IRAQI JOURNAL FOR COMPUTERS AND INFORMATICS
Vol. [46], Issue [1], Year (2020)

PAAD: POLITICAL ARABIC ARTICLES DATASET FOR AUTOMATIC TEXT CATEGORIZATION

Dhafar Hamed Abd [1]
[1]*Department of Computer Science University of Technology, Baghdad, IRAQ*
Dhafar.dhafar@gmail.com

Ahmed T. Sadiq [2]
[2]*Department of Computer Science Al-Maarif University College, Al anbar, IRAQ*
Drahmaed_tark@yahoo.com

Ayad R. Abbas [3]
[3]*Department of Computer Science University of Technology, Baghdad,, IRAQ*
ayad_cs@yahoo.com

Abstract **- Now day's text Classification and Sentiment analysis is considered as one of the popular Natural Language Processing (NLP) tasks. This kind of technique plays significant role in human activities and has impact on the daily behaviours. Each article in different fields such as politics and business represent different opinions according to the writer tendency. A huge amount of data will be acquired through that differentiation. The capability to manage the political orientation of an online article automatically. Therefore, there is no corpus for political categorization was directed towards this task in Arabic, due to the lack of rich representative resources for training an Arabic text classifier. However, we introduce political Arabic articles dataset (PAAD) of textual data collected from newspapers, social network, general forum and ideology website. The dataset is 206 articles distributed into three categories as (Reform, Conservative and Revolutionary) that we offer to the research community on Arabic computational linguistics. We anticipate that this dataset would make a great aid for a variety of NLP tasks on Modern Standard Arabic, political text classification purposes. We present the data in raw form and excel file. Excel file will be in four types such as V1 raw data, V2 preprocessing, V3 root stemming and V4 light stemming.**

Keywords*: Arabic Political Article; Orientation; Sentiment Analysis; Natural language Processing; Opinion Mining.*

I. INTRODUCTION

Nowadays more and more Arabic political is presented in the form of electronics. Arabic political inform us about the current event and explain situations, political articles include education, war, politics, economy and so on. Political articles also records people's opinion and action objectivity from our situations and Ariba countries have many situations as Iraq, Syria and Lebanon and so on. So, this area will reach for Arabic political articles and we have to make analysis.

The Arabic Language is considered one of the most powerful and effective language around the world. According to the latest report, it is the 5th widely used languages around the globe. There are roughly 422 million people as a mother tongue and 250 million an additional language [1]. Arabic alphabet consists of 28 letters. The orientation of writing in Arabic is from right to left [2]. Compared to the high number of studies conducted on English sentiments, opinions, attitudes, and emotions, the number of similar studies on the Arabic language is very small.

In this paper will present four corpus such as V1, V2, V3, and V4 for each corpus will create process then make these corpora available online for who interest in NLP filed. For each corpus will using process except corpus V1 this original corpus but another will make process such as preprocessing, root stemming and light stemming. For these corpora will make statistical analysis for each one by using statistical analysis. Statistical analysis is considered very effective tools to determine any research work. This kind of methods can be applied for a number of research studies that propose some techniques for carrying out comparisons among various algorithms [3-5].

In this research paper, the data analysis applications are concentrated on machine learning models that association with the exploration of the statistical properties of the variables. This can be obtained by plotting the scatter of data, which summarize each class for visualization. In order to obtain the scatter visualization by using To undertake an visualization of the utilised data in these experiments, it computed with summary statistics, followed by visualisation methods including t-distributed Stochastic Neighbourhood Embedding (t-SNE) and Principal Component Analysis (PCA) [6]. Although PCA and t-SNE are powerful in most of advanced statistical methods, they have the great advantage of making distribution of the examined data. This research is presented scatter plots for each input variable and three outputs.

II. LITERATURE REVIEW

Annotation is the process of manually or automatically adding information into text for a given purpose. In computational linguistics, humans called annotators or taggers [7]. Annotation, and its companion activity of corpus creation, has become an important activity in computational linguistics since the widespread application of machine learning algorithms and lexicon. One of the most important preconditions for building an efficient model is to understand the input data characteristics. The specialized models always have potential to outperform the general one. The real contribution of this work is the exploration of characteristics in

the data and its application on the process of building a classification model. For languages using corpus in different language for political as table below.

TABLE 1
SUMMARY OF POLITICAL CORPUS IN DIFFERENT LANGUAGE FROM 2015 TO 2020 EXCEPT ARABIC LANGUAGE.

Reference	year	language	Source	Description	public
Chambers [8]	2015	English	Tweeter	Identifying political relations between nation states	N/A
Du et al [9]	2015	Russian and Chinese	News papers	Extraction sentence from Russian and Chinese news	N/A
Rubtcova [10]	2015	Russian	News papers	The corpus collected from newspapers is used for context of political institutions	N/A
Stranisci [11]	2016	Italian	Tweeter	Online irony debate for Italian political	N/A
Burnap et al [12]	2016	English	Tweeter	Predict UK president in 2015	N/A
Ahmed et al [13]	2016	English and Indian	Tweeter	Indian political election on 2014	N/A
Rashkin et al [14]	2017	English	Newspapers	Analyzing language in fake news and political	N/A
De Smedt [15]	2018	Germany	Tweeter	German political debate	N/A
Stier et al [16]	2018	German	Tweeter and Facebook	Online campaigning of German politicians	N/A
Lai [17]	2018	Italian	Tweeter	Political debate	N/A
Hu [18]	2019	Chine	Newspaper	Chinese democracy strategy in political language	N/A
Mehta and Kavi [19]	2019	English	Newspaper and TV web site	Categorize ideological frames	N/A
Falck et al [20]	2019	English	Newspaper	Measuring between newspapers and political parties	N/A

There are many papers published for political the most of them collect from tweeter [21-23] and few from newspaper [24]. In table above as we can see from different natural language most of researchers using tweeter because ease to collect text because of API [25-27]. In this case we search from 2015 to 2020 for Arabic corpus as we can below in table 2 the years and Arabic corpus with purpose and it publish or not also the source with type of the corpus.

TABLE 2
SHOW THE RELATED WORK WITH ARABIC LANGUAGE CORPS

Author	Year	Source	Description	public
Elsayed et al [28]	2015	News	This corpus building from news from al-youm7	N/A
Ahmed et al [29]	2017	Website king Abdullah	Political parallel corpus for Arabic and English	N/A
Chouigui et al [30]	2018	Tunisian web radio sit	This corpus for Arabic news	ANT[1]
Abooraig et al [31]	2018	Website, social network	This corpus collects from different place and using for classifying political orientation (Arab nationalist, brotherhood, Islamic Shia, liberal and socialist)	N/A
Zeroual et al [32]	2019	News websites	This corpus for Arabic news	CLARIN[2]

In this survey by zaghouani and wajdi [33] all the corpus published none of them with political also the survey for Arabic corpus from 2005 to 2015 by Al-thubaity [34] none of them political corpus. As we can see above there in no corpus for political as we will work for Arabic and another language from 2015 to 2020 also the close paper and also political by Abooraig et al but the category different and also not available online that's make as to prepper our corpus and special with M,S and T because each one will solve problem for human right now.

[1] https://antcorpus.github.io
[2] https://www.clarin.eu

Some problems within Arabic corpus become apparent when examining the previous work. The simple reason that much work is being done on non-public corpus. From the comparison shown in tables 1 it can be seen that of the evaluated 13 corpora none of them publish. In table 2 there are 2 from 5 corpus online. The corpora used also came from several different domains, including newspaper and tweeter.

III. METHODOLOGY

The important unistructural data so have to collect then have to clean and change to structural data steps below show that in figure 1.

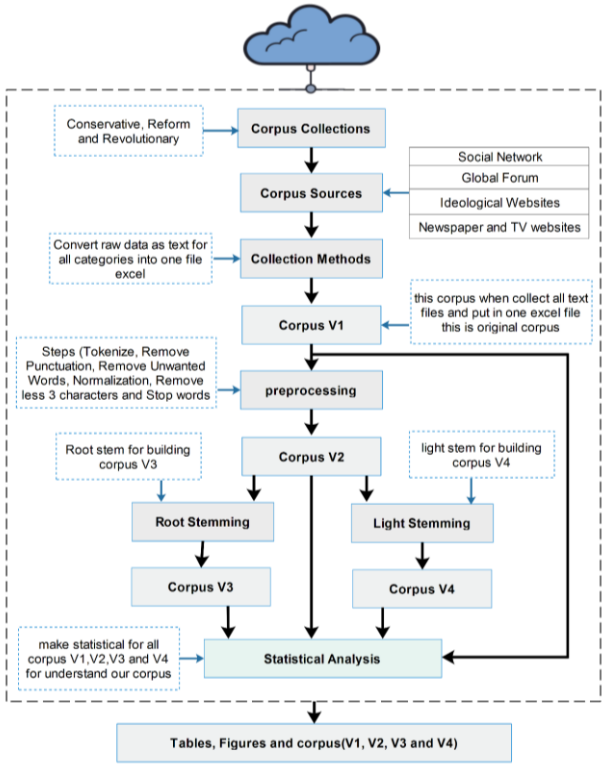

Fig 1: General Model

A. Corpus Collection

There is a crucial challenge during the search on internet for the Arabic articles. This challenge includes the absence of a standard Arabic dataset, so had collected and construct own dataset used in this study. collect 206 political Arabic articles called (PAAD). The dataset is collected from different places such as (social network, global forum, website ideologies, newspaper and TV). PAAD was distributed among three categories such as (conservative, reform, and revolutionary). PAAD has a total number of 206 articles. These categories are summarized in table 3. This corpus available free on Mandalay repository3.

[3] https://data.mendeley.com/datasets/spvbf5bgjs.

TABLE 3
NUMBER OF ARTICLES WITH LABEL

Label of articles	Articles number	Arabic meaning
Conservative	28.2% (58)	التيار المحافظ
Reform	38.8% (80)	التيار الاصلاحي
Revolutionary	33.0% (68)	التيار الثوري
Total	206	

As shown in Table 4, there are three types of articles (short, medium and large).

TABLE 4
Types of articles

Articles	Number of articles	percentage
shortest article	15	7.281%
Medium article	36	17.475%
longest article	155	75.242%
Total	206	100%

The gathering of online articles was based on Modern Standard Arabic. The vast majorities of modern Articles that written in the colloquial Arabic were excluded. A number of different sources was considered to build the corpus manually. There are roughly 60% of our corpus gathered was during Arab Spring revolution, while the 30% for the period preceding the Arab Spring revolution. The remaining percentage 10% is specific for beliefs and ideas that founded the party thought and for the political events.

B. Corpus sources

In this section, we discuss the procedure of our dataset collection resources.

1) Social network: Social networks are considered very effective tools for spreading news and information at any area around the world. Facebook4 was used for crowd gathering during the Arab Spring revolution. There are a number of people are using these platforms to express their support for to the revolution, while other used to express their opposing views. The main purpose of this experimental study is to collect these posts and understanding ideologies and trends in the Arab world.

2) Global forums: The global forums are forums that do not require from a user to be of the same ideology as the forum members, to be a member. We chose global because specific ideological forums due to any member from other ideologies. We also considered collecting the writings of users in global forums source of our dataset. In addition to that, Arab policy forum5 is an example of these forums.

3) Ideological websites: There are a number of ideological websites with various opinions. This kind of website always discuss political events throughout distribution articles or news. The admin of this website is frequently having the same ideology. Eventually, in order to collect articles from ideological websites, it should contain articles of party members.

[4] https://www.facebook.com
[5] https:// www.alsiasat.com

4) TV websites and Newspaper: TV websites and newspaper have corners page that is specified in the articles. We gathered a number of online articles from such sources. Here we will mention to these sources such as BBC6. Table 5 shows the numbers of articles from various sources

TABLE 5
SOURCE OF OUR DATASET

Source	Numbers of articles	percentage
Social network	21	10.194%
Global forum	38	18.446%
Ideological website	93	45.145%
Newspaper and TV	54	26.213%
Total	206	100%

C. Collection Method

The collected data is formatted in three folders (Reform, Conservative and Revolutionary). Each folder involves a list of text files numbered sequentially, in which a file corresponds to one whole article. The articles contain some English symbols, punctuation, digits, and Arabic diacritics as figure 2. There are five important steps that used in our study for collection data. Firstly, we extract and collect online Arabic articles manually using various internet sources. Secondly, it is very significant to remove the unwanted components. Thirdly, correct words that comprise bad mistakes or spelling. Fourthly, place every article in one folder, and then collects the documents that belong to the same ideology in one folder. Finally, raw data will be available for preprocessing.


Reform\25.txt

إن أي عملية اصلاح ، هو تحقيق أقصى قدر من المنافع والتقليل من آثارها السلبية على العمل و الانحراف عن المسار ويعني الانتقال من مرحلة أو حالة غير منتجة إلى مرحلة أو حالة أخرى منتجة وفاعلة وبمعالم اكثر وضوح يفترض فيها أن تكون أكثر ايجابية و يتطلب جهوداً متواصلة ومضنية من المنظمات المدنية ومن المديرين العاملين في التخطيط لعمليات الاصلاح ومجابهة ردود الأفعال الناجمة عنها. من هذا المنطلق يتطلب اتخاذ خطوات فعالة للتعامل مع هذه التغييرات بشكل دائم وتحديد وتصميم الاستراتيجيات المناسبة لنجاح العملية وتجنب فشل الاصلاح أو لإيجاد حل مناسب لأي اضطراب في التنفيذ خاصة إذا كان هدفها البناء و لا يمكن أن تكون إلا بالوسائل الصحيحة وتديرها مجموعات سليمة بعيدة عن المصالح الضيقة والكفيلة بتأمين الاجماع وتكون الاعتراض حول خطوات التغيير اولاً بأول، وتحيطها بحزام الأمان من أي اختراق خارجي ، والا كانت العمليات فاشلة ، بل ودرسا في الضحك على الذقون . بناء المؤسسات ليست لعبة إلكترونية تجنى منها المكاسب الفورية يقوم بها الجالسون في زوايا البيوت وادارتها بالهواتف النقالة او الاجهزة الحديثة ، انما بالممارسة العملية والتواجد في ساحة العمل للتفكير الجدي في المحافظة على الاستمرار بالشراكه في المستقبل القريب والبعيد بحيث كل ما كان هناك عاملين كفء ومحبين للعمل كل ما كان تحقيق الاهداف بسهوله والوصول لتحقيق المكاسب اسرع وبوقت اقل وللموارد البشريه دور كبير في مجال التقييم وتحسين الاداء ودراسه اماكن الضعف والعمل على حلها وتستطيع تحليل الوظائف وتصميمها في استقطاب الكفاءات واختيارها وتعيينها في إدارة الأداء و المزايا والتعويضات وبدلات تطوير الموارد البشرية ونظام تحفيز العاملين الى جانب تخطيط الموارد البشرية وفي وضع الصلاحيات والمسؤوليات مع تحديث الهياكل التنظيمية لوضع أنظمة السلامة و دراسة مشاكل العاملين ومعالجتها بهدف ادارة الموارد البشرية الاستراتجية وتحويلها لتوظيف المهارات والكفاءات العالية التدريب والمتحفزة ماديه ومعنويه وهي اساليب وخطط مدروسة بفعل ممارسات على مستويات متعددة تحتاج في التنفيذ الى السهر والحرص ودأب وعرق ونزيف في أحيانا كثيرة.

Fig 2: Example of reform political article

D. Building corpus

Despite of the labelling process is straightforward and simple; it is crucial part of data preprocessing in Machine learning algorithms. The labelling process requires to be carried out precisely, as any error can render the quality of the dataset, and the general accuracy performance for various kind of machine learning approaches. The articles were collected using excel file and Python scripts written specifically for scraping three class. So, we read the folder from the excel file and read each article as shown figure 3.

Articles	lable
والمحافظين عادة أناس كئيبون ومبنيون، ولكن مع ذلك تكيفوا مع الزمن وأهله، ووضعوا أولويات المجتمع في مقدمة اهتمامهم : الاقتصاد والتربية والتكوين والصحة وتنمية المجتمع ...	Conservative
عرف الحزب في القرن السابع عشر باسم "توريز" والشتير بتأييده للملكية ومعارضته لتوجهات غريمه السياسي في جنبها حزب "ويغز" والذي هيمن على الحياة السياسية البريطانية	Conservative
قد تمسك نصف أعضاء الحزب في جنبها بمبدأ حرية التجارة، فيما طالب النصف الآخر بحماية مصالح المزارعين. Image caption الكتب ويشون تتوشل صفة الفلاح الوطني خلا	Conservative
الالتزام بالحكم الدستور ، واحترام سيادة القانون . المحافظة على استقرار الوطن وأمنه، وصون الوحدة الوطنية وعدم التمييز بين المواطنين . الالتزام بأسس الديمقراطية واحترام التعددية ال	Conservative
ن المحافظين أحد الأحزاب السياسية القوية المؤثرة في المجتمع السياسي المصري والتي تهتم بالشأن المصري على عدة محاور أهمها تأهيل وتمكين الشباب . وأوضح صراة بأن إختياره	Conservative
تكوم قلسفة حزب المحافظين على تضامن مجموعة من الأعضاء المؤمنين بمبادئ الحزب ويسعون لتحقيق أهدافه بطريقة مسؤولة من خلال رؤية لمصر بمثابه عقد اجتماعي ودستور لأعا	Conservative
رى الحزب أن أخطر ما يواجهه المجتمع المصري الآن هو اختلال منظومة القيم الإخلاقية، كنتيجة مباشرة لطول عهده بالفساد . ويؤمن الحزب إن العيشه الصالحة لا تكون إلا في دولة م	Conservative
بدا مفاجأ للبعض أن يعلن وزير الخزانة البريطاني فيليب هاموند، عن إطلاقه موازنته السنوية، عن انتهاء العمل بالتقشف المالي وبالتالي عن رصد مليارات الجنيهات لتعويز الضمان ال	Conservative
قد بات واضحاً ليس في بريطانيا فحسب، بل في صوء أوروبا وفي العالم، أن اتساع الفوارت الإقتصادي بسبب بارز وراء الصعود الشعبوي والقومي المتطرف . أما في بريطانيا تحديداً،	Conservative
للأحزاب السياسية المحافظة في الدول الغربية مواقف متباينة في مجال السياسة والإقتصاد والمجتمع والتربية والتعليم، ونذكر هنا على سبيل المثال حزب المحافظين في بريطانيا، والحزب	Conservative
ن اكثر المجتمعات و الشعوب التي يمكن ان تمور فيها الإفكار و الاعتقادات المحافظة هم من يلتصقون بالمثاليات و الخيال و الروحانيات و الإساطير و الخرافات ، و التجدد صعب فيهم	Conservative
وع من الفلسفة السياسية ينهتم بالمحافظة على القيم التقليدية الثابتة في المجتمع و هيا بشكل عام ضد التغيير ، وطبعا بلاد العالم متنوعة و مختلفه ، لكن رغم الإختلافات دي الا ان المحاف	Conservative
كثيراً ما يوصف الإسلاميون بأنهم تيار " محافظ " والمتدينون بأنهم " محافظون " ، وبعضهم يعتز بكونه " محافظاً " ويراه وصفاً إيجابياً يدل على الإستقامة والأصالة . فهل هم فعلا كذلك؟	Conservative
إذا كان جاذبية الفكر والسياسة المتبذلة في أمريكا تسعى إلى التكيف مع مستجدات الحياة، نجد بأن الفكر المحافظ والليبرالي قدم الكثير من المنتج الفكري في المجال السياسي والإجتماعي و	Conservative
هذه النزعة نحو ماضوية الفكر لم تقتصر على المحافظين، وانما شملت الليبراليين الذين ذهبوا أكثر نحو إعتماد الفكر اليساري، حيث يرون بأن الحكومة يجب أن تتدخل لصالح الإنسان (	Conservative
تكيد استمرارية الإصلاح السياسي والثقافي في مختلف الجوانب ومنها تأكيد سيادة الحكم الجيد وإعلاء قيم الشفافية والمحاسبة والمسئولية، وتحقيق دولة المؤسسات على أرض الواقع، حر	Conservative
تكوم قلسفة حزب المحافظين على تضامن مجموعة من الأعضاء المؤمنين بمبادئ الحزب ويسعون لتحقيق أهدافه بطريقة مسؤولة من خلال رؤية لمصر بمثابه عقد اجتماعي ودستور لأعا	Conservative
عُرف حزب المحافظين بمنهجه الليبرالي المحافظ سياسيا واقتصاديا، وتمسكه بالنظام الملكي في بريطانيا، ورفضه أي تغيير سياسي أو اجت	Conservative

Fig 3: excel file example for label articles

when building excel file, we get our dataset so have to build four datasets as table 6 below. will present the methodology used for building the dataset. The corpus V1 dataset is raw data but put in Excel. From V1 will built three new versions table 6 shows the various dataset.

TABLE 6
show the dataset we build

Corpus	Description
Original V1	It is the version that is built
Pre-processing V2	It is the version that is built with pre-processing
ISRI stemmer V3	It is stemming Appling on V2
Light stemmer V4	It is light stemming Appling on V2

1) Preprocessing for building corpus V2

In this step after building V1 then will building V2 for make clean from unwanted words, noise words and stop words. The V2 is important for see the effective preprocessing on text classification. Corpus V1 need to be cleaned and prepared the text for further classification. The online articles comprise a number of uninformative and noise data like HTML tags and scripts, or advertisements, which hinders the extraction of words. Furthermore, the presence of special Arabic characters or punctuation marks is also not accurately determined. Figure 4 shown the preprocessing steps.

[6] https://www.bbc.com/arabic

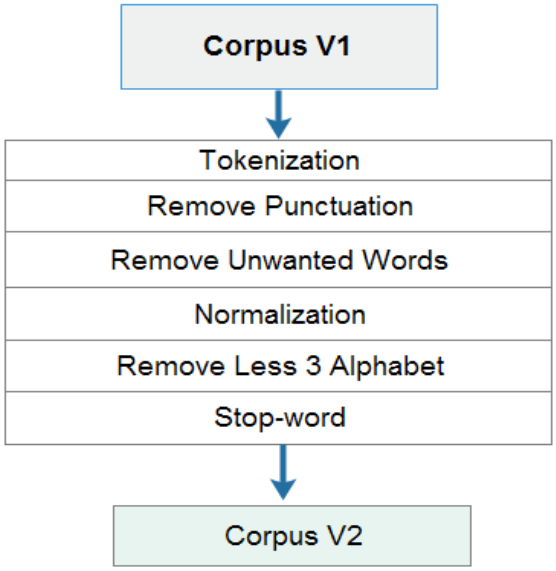

Fig 4: preprocessing steps

The first step is tokenization that will break our sequence of text into pieces such as words. So, the input of our tokenize become the output for another process. The second step is remove punctuation this step is very important because will remove all punctuation such as (:"'؛، - _ ; ? !؟ '{}/*//…[ ],). In third step will remove unwanted words that will show in table 7 by using list of regular expression.

TABLE 7
LIST OF REGULAR EXPRESSIONS

Regular Expression	Results
[a-zA-Z]+	Remove English characters
[0-9]+	Remove English numbers
[0-9]+	Remove Arabic numbers
#$%|@^~(&*)+	Remove another

The most important step in preprocessing for Arabic language is normalization. This important because Arabic language has many shapes for write so our methods involve three steps such as (Diacritics, Tatweel and Latter). Beginning with the removal of a diacritics from an Arabic word, and culminating in the conversion of an alphabetic word into another. An example of this process is displayed in figure 5. If we don't remove then we have many words for same words then vector become large and we need more articles for building then will take time example "تعبيراً" and "تعبيرا" so these words if we remove ("ً") then we have only one word.

Fig 5: The diacritics have to eliminate

Tatweel in Arabic language we can for same character make it long as (ـــــ) this use for make the shape of word pretty so this become problem because for same words we can write in many ways example "الاصـــــــــلاح", "الاصـــلاح" and "الاصلاح" its same word "الاصلاح". This become when press shaft and ("ت") so we have to remove this tatweel for make lass words. The final step in normalization will make unique letter for some letters. In Arabic language there are many ways for write characters as table 8.

TABLE 8
NORMALIZATION LATTER OF ARABIC LANGUAGE

Original	Become	Example	
إأآا	ا	ألاصلاح "reform"	الاصلاح "reform"
ى	ي	قوى "strong"	قوي "strong"
ؤ	ء	اداؤه "performance"	اداءه "performance"
ئ	ء	يتلائم "fits"	يتلاءم "fits"
ة	ه	ثورية " Revolutionary "	ثوريه " Revolutionary"

When complete all steps above then remove any word content only two characters then we can remove. because in Arabic language two words dos not make any sense. Final steps stop words the Arabic stop word can be filtered from an article by removing the token, and matching the word with the stop words listed. The stop words list is a built-in NLTK library. In NLTK all words are sign words.

E. Root Stemming for Building V3

In this section will building from V2 the root stemming corpus V3. Stemming [35] is very important because removal affixes and suffixes then will reduce the number if feature of words to the same stem generated. There are different stemming approaches proposed for many languages. Arabic language is very complexity language that is why need strong stemming to process its complexity morphological. In this study will using Information Science Research Institute's (ISRI) stemmer [36]. This algorithm has 7 steps such as the following algorithm for ISRI stemmer.

Algorithm 1: ISRI

1 **Input**
Remove diacritics which represent Arbic short vowels such as (ً,ْ)

Remove stop words ISRI have 60 stop words

Remove length three and two prefixes

2 Remove length three and two suffixes

Remove connecting و if its in begin word

Replace أ,إ,آ with ا

If 4<= length(word) and length(word)<=7 then

a. If length(word) == 4 then
Extract mutch the list patterns such as
فاعل فعول فعله

b. If length(word) == 5 then
Extract mutch the patterns list such as
تفاعل فعال تفعيل

c. If length(word) ==6 then
Extract mutch patterns list such as استفعل
مفعال افتعال

d. If length(word) == 7 then
Remove one character suffix and preffix if mutch

Pattern list such as ك ا ن ت ل

If length(word) == 6 then

Goto c step

3 **Output**

Corpus V3

F. Light stemming for building V4

From corpus V2 will using light stemming to build corpus V4. In light stemming all rules we have here taken from ISRI stemmer. Using light stemming if we looking for the meaning word and keep it clear for human to understand the word. ISRI Light stemmer, which is a number of conditions that determine how to apply the stemming on a certain word or not, the rules of removing suffixes, prefixes and waw. The table 9 show the three conditions with length of words that would be stem.

TABLE 9
RULES OF LIGHT STEMMING FOR WAW, PREFIX AND SUFFIX

Condition	Length of word	Remove from word	Letters
Waw	w >= 4	1	و
Prefix	p>=6	3	كال , بال , ولل , وال
	p>=5	2	لل , ال
Suffix	s>=6	3	تمل , همل , تان , كمل , تين
	s>=5	2	ون , ات , ان , ين , تن , كم , هن , نا , يا , ها , تم , كن, ني, وا , ما,هم

The conditions will be applying of the rule. The first apply waw and prefix finally suffix.

G. Statistical analysis

Statistical is very important for make analysis for data and see the data is good or not. In this case will using tables to show the frequency words belongs to each class and another thing. Also, will using visualization such as PCA, t-SNE and cloud word. Corpus visualization is very important step in the different approach in opinion mining, allowing the human advisor to gain an intuition of the data and the potential learn ability of such data. The results from corpus visualization can be used to guide the modelling phase, since a major component of learn ability is known to be a function of the correspondence between the different approaches and the type of representation it is supplied with. To undertake a visualization of the utilised data, it computed with summary statistics, followed by visualisation methods t-SNE and PCA. The main visualization analysis tools are discussed in the following sections.

1) *Principal component analysis (PCA)*

PCA can reduce the dimensionality of the data easily by discovering the orthogonal linear integrations from the original feature with the largest variance[37]. Given a sample of P observations on vector N variables to $\{x_1, \dots, x_p\} \in R^N$. For each observation, n dimensional vector representing the n features. The main purpose is to find the mapping form x to, where z is m dimension. In order to identify the initial principal component of the sample by the linear transformation in Equation 4 [38, 39].

$$z' = W^T x_j = \sum_{i=1}^{N} w_{i1} x_{ij} \quad j = 1,2,3, \dots, p. \tag{4}$$

Where the vector

$$w_1 = (w_{11}, w_{21}, w_{31}, \dots, w_{N1})$$

$$x_j = (x_{1j}, x_{2j}, x_{3j}, \dots, w_{Nj})$$

Var $[z_1]$ selected as maximum.

So, it is required to choose the feature where the variance of z_1 is maximum. The value of W^T for which projection that obtain correspondence to the largest variance of z_1. The principal component analysis is an effective process in terms of selecting a suitable number of features with accurate mapping dimensional space. In order to recover the original instances from the reduced presentation, the principal components are constructed error rate with minimum value[39].

2) *T-distributed Stochastic Neighborhood Embedding (T-SNE)*

A popular method for exploring high-dimensional data is something called t-SNE, introduced by Van and Hinton in 2008[40]. Exploratory analysis is considered vital step in the machine learning algorithms, permitting the humankind to acquire an intuition of the data and the potential learn ability of such data. In addition to that, the outcomes from data exploration can be utilized to monitor the modelling level, since a major component of learn ability is known to be a

function of the correspondence between the algorithms and the type of representation it is supplied with.

3) *Word Cloud*

In this paper will introduce words cloud that will visualization most frequency words in corpus also provide simple and effective meaning to visualize the most frequency words in corpus [41, 42].

IV. Experimental Result

In this result will using python language for all building corpus and test. Our test will be occurred in four steps because we have four corpus and for each corpus will make test.

A. Corpus V1

In this empirical study, we presented in details the statistical analysis and characteristic of V1. The characteristics are illustrated in Table 10. Revolutionary articles are the longest in terms of both the number of sentence and the number of digits, followed by Reform articles and then Conservative articles, whereas Conservative articles are the shortest. The same applies to the number of punctuations the Reform articles are the longest followed by Revolutionary articles and Conservative articles. Now, regarding the length of English words, Reform articles have the longest words, followed by Conservative articles and less English words for Revolutionary articles. For the number of unique words, the longest unique words are for Reform articles, then Revolutionary articles and Conservative articles have the less words. For the number of longest words, the longest words are for Reform articles, then Revolutionary articles and Conservative articles have the less words.

TABLE 10
SHOW THE DIFFERENT STATISTIC FOR EACH CLASS FOR CORPUS V1

statistical	Conservative	Reform	Revolutionary	Total
Number of articles	58	80	68	206
Number of sentences	511	738	873	2122
Highest article sentence	53	40	56	149
Lowest article sentence	1	3	1	5
Number of all token	14111	29607	23853	67571
Highest article token	626	1130	1208	2964
Lowest article token	54	82	63	199
Number of unique tokens	3246	6782	5303	15331
token occurs more than one time	1997	3319	2646	7962
Number of English words	65	110	7	182
Number of punctuations	616	1271	1054	2941
spatial character is dot.	370	586	705	1661
How many Digits	88	207	211	506
Word equal six alphabetic	1768	3687	3297	8752
Word less than six alphabetic	7327	14772	12157	34256
Word great than six alphabetic	3101	7289	5067	15457

In table above we have dot ('.') as spatial character this character very important for detection the sentence or split sentence in farther work. The figure 7 show the most frequent words in word cloud.

Fig 6: Most words in word cloud for corpus V1

As we can see in figure above the most frequency words present in cloud words are the commonly Arabic words. The table 11 below show the most 10 frequent words in PAAD corpus V1 for each class.

TABLE 11
MOST 10 FREQUENCY WORDS FOR CORPUS V1 FOR EACH CLASS

No	Conservative		Reform		Revolutionary	
1	في	418	في	782	في	789
2	.	370	من	620	.	705
3	من	306	.	586	من	538
4	على	249	على	384	على	355
5	أن	152	،	361	اليسار	234
6	أو	89	الاصلاح	285	الى	233
7	التي	88	الى	283	و	190
8	الى	82	أن	266	التي	187
9	``	81	أو	260	الثورة	162
10	اما	79	السياسي	233	أن	139

The table above present the high frequency punctuation. In the figure 7 will show articles distribution in corpus V1 for each class.

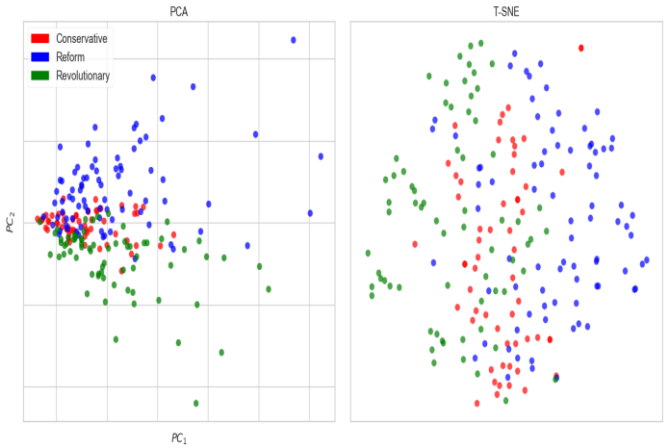

Fig 7: T-SNE and PCA visualization for corpus V1 show the articles distribution cluster view

As we can see in table above for both visualization PCA and T-SNE the articles distribution is non-linear. As we can see the distribution return the scatter plot so for each point will present the article. The distribution shows the similarity clusters of articles as scatter plot.

B. CORPUS V2

In this section show how preprocess steps effect on corpus and reduce number of words also keep only important words after make preprocessing. Table 12 show number of words during the preprocessing steps.

TABLE 12
EFFECTIVE PREPROCESSING STEPS ON CORPUS V1

Preprocessing steps	Conservative	Reform	Revolutionary
Numbers of articles	58	80	68
Tokenize numbers	14111	29607	23853
Remove punctuation	13323	28109	22605
Remove unwanted words	12177	25698	20607
Normalization	12177	25698	20607
Remove less than 3	10681	22553	18504
NLTK Stop words	7663	17796	14781
Final words	7663	17796	14781

As we can see in table above the final words. The highest words with Revolutionary article flow by Reform article and the lest Conservative article. In the figure below show the most frequency words in corpus V2 and present in cloud word.

Fig 8: show the most frequent words in cloud words

The figure 8 show the most frequency words and as we can see the words will affect the decision will presented and different from figure 6 in corpus V1. In table 13 will present the most 10 frequency words belong to each class.

TABLE 13
MOST 10 FREQUENCY WORDS FOR EACH CLASS IN CORPUS V2

No	Conservative		Reform		Revolutionary	
1	المحافظين	79	الاصلاح	360	اليسار	237
2	المحافظه	76	السياسي	268	الثوره	165
3	السياسي	73	السياسيه	218	اليساريه	86
4	المحافظ	71	النظام	140	السياسي	80
5	السياسيه	50	المجتمع	80	الاجتماعي	76
6	المجتمع	46	العربيه	71	السياسيه	62
7	الحزب	40	عمليه	70	المجتمع	59
8	التيار	39	التحول	65	الثورات	55
9	حزب	33	العربي	58	الشيوعيه	54
10	الثوره	32	الدول	58	اليمين	54

As we can see in table above most of words same meaning but the write shape different so in next, we will present root and light stemming for reduce all these words from corpus V2. In figure 9 show the PCA and T-SEN.

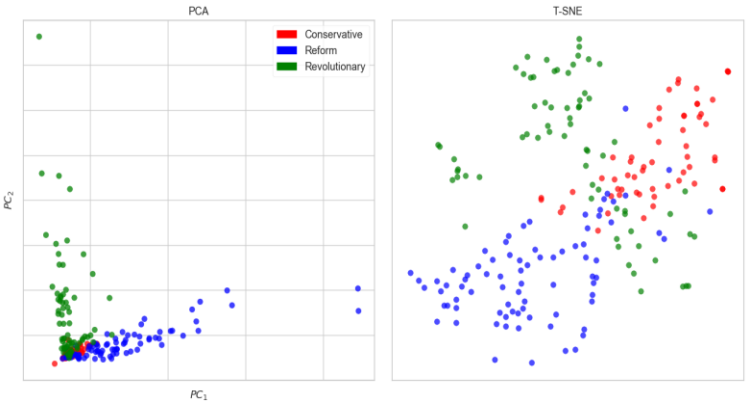

Fig 9: distribution of corpus V2 using PCA and T-SNE

Figure 9 show the distribution of our corpus V2 and as we can see the distribution show corpus V2 non-linear distribution. In this figure the distribution will be better than in figure 7 in corpus V1 that will make decision high accurate.

C. *Corps V3*

In this section show how root stemming effect on corpus V2 and reduce number of words. Table 14 show number of words when apply root stemming.

Table 14
Most 10 frequency words corpus V3

No	Conservative		Reform		Revolutionary	
1	حفظ	326	صلح	659	يسر	482
2	سيس	196	سيس	606	ثور	378
3	جمع	141	نظم	316	سيس	260
4	حزب	118	جمع	302	جمع	260
5	فكر	100	عمل	241	عمل	254
6	تير	87	دول	189	طبق	167
7	حكم	79	حكم	162	حزب	154
8	علم	79	حقوق	151	شيع	137
9	عمل	67	علم	147	علم	124
10	دول	63	عرب	138	نظر	121

In table 13 for corpus V2 there are many words in same root but different shape so after apply root stemming and se we can see in table 14 the words back to the original root. The figure 10 show the articles distribution for corpus V3.

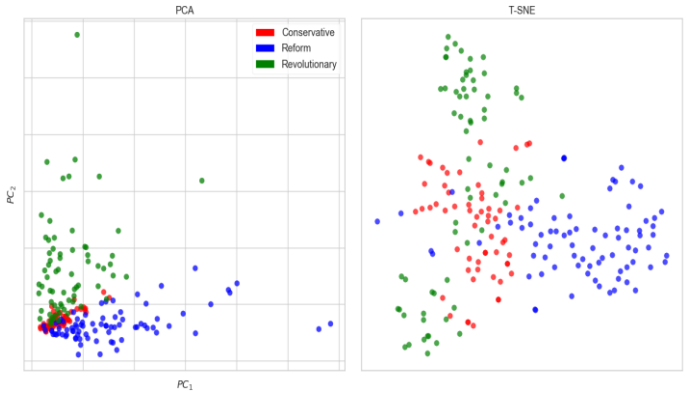

Fig 10: Corpus V3 distribution using PCA and T-SNE

As we can see in figure above the corpus V3 is non-linear distribution.

D. *Corpus V4*

In this section show how light stemming effect on corpus V2 and reduce number of words. Table 15 show number of words after apply light stemming.

TABLE 15
MOST 10 FREQUENCY WORDS CORPUS V4

No	Conservative		Reform		Revolutionary	
1	محافظ	197	اصلاح	507	يسار	299
2	حزب	88	سياسي	294	ثوره	233
3	سياسي	86	سياسيه	248	يساريه	109
4	تيار	86	نظام	212	سياسي	108
5	محافظه	84	مجتمع	149	سياسيه	103
6	مجتمع	78	تحول	92	اجتماعيه	102
7	سياسيه	65	عمليه	92	مجتمع	94
8	دوله	45	تغيير	87	حزب	91
9	تغيير	35	دوله	84	ثور	81
10	ثوره	35	حكم	79	تيار	80

In table 14 for corpus V3 root stemming reduce the root words but the meaning of word occurs almost lost. In table above for light stemming the meaning still there. Figure 11 show the distribution articles of corpus V4.

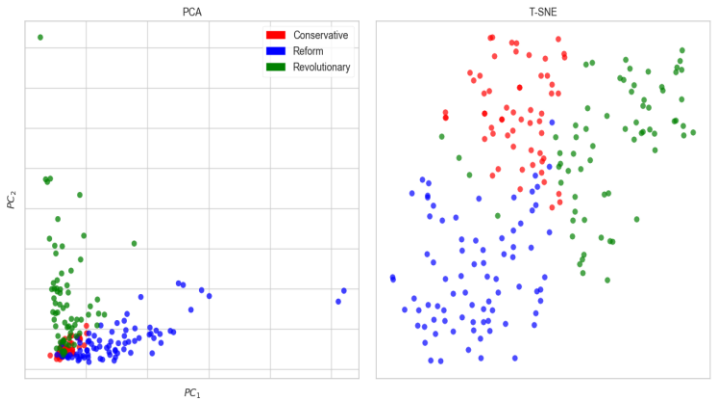

Fig 11: corpus V4 distribution using PCA and T-SNE

CONCLUSION

In this article, we have presented the Political Arabic article dataset (PAAD). We address the research problem by collecting and building four types of corpus from PAAD and make this corpus available online. These corpus as V1 the original raw data and from this corpus will building corpus V2 that will apply on its preprocessing steps. From corpus V2 will building two another corpus such as V3 apply root stemming and corpus V4 apply light stemming. This dataset PAAD will use for how interest in NLP or text categorization.

REFERENCES

[1] M. S. Khorsheed, "Off-line Arabic character recognition–a review," Pattern analysis & applications, vol. 5, no. 1, pp. 31-45, 2002.
[2] B. Brahimi, M. Touahria, and A. Tari, "Data and Text Mining Techniques for Classifying Arabic Tweet Polarity," Journal of Digital Information Management, vol. 14, no. 1, 2016.
[3] M. Markatou, H. Tian, S. Biswas, and G. Hripcsak, "Analysis of variance of cross-validation estimators of the generalization error," Journal of Machine Learning Research, vol. 6, no. Jul, pp. 1127-1168, 2005.
[4] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Classifying Political Arabic Articles Using Support Vector Machine with Different Feature

Extraction," in International Conference on Applied Computing to Support Industry: Innovation and Technology, 2019, pp. 79-94: Springer.
[5] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Political Articles Categorization Based on Different Naïve Bayes Models," in International Conference on Applied Computing to Support Industry: Innovation and Technology, 2019, pp. 286-301: Springer.
[6] M. I. Khalaf, "Machine Learning Approaches and Web-Based System to the Application of Disease Modifying Therapy for Sickle Cell," Liverpool John Moores University, 2018.
[7] E. Hovy, "Corpus Annotation," in The Oxford Handbook of Computational Linguistics 2nd edition, 2015.
[8] N. Chambers et al., "Identifying political sentiment between nation states with social media," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 65-75.
[9] W. Du, W. Liu, J. Yu, and M. Yi, "Russian-Chinese sentence-level aligned news corpus," in Proceedings of the 18th Annual Conference of the European Association for Machine Translation, 2015.
[10] M. Rubtcova, O. Pavenkov, V. Pavenkov, and E. Vasilieva, "Representations of trust to public service in Russian Newspapers during election time: Corpus-based content analysis in public administration sociology," Mediterranean Journal of Social Sciences, vol. 6, no. 4, p. 436, 2015.
[11] M. Stranisci, C. Bosco, H. FARIAS, D. IRAZU, and V. Patti, "Annotating sentiment and irony in the online Italian political debate on# labuonascuola," in Tenth International Conference on Language Resources and Evaluation LREC 2016, 2016, pp. 2892-2899: elra.
[12] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using Twitter to predict the UK 2015 General Election," Electoral Studies, vol. 41, pp. 230-233, 2016.
[13] S. Ahmed, K. Jaidka, and J. Cho, "The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties," Telematics and Informatics, vol. 33, no. 4, pp. 1071-1087, 2016.
[14] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2931-2937.
[15] T. De Smedt and S. Jaki, "The Polly corpus: Online political debate in Germany," in of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018), 2018, p. 33.
[16] S. Stier, A. Bleier, H. Lietz, and M. Strohmaier, "Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter," Political communication, vol. 35, no. 1, pp. 50-74, 2018.
[17] M. Lai, V. Patti, G. Ruffo, and P. Rosso, "Stance evolution and twitter interactions in an italian political debate," in International Conference on Applications of Natural Language to Information Systems, 2018, pp. 15-27: Springer.
[18] Y. Hu, "Refocusing democracy: the Chinese government's framing strategy in political language," Democratization, pp. 1-19, 2019.
[19] K. Mehta, "Underpinnings of Political Leaning: Using Collocation Extraction and Semantic Analysis to Categorize Ideological Frames," 2019.
[20] F. Falck et al., "Measuring Proximity Between Newspapers and Political Parties: The Sentiment Political Compass," Policy & Internet, 2019.
[21] C. Bosco, M. Lai, V. Patti, and D. Virone, "Tweeting and Being Ironic in the Debate about a Political Reform: the French Annotated Corpus TWitter-MariagePourTous," in Tenth International Conference on Language Resources and Evaluation LREC 2016, 2016, pp. 1619-1626: elra.
[22] S. Hegelich and D. Janetzko, "Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet," in Tenth International AAAI Conference on Web and Social Media, 2016.
[23] L. Rheault, "Expressions of anxiety in political texts," in Proceedings of the first workshop on nlp and computational social science, 2016, pp. 92-101.
[24] K. Lazaridou and R. Krestel, "Identifying political bias in news articles," Bulletin of the IEEE TCDL, vol. 12, 2016.
[25] S. Lomborg and A. Bechmann, "Using APIs for data collection on social media," The Information Society, vol. 30, no. 4, pp. 256-265, 2014.
[26] M. Oussalah, F. Bhat, K. Challis, and T. Schnier, "A software architecture for Twitter collection, search and geolocation services," Knowledge-Based Systems, vol. 37, pp. 105-120, 2013.
[27] J. M. Kevan and P. R. Ryan, "Experience API: Flexible, decentralized and activity-centric data collection," Technology, knowledge and learning, vol. 21, no. 1, pp. 143-149, 2016.
[28] H. Elsayed and T. Elghazaly, "Information Extraction from Arabic News," IJCSI International Journal of Computer Science Issues, vol. 12, no. 1, pp. 1694-0814, 2015.
[29] A. A.-S. Ahmad, B. Hammo, and S. Yagi, "ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 3, no. 3, 2017.
[30] A. Chouigui, O. B. Khiroun, and B. Elayeb, "A TF-IDF and co-occurrence based approach for events extraction from arabic news corpus," in International Conference on Applications of Natural Language to Information Systems, 2018, pp. 272-280: Springer.
[31] R. Abooraig, S. Al-Zu'bi, T. Kanan, B. Hawashin, M. Al Ayoub, and I. Hmeidi, "Automatic categorization of Arabic articles based on their political orientation," Digital Investigation, vol. 25, pp. 24-41, 2018.
[32] I. Zeroual, D. Goldhahn, T. Eckart, and A. Lakhouaja, "OSIAN: Open Source International Arabic News Corpus-Preparation and Integration into the CLARIN-infrastructure," in Proceedings of the Fourth Arabic Natural Language Processing Workshop, 2019, pp. 175-182.
[33] W. Zaghouani, "Critical survey of the freely available Arabic corpora," arXiv preprint arXiv:1702.07835, 2017.
[34] A. O. Al-Thubaity, "A 700M+ Arabic corpus: KACST Arabic corpus design and construction," Language Resources and Evaluation, vol. 49, no. 3, pp. 721-751, 2015.
[35] J. Singh and V. Gupta, "Text stemming: Approaches, applications, and challenges," ACM Computing Surveys (CSUR), vol. 49, no. 3, pp. 1-46, 2016.
[36] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II, 2005, vol. 1, pp. 152-157: IEEE.
[37] D.-C. Li, C.-W. Liu, and S. C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets," Artificial Intelligence in Medicine, vol. 52, no. 1, pp. 45-52, 2011.
[38] Y. J. Lee, "an introduction to Principal Component Analysis (PCA)," Available online : http://jupiter.math.nctu.edu.tw~yuhjye/assets/file/teaching/2017_machine_learning/PCAsubset.pdf Accessed [02 Fab 2018].
[39] S. K. Saha, S. Sarkar, and P. Mitra, "Feature selection techniques for maximum entropy based biomedical named entity recognition," Journal of Biomedical Informatics, vol. 42, no. 5, pp. 905-911, 2009/10/01/ 2009.
[40] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of machine learning research, vol. 9, no. Nov, pp. 2579-2605, 2008.
[41] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization," in 2010 IEEE Pacific Visualization Symposium (PacificVis), 2010, pp. 121-128: IEEE.
[42] S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl, "Concentri cloud: Word cloud visualization for multiple text documents," in 2015 19th International Conference on Information Visualisation, 2015, pp. 114-120: IEEE.