 15, Fig. 16, and Fig. 17 visualize the results from each step, while Fig. 18 visualizes a classification of each day individually in total, and Fig. 19 visualizes the frequency for each class as a whole.

Accuracy = 0.85

Misclassification = 0.15

Recall (location) = 0.71

Recall (organization) = 0.95

Recall (non) = 0

Recall (person) = 0

Precision (location) = 0.81

Precision (organization) = 0.86

Precision (non) = 0

Precision (person) = 0

TABLE VI. CONFUSION MATRIX FOR AUCADH ENTITY USING THE WHOLE SYSTEM

N=114	**Predicted: location**	**Predicted: person**	**Predicted: organization**	**Predicted: non**	**Total**
Actual: location	22	0	4	1	27
Actual: person	0	0	0	0	0
Actual: organization	10	2	75	0	87
Actual: non	0	0	0	0	0
Total	32	2	79	1	

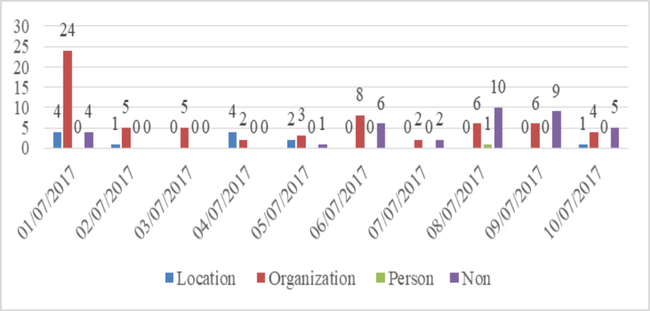

Fig. 15. Visualization of Results from First Step in Testing Aucadh Entity through the whole Model.

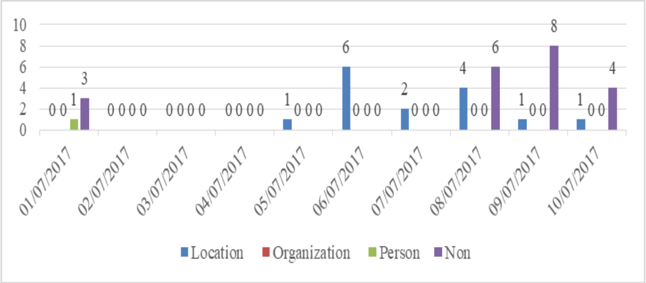

Fig. 16. Visualization of Results from Second Step in Testing Aucadh Entity through the whole Model.

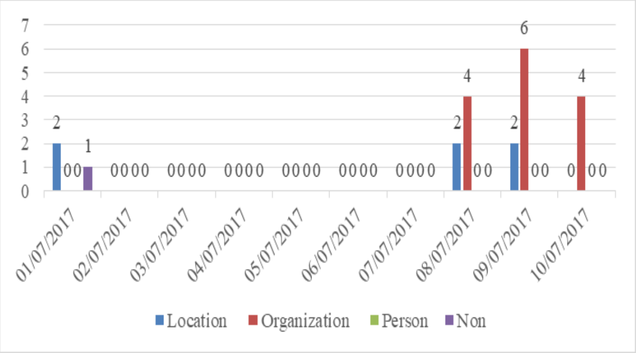

Fig. 17. Visualization of Results from Third Step in Testing Aucadh Entity through the whole Model.

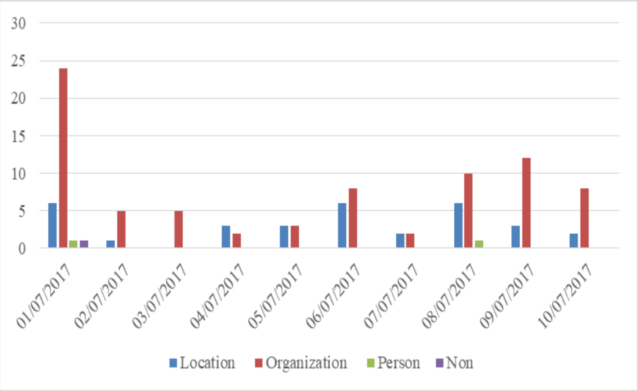

Fig. 18. Visualization of Classification for Aucadh Entity Day by Day using the whole System.

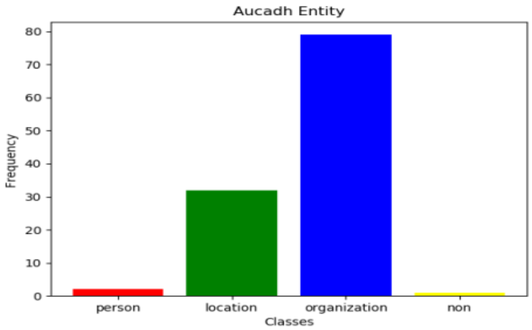

Fig. 19. Visualization of Aucadh Entity as a whole using whole System.

Entity 3: فورد – Ford

Table VII shows the confusion matrix for testing this sample. The corpus method classified 20 entries, with the remaining 37 being classified through the gazetteer method as an organization. Fig. 20 and Fig. 21 visualize the results from each step, while Fig. 22 visualizes a total classification of each individual day, and Fig. 23 visualizes the frequency of each class as a whole.

Accuracy = 0.88

Misclassification = 0.12

Recall (organization) = 0.9

Recall (Person) = 1

Precision (organization) = 0.98

Precision (Person) = 0.33

TABLE VII. CONFUSION MATRIX FOR FORD ENTITY USING WHOLE SYSTEM

N=57	**Predicted: person**	**Predicted: organization**	**Predicted: location**	**Tot al**
Actual: person	3	6	0	9
Actual: organization	0	47	1	48
Actual: location	0	0	0	0
Actual:non	0	0	0	0
Total	3	53	1	

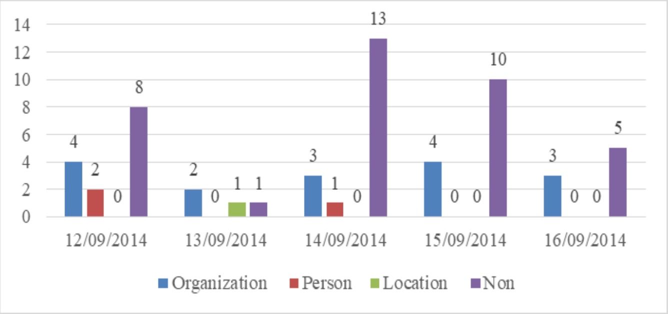

Fig. 20. Visualization the Results from First Step in Testing Ford Entity through the whole Model.

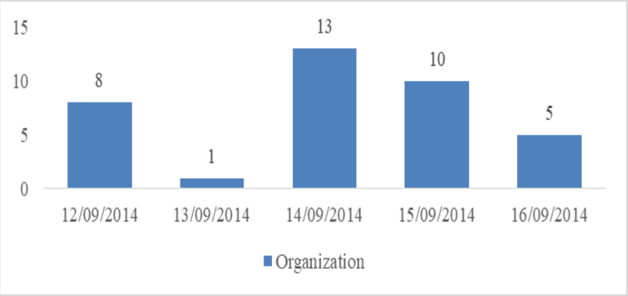

Fig. 21. Visualization the Results from Third Step in Testing Ford Entity through the whole Model.

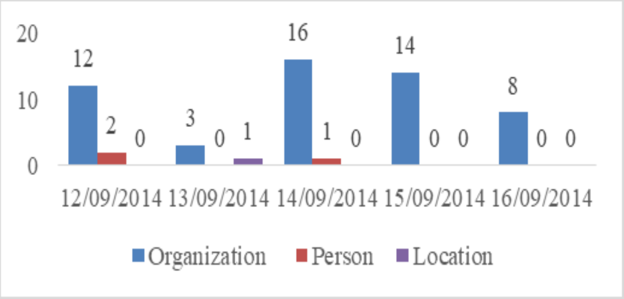

Fig. 22. Visualization of Ford Entity Classification Day by Day using whole System.

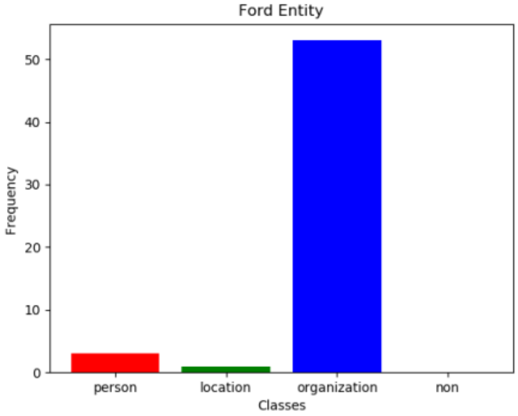

Fig. 23. Visualization of Ford Entity Classification as a whole using whole System.

The accuracy of each model was measured by testing the system and showing the difference between actual classes and the predicted classes of the testing sample using a confusion matrix and then calculating accuracy. The accuracy of testing samples on a single model and the whole system are set out in Table VIII. Using the designed NER system improved the accuracy in some testing samples (Jamil, Aucadh, Khalid, and Ford). The system is 90% accurate, and 10% misclassification.

TABLE VIII. COMPARING ACCURACY OF SINGLE MODEL AND WHOLE SYSTEM

Entity	Testing method	Accuracy
Jamil	Corpus model	61%
	Whole system	93%
Aucadh	Corpus model	71%
	Whole system	85%
Ford	Gazetteer model	84%
	Whole system	86%

V. CONCLUSION

In this research, the researcher developed an Arabic NER system that classifies selected entities mentioned in tweets posted during a specific period as POS (person name, organization, or location). The system uses a machine learning technique to build the classifier, which follows three steps arranged by priority for tagging—using similarity as in customized corpus, using an order of entity and the indicator if found, and, finally, using a list of gazetteers for each class. The system is 90% accurate. One of the main problems encountered during the study was building the customized corpus from tweets. Searching and annotating tweets that contained entity references that can fall into multiple categories depending on their location in the sentence was time-consuming.

In future work, the researcher plans to increase the classification types to more than three types. Furthermore, the researcher aims to make it possible to recognize more than one entity (word) and hopes to solve the limitation problem in retrieving tweets posted after 17/06/2016.

ACKNOWLEDGEMENT

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology – the Kingdom of Saudi Arabia, award number (1-18-02-008-0003).

REFERENCES

[1] Abdel Rahman S., Earnest M., Mandy M., and Filmy A., “Integrated machine learning techniques for Arabic named entity recognition” International Journal of Computer Science Issues, vol. 7, no. 4, pp. 27–36, 2010.

[2] Ages A. P., and Ridicule, S. M., “A named entity recognition system for Malayalam using neural networks” Procedia Computer Science, vol. 143, pp. 962–969, 2018.

[3] Al-Jumaily H., Martínez P., Martínez-Fernández J., and Goot, E., “A real time named entity recognition system for Arabic text mining” Language Resources and Evaluation, vol. 46, no. 4, pp. 543–563, 2012.

[4] Alanazi S., Sharp B., and Stanier, C., “A named entity recognition system applied to Arabic text in the medical domain” International Journal of Computer Science Issues, vol. 12, no. 3, pp. 109–117, 2015.

[5] Alkharashi I., “Person named entity generation and recognition for Arabic language” In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, pp. 205–208, 2009.

[6] Alsayadi H. A., and Elkorany A. M., “Integrating semantic features for enhancing Arabic named entity recognition” International Journal of Advanced Computer Science and Applications, vol. 7, no.3, 2016.

[7] Benajiba Y., Diab M., and Rosso, P., “Arabic named entity recognition using optimized feature sets” In M. Lapata, H. T. Ng, & X. Wan (Eds.), Proceedings of the Conference on Empirical Methods in Natural Language Processing – EMNLP 08, pp. 284–293, 2008.

[8] Elgibali A., Investigating Arabic: Current parameters in analysis and learning (Studies in Semitic languages and linguistics), Boston, MA: Brill Academic Publishers, 2005.

[9] Habib M. B., and Keulen M. V., “Twitter NEED: A hybrid approach for named entity extraction and disambiguation for tweet” Natural Language Engineering, vol. 22, no. 3, pp. 423–456, 2015.

[10] Hjouj M., Alarabeyyat A., and Olab I., “Rule based approach for Arabic part of speech tagging and name entity recognition” International Journal of Advanced Computer Science and Applications, vol. 7, no.6, pp. 331–335, 2016.

[11] Kanan T., Kanaan R., Al-Dabbas O., Kanaan G., Al-Dahoud A., and Fox, E., “Extracting named entities using named entity recognizer for Arabic news articles” International Journal of Advanced Studies in Computers, Science and Engineering, vol 5, no.11, pp. 78–84, 2016.

[12] Korayem M., Crandall D., and Abdul-Mageed, M., “Subjectivity and sentiment analysis of Arabic: A survey” In A. E. Hassanien, A. M. Salem, R. Ramadan, & T. Kim (Eds.), Advanced machine learning technologies and applications, AMLTA 2012, Communications in Computer and Information Science vol. 322, pp. 128–139, 2012.

[13] Kucuk D., Jacquet G., and Steinberger, R., “Named entity recognition on Turkish tweets” In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani … S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland: Association for Computational Linguistics (ACL), pp. 450–454, 2014.

[14] Liu X., Wei F., Zhang S., and Zhou, M., “Named entity recognition for tweets” ACM Transactions on Intelligent Systems and Technology, vol. 4, no. 1, pp. 1–15, 2013.

[15] Martínez-Cámara E., Martín-Valdivia M. T., Ureña-López L. A., and Montejo-Ráez A. R., “Sentiment analysis in Twitter” Natural Language Engineering, vol. 20, no. 1, pp. 1–28, 2012.

[16] Monem A. A., Shaalan K., Rafea A., and Baraka, H., “Generating Arabic text in multilingual speech-to-speech machine translation framework” Machine Translation, vol. 22, no. 4, pp. 205–258, 2008.

[17] Nguyen V. H., Nguyen H. T., and Snasel, V., “Named entity recognition in Vietnamese tweets” In M. Thai, N. Nguyen, & H. Shen (Eds.), Computational Social Networks, Lecture Notes in Computer Science, pp. 205–215, 2015.

[18] Oudah M., and Shaalan K., “NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic” Natural Language Engineering, vol. 23, no. 3, pp. 441–472, 2016.

[19] Patawar M., and Potey, M., “Named entity recognition from Indian tweets using conditional random fields based approach” International Journal of Advanced Research in Computer Engineering & Technology, vol 5, no.5, pp. 1541–1545, 2016.

[20] Ritter A., Clark S., Mausam., and Etzioni O., “Named entity recognition in tweets: An experimental study” In R. Barzilay & M. Johnson (Eds.), Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534, 2011.

[21] Shaalan K., “A survey of Arabic named entity recognition and classification” Computational Linguistics, vol. 40, no. 2, pp. 469–510, 2014.

[22] Shaalan K., and Oudah M., “A hybrid approach to Arabic named entity recognition” Journal of Information Science, vol. 40, no. 1, pp. 67–87, 2013.

[23] Shaalan K., and Raza H., “Arabic named entity recognition from diverse text types” In B. Nordström & A. Ranta (Eds.), Advances in natural language processing. Lecture notes in computer science, vol. 5221, pp. 440–451, 2008.

[24] Shaalan K., and Raza, H., “NERA: Named Entity Recognition for Arabic” Journal of the American Society for Information Science and Technology, vol. 60, no. 8, pp. 1652–1663, 2009.

[25] Zaghouani W., “RENAR: A Rule-Based Arabic Named Entity Recognition System” ACM Transactions on Asian Language Information Processing, vol. 11, no. 1, pp. 1–13, 2012.

[26] Zayed O., and El-Beltagy S., “Named entity recognition of persons’ names in Arabic tweets” Proceedings of Recent Advances in Natural Language Processing, pp. 731–738, 2015.

[27] AlAjlan, S.A., Saudagar, A.K.J., “Machine learning approach for threat detection on social media posts containing Arabic text” Evolutionary . Inteligence. 2020. https://doi.org/10.1007/s12065-020-00458-w.

[28] AlAjlan S.A., Saudagar A.K.J., “Threat Detection in Social Media Images Using the Inception-v3 Model” Proceedings of Fifth International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing, vol 1184. 2021. Springer, Singapore. https://doi.org/10.1007/978-981-15-5859-7_57.