## INCREASING THE INVESTMENT'S OPPORTUNITIES IN KINGDOM OF SAUDI ARABIA BY STUDYING AND ANALYZING THE SOCIAL MEDIA DATA

Sherif Kamel Hussein<sup>1</sup>, Sultanah AL-Sedairi<sup>2</sup> and Nawal AL-Rahimi<sup>3</sup>

 <sup>1</sup>Associate Professor – Department of Communications and Computer Engineering -October University for Modern Sciences and Arts, Giza- Egypt
 <sup>1</sup>Head of Computer Science Department, Arab East Colleges for Graduate Studies, Riyadh, KSA
 <sup>2</sup>Master of Computer Science, Arab East Colleges for Graduate Studies, Riyadh, KSA
 <sup>3</sup>Master of Computer Science, Arab East Colleges for Graduate Studies, Riyadh, KSA

#### ABSTRACT

Social networking sites are a significant source of information to know the behavior of users and to know what is occupying society of all ages and accordingly helpful information can be provided to specialists and decision-makers. According to official sources, 98.43% of Saudi youth use social networking sites. The study and analysis of social media data are done to provide the necessary information to increase investment opportunities within the Kingdom of Saudi Arabia, by studying and analyzing what people occupy on the communication sites through their tweets about the labor market and investment. Given the huge volume of data and also its randomness, a survey of the data will be done and collected from through keywords, the priority of arranging the data, and recording it as (positive - negative - mixed). The study analysis and conclusion will be based on data-mining and its techniques of analysis and deduction.

## **KEYWORDS**

Big data, Data analytics, Data Mining, Social media, Kdd, Facebook, Twitter, Instagram.

## **1. INTRODUCTION**

Data mining is one of the best and most powerful tool that can be used, when the data is very large and also randomly as is the case in social media, where data mining is linked to several areas including machine learning, statistics, information retrieval, databases, and even data visualization [1], There is one formal definition of data mining registered at Princeton University, as follows: "Data manipulation using advanced data search capabilities and statistical algorithms to discover patterns and correlations in pre-existing large databases; a method for discovering new meaning in data". The main idea of using data mining through specific algorithms is to obtain new information that is better understood by a large set of hidden or latent data [2].

The representation, analysis and extraction of actionable patterns from social media data is the basis of the concept of social media mining. It introduces basic concepts and key algorithms suitable for investigating big social media data, and it discusses theories and methodologies in various branches and sciences: computer science, data mining, machine learning., Social network analysis, network science, sociology, descriptive and analytical research (ethnography), statistics, mathematics, and others. It also includes tools for representing, measuring, modeling and mining meaningful and semantic patterns from large-scale social media data [3].

Data mining in social media can expand researchers' ability to understand new phenomena due to the use of social media and improve business intelligence to provide better services and develop innovative opportunities. For example, data mining techniques can help in identifying influential people in a particular specialty or field, discovering implicit or hidden groups in the social networking site, sensing user sentiments for proactive planning, developing systems and creating different entities such as buying specific products or making new friends, or creating new jobs, opening suitable markets, creating a new product, or building and strengthening trust among users or between users and entities. Social media mining is a thriving interdisciplinary field where researchers from different backgrounds can make important contributions to social media research and development. Data mining is an essential part in many related fields including statistics, machine learning, pattern recognition, database systems, visualization, data warehouse, and information retrieval [4].

Data mining algorithms are categorized into two classes: supervised learning algorithms and unsupervised and semi-supervised learning algorithms.

- Classification: the common example of supervised learning algorithms, in which a particular data set is typically divided into two parts: Training and Test data sets with known class labels. Supervised algorithms generate classification models from training data and use learned models for prediction. This is to evaluate the performance of the classification model, where the model is applied to the test data, in order to obtain the classification accuracy. Also, typical supervised learning methods include: decision tree induction, Bayes classification, K-means, and vector support machines.

- Clustering: the common example of unsupervised learning algorithms, in which a model is constructed based on similarity or difference between data objects. Similarity or difference between data objects can be measured using measures of proximity including Euclidean distance, Minkowski distance, and Mahalanobis distance.

- Data mining includes other techniques such as correlation mining, anomaly detection, feature selection, instance selection, and visual analytics [5].

The general objective of the paper is to use data mining techniques to extract information that can be used to promote investment in the Kingdom of Saudi Arabia and provide job opportunities for young people through a huge set of data, which can be extracted from user platforms on social media in the Kingdom of Saudi Arabia and turned into an understandable and easy-to-use structure. The study is based on making at first: a comprehensive survey about everything related to investment, its concept, products, manufacturing or market. Second: a manual research study on social networking sites of that interested people in the field of investment will be done. Third: creating a dictionary that includes all the words and phrases that serve the search. Then, initial preparation of the program used to extract the information (based on the dictionary). Finally initiate the preparation of exploration techniques for analysis and conclusion. Smart investing and machine learning. Artificial intelligence, big data analytics, deep learning, and machine learning are increasingly being used to improve the customer experience in particular by getting more insights into their behavioral pattern. AI-assisted software can collect historical and geographical data on past products and customer needs and help officials draw conclusions to facilitate decision-making. Use the acquired data form to validate the tested data or target data. It is increasingly used in social media analysis applications to understand and then predict valuable data. Where descriptive data mining tasks describe the general properties of data, predictive data mining tasks make inferences based on the existing data set to predict how a new data set would behave. Clustering is a technique for identifying data items that are identical. It consists of an interconnected set of artificial neurons and information processes using a connection approach to

computation. From the deep learning family, it is classified as artificial neural networks, and is commonly used in speech recognition and natural language processing. RNNs are designed to recognize the characteristics of sequential data and use patterns to predict the next possible scenario.

## 2. REVIEW OF LITERATURES

#### 2.1. Social Media in Data Mining Review Paper [6]:

This paper is based on 3 techniques to extract data for analyzing feelings on Twitter. The paper reviews the concept of Media Platform and that it is a web-based application resource that allows for the structure and employment of customer-generated content. Throughout the paper, the author provided an examination of the background of SM and the challenges of data query techniques in presenting analysis in SM, and a discussion of mining techniques through various data in sentiment analysis. List of data extraction techniques currently used in sentiment analysis. While social networks are one of the most important networks used in data mining techniques. Unsupervised machine learning: where there is no teacher. Algorithms are left to their own innovations in discovering and presenting an interesting structure in data. An example of a compound application: clustering problems in k-mean. Correlation rules the learning problem in the apriority algorithm. The lexical resource is used for attitude analysis, it is a record of the lexical units of a language, the lexical units may be words, the senses of words, phrases, etc.

## 2.2. Overview on Data Mining in Social Media [7]:

This paper conveys several issues in data mining and also focuses on the scope of data mining in social media which will be useful in further research. The research cited in year 2017, when the number of social media users reached 2.34 billion, according to «statista.com» Due to this large number, huge amounts of data are available in various format. To make the use of this data more efficient, it requires a special tool to summarize, analyze, and extract information, and to discover pattern and correlation between data. The solution, above all, is «data mining». Data mining refers to «information mining» The purpose of a predictive model is to allow the data mine to predict an unknown value of a given variable. Finally, the author acknowledges that each field requires a different type of knowledge and uses a different data warehouse. Data in data warehouses is heterogeneous, disorganized, dynamic and noisy. The speed and volume of data is unrealistic. Several data mining algorithms are used to extract social media data. The most used algorithm is SVM. Nonlinear primary selection, training timing, storage, and uncertainty in data points are major issues in SVM.

# 2.3. A Study on Social Network Analysis through Data Mining Techniques – A detailed Survey [8]:

The author presented a set of important concepts in the study of social network analysis, using data mining. Data collection: Data were collected through the Sina API crawler and functions.

Noise Text Filter was performed through event correlation analysis, as irrelevant texts were filtered. The goal of event correlation analysis is to remove unwanted data and this is done in three steps:

1. Preprocessing: junk characters and URLs have been removed at this point.Correlation.\

- 2. Calculation: correlation is calculated by VSM, chain matching, Levenshtein distance and SimHash algorithms.
- 3. Filter: this is achieved when the link is less than 0.2 or when the text is no more than 8 characters long.

## 2.4. Survey of data mining techniques for social networking website [9]:

The author discussed various data mining techniques used in a social network. It talks about the privileges that social networks have granted to its users to access uncensored, easily available and never-ending information based and structural analysis.

1) Graph based K-means clustering: This approach is applied to social network analysis in order to identify important features of the network such as: nodes and links and it is composed of:

- Community detection using hierarchical clustering: The community detection method is able to assign node not only to one community but also to many communities.
- Semantic web of social network: The semantic web environment makes knowledge sharing and reuse possible across different applications and edges of society. The discovery of the evolution of the Semantic Web enhances knowledge of the importance of the Semantic Web community and emphasizes the synthesis of the Semantic Web.

2) Aspect based / Feature based opinion mining: Also known as feature-based analysis is the process of excavating the entity area that customers reviewed.

3) Sentiment analysis of social network: Sentiment analysis is also called opinion mining.

4) Classification of Social network data: An artificial neural network is a mathematical model, consisting of a group of processing units that communicate together by sending signals to each other over a large number of balanced connections. There are three types of learning methods: Learning, as the network is trained by providing it with matching inputs and outputs patterns - a drawing.

## 3. THE NEWLY PROPOSED SYSTEM

## **3.1. System Description:**

In this paper a tool will be created for searching and analyzing words on social media, helping investors to choose the best products and regions, in establishing or improving their projects in the Kingdom of Saudi Arabia. Through a search engine across social networks to collect all comments, reviews, and tweets with the user's search word. Then, through smart processing technology, to extract the true meanings of people's comments, define them, and classify them in terms of positive, negative or neutral, and thus know the majority of people who like or hate the required product, and if possible, to know their areas of presence and classify them according to the regions of the Kingdom. This research will be started with Twitter, it will be the main source of data for the prototype, and then integration will be required to include more sources such as: Facebook and other well-known social networking pages.

## **3.2. Flow Chart:**

Flow chart clarify all the components of the process in the project and its multiple stages and the progress of work steps from beginning to end, and also help in determining the actual path of the process, and identifying the recurring operations and inspection stages in order to raise the degree of efficiency of the various processes as shown in figure 1.



Figure 1. Flow Chart

## **3.3. Use Case Diagram:**

The use case diagram in figure 2 shows the relationship between actors and use cases. The use case project the Actor User, creates a search based on the required keywords during the use case Tweet Search, thus accessing Connect and Extract Tool Twitter. In order to get the Tweet links, it is needed to access Twitter API Authorization Which redirects get Link response, from Twitter

then use case " And Extract Tool Twitter ". Tweets are removed and stored back in the database by "Store data in database".



Figure 2. Use Case Diagram

## **3.4. Sequence Diagram:**

Sequence diagram represents the interaction between objects in sequential order, that is, the order in which these interactions take place. A sequence diagram may also be referred to as an event diagram or an event scenario. Sequence diagrams show how and in what order various objects in a system work. The sequence diagram is an important part for software developers to document and understand the requirements of new and existing systems [11]. Figure 3 shown a project progress steps starting with User, where creates Search Twitter through Twitter and then Store Data in the database, Data is called from Database and send Data to train through Train Data, the data is also called from the Database to the Test Data, the results are sent from both Train Data and Test Data to the Feature Extract, then the results are sent to the Classification who in turn sends the results to the User.



International Journal of Computer Science & Information Technology (IJCSIT) Vol 14, No 5, October 2022

Figure 3. Sequence diagram for project

## 3.5. Class Diagram:

Class diagram (classes) is a very important part of the UML, and it is an organizational diagram that aims to display classes in a specific system with all the relationships that link them, and it is - in my opinion - the most popular type of diagram in software engineering [12].

The project includes the following main categories shown the figure 4:

- TiwtterSearch: Through it, it receives the user's words, to conduct the search.
- PreProcessing: This category contains a set of procedures. Such as: the Tokenizer, the stopword deletion procedure, and the data-storing procedure.
- Databse\_help: Is responsible for sending to the database, reviewing, deleting, and more Tweets API Collection, responsible for managing Tweets in Twitter.
- Extracting\_influential\_members: Responsible for getting influential members in Tweets who are the main Tweet owners or first tweeters.
- Feature\_Extract\_Topic: Responsible for extracting requested threads by word.
- Classifier: Responsible for classifying tweets according to search words in terms of positive, negative or neutral.
- Predict\_check: Tests the final results and the extent to which the goal is achieved.
- Result Display: Responsible for displaying the results.



Figure 4. Class Diagram

## 4. IMPLEMENTATION AND TESTING

In this section, will describe software used in the development process and present code fragments for the main functions of the system.

## 4.1. Project tools:

The developer's software requirement is a description of the software necessary before implementing the proposed framework. The system is analyzed keeping in mind the use of tools that provide a simple user-friendly interface for the user for effective use, also providing accurate insights for opinion mining.

- Programming Language: Python.
- Integrated Development Environment (IDE): Spyder.
- Application Programming Interface (API): Twitter Search API.
- Package Manager: Anaconda.
- Libraries/Toolkits: Tweepy, Sci-kit learn, Matplotlib and NLTK.

#### 4.2. Prepare Pre-data sources:

1. Twitter's tweet file: It includes approximately 10100 tweets based on people's feelings towards investment and commercial activity and the search for various job opportunities. All tweets were classified according to their impact (positive - negative - neutral), which was done by us.

2. Investment Keyword Dictionary: A file has been created containing all the words related to investing and are stored in a file, which will be used to search for the tweets that made up the previous file.

3. Dictionary of stop words: This file is used to delete unwanted words that do not serve tweets, such as prepositions and some irregular words and other words related to the Saudi dialect.

4. Ugly or Bad Words Dictionary: This file is used to delete bad and obscene tweets and contrary to the morals of society.

#### 4.3. Implementation Steps:

#### 4.3.1. Data Collection:

Reviews used in this project were collected from Twitter.com. It pertains to the field of investment in Saudi Arabia, where the rating is made on live audit data, as shown in figure 5.

Figure 5. Data collection

## **4.3.2. Pre-processing of Data: 4.3.2.1. Turn off word filter:**

NLTK includes a list of stop words in English but an Arabic dictionary is used to filter out the words, stop words have been removed from the reviews, since the stop words don't have any meanings that are useful for analysis, so they are never useful for classification, as shown in figure 6.

```
def StopWords(filePath):
    stop_words=[]
    f=open(filePath,'r',encoding="utf-8")
    for i in f:
        i=i.strip()
        stop_words.append(i)
    return stop_words

def removeStopWords(self,tweet,stop_words):
    new_tweet=''
    for word in tweet.split():
        if word not in stop_words:
            new_tweet+=word+' '
    return new_tweet
```



#### 4.3.2.2. Remove bad word:

Since there are tweets containing bad and offensive words, that will be deleted through the bad word dictionary, as shown in figure 7.

```
def UndesirableWords(self,filePath):
    undesirable_words=[]
    f=open(filePath,'r')
    for i in f:
        i=i.strip()
        undesirable_words.append(i)
    return undesirable_words
def removeTweetsUndesirable(self,tweet,undesirable_words):
    result=False
    for word in tweet.split():
        if word in undesirable_words:
            result=True
    return result
```

Figure 7. Remove bad word

#### 4.3.2.3. Delete links, "#", "\_" and "@":

Links and private tags are among the obstacles to understand, analyze and delete tweet text, as shown in figure 8.

Have 4 preprocessing:

- Any link replaced by the word link.
- Any @ replaced by the word Mention.
- Delete all repeated characters such as: Heeelllllooo.
- •Any Underscore \_ replaces by " ".
- Any Hashtaq # replaces by " ".

```
def preProcessing(self,tweet):
    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','ueet)
    tweet = re.sub('@[^\s]+','uimu',tweet)
    tweet=self.removeDuplicates(tweet)
    tweet = tweet.replace("_", " ")
    tweet = tweet.replace("#", " ")
```

Figure 8. Delete links, #, \_ and @

#### 4.3.3. SVM classification:

The support vector machine is used because it is a binary classifier, so it works best for classifying reviews into positive and negative category. The classified data is divided into two parts for training and testing. Different SVMs have different kernels, Linear SVM Kernel is used in this work, as shown in figure 9 and figure 10.

```
def Training_testData():
        X = []
y = []
        path = os.getcwd()
f=open(path + '/output.csv',encoding='utf-8')
reader = csv.reader(f)
             for row in reader:
             rint("row[1])
print("row[1]: ",
if (row[0]=='1'):
             y.append(2)
        X_train, X_test, y_train, y_test = sklearn.model_selection.train_test_split(X,y,test_size=0.20, random_state=10)
return X_train, X_test, y_train, y_test
print("X_train : ",X_train)
print("y_train : ",y_train)
                                                         Figure 9. Train data
 def classifier(X_train,y_train):
           vec = TfidfVectorizer(min df=5, max df=0.95, sublinear tf = True,use idf = True,ngram range=(1, 2))
           svm_clf =svm.LinearSVC(C=0.1)
           vec_clf = Pipeline([('vectorizer', vec), ('LSVC', svm_clf)])
           vec_clf.fit(X_train,y_train)
           joblib.dump(vec clf, 'svmClassifier.pkl', compress=True)
```

Figure 10. Classifier data

#### 4.3.4. Prediction:

return vec clf

It is an approach to parse tweet texts based on taxonomy preset by SVM taxonomy algorithms, containing words with their sentiment scores. In this work, the SVM classifier file is used. This file gives three types of positive, negative and neutral scores, as shown in figure 11.

```
def predict(self, tweet,classifier):
    X = [tweet]
    opinion = classifier.predict(X)
    return (opinion[0])
```

Figure 11. Prediction

#### 4.3.5. User interface:

The newly proposed system interface is a web page based design, to make it easier for users to interact with it, and to get results in an easy and smooth manner. It will be divided it into two pages, the first for query and the second for showing results. Figure 12 shows the user interface.



Figure 12. User interface

#### 4.4. Testing and Result:

The test stage is the most important phase of the proposed system. It allows to see if the system as a whole is working or not and whether all the systems functions are working properly.

#### 4.4.1. Accuracy Measure:

To calculate accuracy, the metrics. accuracy\_score module is used in Python. The results shown in figure 13 display the accuracy of the approach (SVM), which gives high results of up to 93.5% and this is a satisfactory result through the project.

	precision	recall	f1-score	support	
e	0.00	0.00	0.00	7	
1	0.92	0.54	0.68	244	
2	0.94	0.99	0.96	1769	
micro avg	0.94	0.94	0.94	2020	
macro avg	0.62	0.51	0.55	2020	
weighted ave	0.93	0.94	0.93	2020	
SVM accuracy	0.935148514	8514852	-		

Figure 13. Accuracy Measure

#### 4.4.2. Classification Report:

To calculate other metrics such as precision, recall, f1-score and support, the Python classification\_report unit is used. SVM has shown excellent results as shown in figure 14.

support	f1-score	recall	precision		
7	0.00	0.00	0.00	0	
244	0.68	0.54	0.92	1	
1769	0.96	0.99	0.94	2	
2020	0.94	0.94	0.94	micro avg	micro
2020	0.55	0.51	0.62	macro avg	macro
2020	0.93	0.94	0.93	ghted avg	weighted

figure 14. Classification Report

### 4.4.3. Example:

## 4.4.3.1Query for " best stock ":

Figure 15 shows the entries, where the number of results of 298 tweets was recorded, of which 296 are neutral tweets, 2 are positive tweets and 0 negative tweets.



Figure 15. Analysis results

• The number of tweets obtained from the query process is shown in figure 16.





• A graph showing the percentages of results is shown in figure 17.



Figure 17. percentages of results

• Viewing tweets and rating each tweet (positive-negative-neutral)is shown in figure 18.



Figure 18. Viewing tweets and rating (positive-negative-neutral)

#### 4.4.4. Result and Analytics:

The output of the system about the opinions of tweeters regarding investment in products, employment, projects, goods, etc. The system displays a graph describing opinions measured over time and a pie chart that shows the percentage of positive, negative and neutral tweets in the total number of tweets, as well as the most important tweets. This app has a web page interface already implemented for the front end of tweet parsing. Depending on the decision made in the classification section, this interface allows to perform several queries based on the words entered by the user, and perform many queries and analysis until reaching the satisfactory results for him in the field he wants.

## **5.** COST ANALYSIS

### 5.1. Research Requirements and Budget:

In terms of requirements, Anaconda Navigator is the main software used for implementing the newly proposed system, Server: XAMPP, in addition to laptop, HyperText Markup Language (HTML), Cascading Style Sheets (CSS) and Spyder. Table 1 shows the detailed information about the cost of the project.

Requirement	Estimated Price
Laptop	2,000 USD
Anaconda	Free
Server (XAMPP)	Free
HyperText Markup Language (HTML)	Free
Cascading Style Sheets (CSS)	Free
Spyder	Free
Total	2,000 USD

Table 1. Requirements Cos	Table	1. Re	quirements	Cost
---------------------------	-------	-------	------------	------

Table 2. Time Plan

## **6.** CONCLUSION AND FUTURE WORK

This paper is introduced to solve the problem ofhow to provide the necessary information to increase investment opportunities within the Kingdom of Saudi Arabia, by studying and analyzing what people occupy on the communication sites through their tweets about the labor market and investment. A set of research papers had been discussed in this paper. A newly proposed system is explained to analyze the tweets of users on twitter in the field of investment. The system provides necessary and clear information to investors that helps them to establish special projects based on offering a special dictionary for investment in tweets written in the Saudi Arabian dialect and Arabic language. The proposed system depends on using support vector machine which is the most common algorithms in the field of word analysis on social media to improve sentiments Analyzing. The system displays a graph describing opinions measured over time and a pie chart that shows the percentage of positive, negative and neutral tweets in the total number of tweets, as well as the most important tweets. After testing, the results showed that the neutral class is more than positive and negative class due to the fact that

the data set was created by the authors and not by companies or specialized agencies, because this task will takes a lot of time and effort. In the future will enhance the data set dictionary, build English dictionary and collect the data set at different periods with trend topics.

#### REFERENCES

- [1] J. Han, M. Kamber, J. Pei, "Data Mining Concepts And Techniques", Diego, 2011.
- [2] Daniel T. Larose, "Discovering Knowledge In Data An Introduction To Data Mining", 2005.
- [3] R. Zafarani, M.A. Abbasi, And H. Liu, "Social Media Mining: An Introduction", Cambridge University Press, 16.
- [4] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, 2011.
- [5] P.N. Tan, M. Steinbach And V. Kumar, "Introduction To Data Mining", Pearson Addison Wesley, Boston, 2006.
- [6] Prabhjot Kaura, Anu Goyal, Kavita Sharma, Pritia, Anupriya Jain, Prasenjit Banerjee, "Social Media In Data Mining Review Paper", Manav Rachna International University, Faridabad, India, 2018.
- [7] Amali Pushpam, Gnana Jayanthi Joseph, "Over View On Data Mining In Social Media", International Journal Of Computer Sciences And Engineering, November 2017.
- [8] Annie Syrien, M.Hanumanthappa, "A Study On Social Network Analysis Through Data Mining Techniques – A Detailed Survey", International Journal Of Advanced Research In Computer And Communication Engineering, Vol. 5, Special Issue 2, October 2016.
- [9] D.Kavitha, Mca, M.Phi., "Survey Of Data Mining Techniques For Social Networking Websites", Assistant Professor, Kg College Of Arts And Science, Coimbatore, Ijcsmc, Vol. 6, Issue. 4, April 2017, Pg.418 – 426.
- [10] Flow Chart Definition, Https://Www.Edrawsoft.Com/Flowchart-Definition.Html?Gclid=Cj0kcqjwjpacbhdkarisaiszn7qrrt9bx9cakidk2mpap2mfkgs8x4bfgw-Av3mqvcfdcdn7ctb720kaagjfealw Wcb,2020.
- [11] Unified Modeling Language (Uml) | SequenceDiagrams,Https://Www.Geeksforgeeks.Org/Unified-Modeling-Language-Uml-Sequence-Diagrams/.
- [12] What Is Class Diagram?, Https://Www.Visual-Paradigm.Com/Guide/Uml-Unified-Modeling-Language/What-Is-Class-Diagram/.