Analysing fake news titles for 2016 Trump-Hillary campaign using contextual-based approaches in text analytics

Azwa Abdul Aziz[1], Andrew Starkey[2]

[1]*Universiti Sultan Zainal Abidin, Faculty Informatics & Computing, Tembila Campus, 22200, Malaysia*
[2]*School of Engineering, University of Aberdeen, AB24 3FX, Aberdeen, United Kingdom*

ABSTRACT

Text analytics is the process of transforming unstructured text data into meaningful information that can be used for fact-based decision making. It is widely used for sentiment analysis, summarising text or searching for useful information from the web. Existing approaches such as machine learning or natural language processing techniques have been proven to obtain significant information from massive amounts of text data. However, these approaches can have issues of obtaining sufficiently accurate results during training or the limitation of linguistic resources for the understanding of slang or acronyms for example. Thus, we propose a new method called Contextual Analysis (CA) that accentuates the relationship of the words and sources that are used for analysis. This approach will create a self-learned knowledge tree of contextual information, based on where words appear in the underlying sources. CA provides an understanding of the degree of relationship between the context of words which is a new technique to understand textual data sources. To evaluate CA techniques, 2000 news items are used that contain fake and actual news during 2016 Trump-Hillary campaign. The results are compared with other prominent Supervised Machine Learning (SML) techniques. CA matched the best classification performance and achieved the best performance of 0.81 accuracy for fake news prediction. Moreover, CA provides a Hierarchal Knowledge Tree (HKT) that helps to understand the context of words used in both fake and real news and is one of the important findings of this method. The experimental results demonstrate that CA has the potential to undertake classification tasks and at the same time reveal the contextual relationship and hierarchy of words which improves upon existing ML methods that treat each word as independent.

Keywords: *Contextual Analytics, fake news, hierarchical knowledge tree, text analytics, supervised machine learning.*

I. INTRODUCTION

Text analytics is a crucial process to convert unstructured data into meaningful data for the purpose of knowledge discovery. It has been applied in numerous research areas such as information extraction, text summarization, text machine learning, opinion mining and transfer learning. Research conducted by Hu, Chen, & Chou [1] is an example of how travellers share important information in text such as their opinion, experiences and perspectives on social networking sites.

However, a major problem with text data is that it is sparse and has high dimensionality [2]. It is also referred to as a feature space problem [3], which is the process of selecting important features (words) to represent documents. The main challenges are not only for filtering the information that is required (e.g. pre-processing text, removing stop words, languages, tokenized), but also dealing with 'slang', sarcasm or a different domain adaption.

Data mining also known knowledge discovery is the process of extracting information from large data sets using various techniques [27]. The past decade has seen the development (data mining process) of the most common approaches used to discover knowledge in text: Machine Learning (ML) and Natural Linguistic Processing (NLP). ML is widely used for text classification and prediction while NLP refers to the ability of a computer to understand human speech or text. Both techniques have shown a good result in extracting important elements of the text. However, the limitation of machine learning is it depends strongly on the initial training dataset (supervised ML) whereas NLP requires good linguistic resources to achieve good results. NLP also has difficulty in understanding text containing sarcasm or 'slang'. Therefore, we propose a new technique known as Contextual Analysis (CA) to overcome the limitation in both approaches.

So far, there has been little discussion about how to understand text based on the relationship between text sources and words they contain. CA method is an approach that emphasises the relationship of the words and groupings that words with similar contexts create based on the aggregation of their sources (e.g. where different words appear in the same context). The method is inspired by the Self-Organizing Map (SOM) [4], which is a type of Artificial Neural Network (ANN). Words and sources are embedded together in nodes within a SOM based on the relation and intersection of sources. Branches of the analysis are further created using parent node-child SOMs that allow the contextual analysis of sources to be undertaken. Fig. 1 shows an example of the knowledge tree structure created by CA.

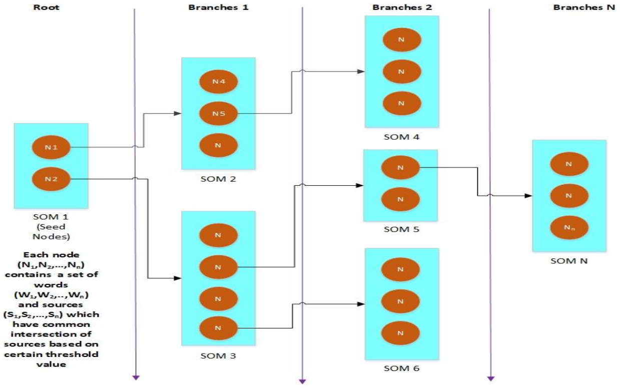
Root
Branches 1
Branches 2
Branches N
N1
N2
SOM 1 (Seed Nodes)
Each node (N_1,N_2,...,N_n) contains a set of words (W_1,W_2,..,W_n) and sources (S_1,S_2,...,S_n) which have common intersection of sources based on certain threshold value
N4
N5
N
SOM 2
SOM 3
SOM 4
SOM 5
SOM 6
N_n
SOM N

Fig 1. Knowledge Tree Creation

To evaluate the technique, we used 2000 news titles (1000 fake; 1000 real) during the 2016 US President Election Campaign between Donald Trump and Hillary Clinton that can be obtained from the Kaggle website [5]. The aim is to create a Hierarchical Knowledge Tree (HKT) for both news title types in order to undertake a relationship comparison between fake news and real news. A number of examples of titles extracted from the dataset can be seen in Table 1.

Table 1: example news titles

FAKE TITLES	REAL TITLES
Homeless Woman Protects Trump Walk of Fame Star From Violent Leftists!'	*Hillary Clinton Makes A Bipartisan Appeal on Staten Island*
Ex-Assistant FBI Director: Clintons Are a Crime Family	*Anti-Trump forces seek last-ditch delegate revolt*
*Hillary Clinton in HUGE Trouble After America Noticed SICK Thing Hidden in this Picture... * LIBERTY WRITERS NEWS*	*Sanders Trounces Clinton in W. Va. -- But Will It Make a Difference?*

The main contribution of the research are as follows:

- A new technique is proposed to perform classification tasks known as CA that has less dependence on the training dataset or on linguistics resources.
- The text is considered as a relation of words that can be summarised using HKT which helps to understand particular concepts of words and their relationship to each other. For example the word 'war' may have a different sentiment orientation in fake or real news
- Identify the relation of words can lead to a dynamic real time classification process whereby the system flags when classification is not possible, identifying when new words being used – this can be used to identify when predictions using other approaches may also fail

This paper has been divided into five parts. The first section deals with a simple introduction of the research. The next section briefly describes the past work done in text analytics and fake news detection. Section III illustrates the framework of the CA methods and elaboration of the process. Section IV discusses the results of experiments conducted. Finally, in Section V, the conclusions and discussion of potential improvements is made.

II. RELATED WORKS

The two most common approaches in text analytics are using ML and NLP. ML is divided into two: supervised Machine Learning (SML) and unsupervised ML. The paper focus in SML for text analytics for comparison. Generally, SML exploits training data to train a classifier and predict unknown data in testing data. Classifiers are mostly trained using a set of features comprised of n-grams [6] which is a contiguous sequence of n items from a given sample of text or speech.

Many researchers have used various SML techniques such as Random Forest Classifier (RFC), Support Vector Machine Stochastic Gradient Descent (SGD), and Naive Bayes (NB) for text analysis [7], [8], [9], [10], [11], [25], [26]. The main limitation of the technique is the requirement for training data of sufficient size and scope for the particular classification task. Kotsiantis [12], points out that the training dataset may suffer from noise and incomplete data. The option for using 'brute-force' which involves measuring everything available in the dataset leads to poor accuracy of classification and the wrong interpretation of data.

Alternatively, unsupervised ML does not need any labeled data, thus can be applied to any text data without manual effort [2]. Some examples of unsupervised ML techniques are k-means and self organising map (SOM). However, these approaches generally require more computational resources and result in slow performance although there are some techniques proposed indicate high performance and accuracy results.

NLP techniques, also known as lexical approaches, use linguistics resources (e.g. MPQA corpus, General Inquirer Corpus) to extract the important elements of text data. The techniques concern the grammatical and language structure of sentences. Vani and Gupta [10] use the approach to detect plagiarism. In their major study, they are using the WordNet lexical database to extract the semantic concepts that are used for semantically relevant comparisons. However, this method also has a constraint in dealing with new language structures such as 'slang' that can be commonly found on the web (Twitter, Facebook).

Fake news has resulted in sophisticated problems in the new digital era. Analysing fake news is a crucial process to avoid for example malign influence in political elections. False information is a part of the contemporary media system whether it is in the form of conspiracy theories or unsubstantiated rumours [13]. Hyman [14] highlights that fake news proliferation and distrust in news media have become an endemic problem in the American society and has aided in the corruption of civil political discourse. He also suggests developing an open standard for the identification of fake news by blending autonomous analysis with the human-driven process. Alcott & Gentzkow [15] clearly state that the users of social media play a vital role in the fake news dissemination process. Following the 2016 US Presidential Election, 62% of US adults obtained news from social media that resulted in the most popular fake news stories being more widely shared compared to mainstream news items. Brigida & Pratt [16], in their research observed the reaction of fake news toward the equity market. The finding shows that the stock price reacts to the news faster than option prices.

Figuirea & Oliveira [17] highlighted one of critical challenges for established reliability of online information is digital content monitoring. Thus, advanced analytical techniques must be developed as counter measures to ease text data analysis. Jang et al. [18], used a computational network approach known as evolution tree analysis to examine the roots and spreading pattern. Research conducted by Cardoso, Silva, & Almeida [19], presents a comprehensive analysis for the automatic filtering of fake reviews, as bad reviews can damage the reputation of brands and manipulate users' perception about products or companies. Although there are differences between fake reviews and fake news, their purposes are similar in that they seek to damage the reputation of certain parties. An interesting finding is that the prediction models fail with real-world data because the models overestimate results achieved during training and in artificial datasets. This is one of the main motivations for using CA approaches.

In fake news text analytics research, Gilda [20] uses NB and RFC to predict classification of news items which leads to 67% and 56% accuracy respectively. The study also includes other feature selection criteria such as the reliability of sources. Meanwhile, Granik & Mesyura [21] choose NB for fake news detection which gives 74% accuracy. Both techniques require the manual labeling of the training dataset. In another major study, Buntain & Golbeck [22] proposed automatically identifying fake news in popular Twitter threads using two credibility focus datasets: CREDBANK and PHEME. The research focuses on the validity and availability of both datasets. PHEME features give classification of 66.93% for potential false threads while CREDBANK is slightly higher with 70.28%.

In conclusion the current approaches do not achieve high predictive classification and given that they are mainly ML type approaches they are also "black box" in that it can be hard to know what words the model has used to give predictive capability.

III. CA APPROACHES

CA uses Bag-of-Words (BoW) approaches in separating the words in the sources. In this model, the words are separated and counted without concern on the grammar or structure of sentences. However, text pre-processing and Part-of-Speech Tagger (POS) techniques are applied in order to remove irrelevant words from the sources. Fig. 2 illustrates the flow of overall CA process.

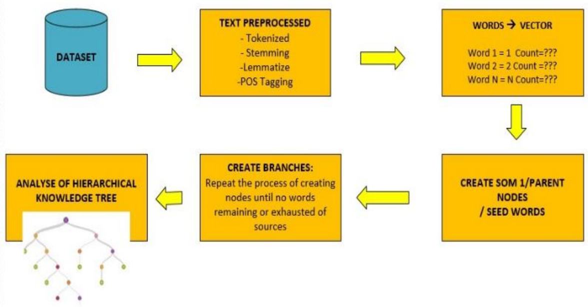
DATASET
TEXT PREPROCESSED
- Tokenized
- Stemming
-Lemmatize
-POS Tagging
WORDS → VECTOR
Word 1 = 1 Count=???
Word 2 = 2 Count =???
Word N = N Count=???
CREATE SOM 1/PARENT NODES / SEED WORDS
CREATE BRANCHES: Repeat the process of creating nodes until no words remaining or exhausted of sources
ANALYSE OF HIERARCHICAL KNOWLEDGE TREE

Fig. 2. Knowledge CA Process Flow

The first step of the CA process is to clean the dataset using text pre-processing techniques. The text will be tokenized into smaller pieces or tokens. Then, the process continues to remove stop words; (e.g. 'the', 'a') or any words that are less than 4 characters in length; (e.g. 'and', 'but', 'so'). Next, each word needs to be normalised. The normalisation approach converts text to the same case, removes punctuation, replaces characters, and so on. There are two crucial processes in text normalization: stemming and lemmatization. Stemming is the process of eliminating affixes (suffixes, prefixes, infixes) from a word in order to obtain a word stem. In contrast, lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Finally, we applied POS tagging techniques using PennTreebank tag set to choose only adjectives, nouns, adverbs and verb groups [24]. A treebank is a parsed corpus that annotates syntactic and semantic in a sentence structure.

The second stage of the CA process is to convert all words and sources into a vector. Each word and source is transformed into a number using a lookup table. This allows a faster comparison between words and sources to be made. Transforming words to numbers also speeds up other calculations made during the CA algorithm.

The third and fourth stages create the HKT which is the most critical part of the process. This is created by the two most important data obtained from the dataset; sources and words. For a given data *D* containing a set of sources denoted by $D = \{ S_1, S_2, \ldots S_n \}$, each source ($S_i$) has a set of unique words which in a sequence of N words denoted by $S_i = \{W_1, W2, \ldots W_N \}$. The nodes are created based on the count of unique words appearing in the sources depicted in the formula below:

$$f_{kl} \quad (w, s) = \sum_{d=1}^{D} \emptyset(S_d = l) \emptyset(k \in w) \qquad \textbf{(1)}$$

where Ø (x) is an indicator function which takes a value of 1 if x is true and 0 otherwise. The equation calculates how often word k appears in sources (l) in the dataset. Thus, we obtain a ranking of the words based on sources using the formula. The word with the highest source count will create the first node of the SOM where each node contains words and their associated sources. Then, the 2nd highest word will be compared against this newly created node. If the sources from the 2nd word and the node have at least 50% (say) in common, then the word will be added to the node, otherwise a new node is created. The process continues to the next word which will be compared against all existing nodes until the number of sources for the word falls below a defined percentage of the first word. On completion of the first SOM, the process continues by creating branches using the words and sources that are left over from this first phase of the analysis. As a result, three types of relationship are obtained from the hierarchy knowledge tree that can be explored; words in the same node (identical-relation), words in the same SOM but in different nodes (genealogical-relation), words that are related by parent node-child SOM relation (inheritance-relation). The overview of this algorithm is shown in Table 2 for creating first SOM.

Table 2: pseudo code for creating first som

```
PROCESS: CREATE FIRST SOM

INPUT:
    S is the list of sources
    W are the words in the sources (S)
        Y is the total of the number of unique sources
        Z is the total of the number of unique words
        D is number of sources (S) in a document

1   Convert all sources (S) contain in documents (D) into a unique number.
        $\forall s, s = 1, \ldots\ldots\ldots, Y \; where \; s \in D$
2   Convert all words (W) contain in sources (S) into a unique number
        $\forall w, w = 1, \ldots\ldots\ldots, Z \; where \; w \in S$
3   Provide a ranking of words based on the number of sources using equation (1)
4   The word with the highest count will be labeled as $w_1$.
5   Find all words that have sufficiently overlapping sources based on total count ($tc$) of S in $w_1$.

        Eg. $w_2, w_3 \ldots w_z > (tc/2))$

7   Create first node ($n_1$) which contains $w_1$ and sources ($s_1, s_2 \ldots, s_y$)
8.  Search for 2nd word ($w_2$) sources and compare with sources it first node ($n_1$)
9.  IF the sources overlapped (e.g.: the sources have more than 50% similarity with sources in first node ($n_1$).
        ADD 2nd word, $w_2$ (and its sources) to this node ($n_1$)
    ELSE
        CREATE second node ($n_2$)
    END IF
10. CONTINUE for remaining words and verify the words and sources to all existing nodes.
11. Once the process is exhausted, all the sources need to be mapped to at least one node.  If any have not, then they are written to a new node, which is equivalent to 'none of the above'.
```

The process of Table 2 will be repeated for remaining words in the nodes and by creating branches. If the number of sources and the number of unique words for particular nodes are sufficiently high, by using certain threshold setting, new analysis will be conducted.

The tree structure can be described as a combination of SOM (M) for a given tree (*T*) denoted as $T = \{m_1, m_2, \ldots m_n\}$ where a sequence of nodes $N = \{n_1, n_2, \ldots i_n\}$ is subset of SOM; $n \in M$. The relation between SOMs and nodes that can represented by the formula below:

$$T = \{N\},\ where\ N \subseteq SOM, \{n_1, n_2, \dots, n_m\} \in M\ .$$
$$SOM = \{som_1, som_2, \dots, som_m\} \in T$$

(2)

For each node, it contains words (W) and sources (S).

$$N = \{W, S\}, \{w_1, w_2, \dots, w_m\} \in W \quad \text{and}$$
$$\{s_1, s_2, \dots, s_m\} \in S$$

(3)

The size of the tree can (x) be determined by accumulating the sum of SOMs or total nodes.

$$f(x) = \sum_{i=1}^{M} x_i \ \ , x = \{hkt_1, hkt_2, .., hkt_m\} \mid x = \{n_1, n_2, \dots, n_m\}$$

(4)

Based on the equation, therefore, total nodes (β) also can be derived by using this equation:

$$\beta = \sum_{h=1}^{K} \sum_{i=1}^{J} N_{hi}$$

(5)

The CA approach also has the ability to undertake prediction whether the news is fake or real similar to others supervised ML approaches (e.g. RFC, NB, SVM, SGD). To do the prediction, a single tree is created by combining fake and real news (training data). The CA prediction for testing data is based on mapping test data to the 'influential nodes' which are nodes that can determine whether the news is fake or not based on the number of sources that attach to each node. The 'influential nodes' will be triggered if the number of sources exceeds a specified threshold as shown in Fig. 3.

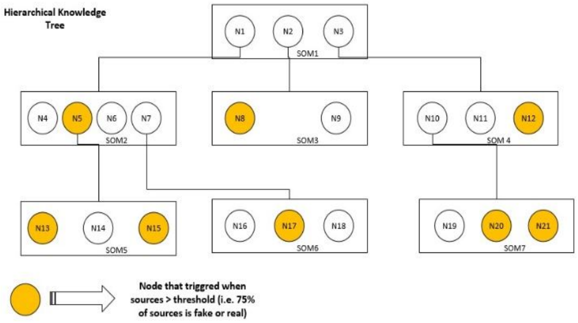
Hierarchical Knowledge Tree
N1
N2
N3
SOM1
N4
N5
N6
N7
SOM2
N8
N9
SOM3
N10
N11
N12
SOM 4
N13
N14
N15
SOM5
N16
N17
N18
SOM6
N19
N20
N21
SOM7
Node that triggred when sources > threshold (i.e. 75% of sources is fake or real)

Fig 3: Burst Nodes in Tree

To compare the accuracy of CA methods, we use 4 main ML techniques. The techniques are Random Forest Classifier (RFC), Multinomial Naïve Bayes (MNB), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) – Random Gradient Descent, Linear Kernel. The ML techniques that are used for comparison are based on the results achieved from previous experiments by Mahalakshmi & Sivasankar [9]. This has been discussed in detailed in a previous research paper [23]. The most obvious enhancement that was proposed in this paper is in using Term Frequency-Inversed Document Frequency (TFIDF) that replaces term frequency for feature selection. The result shows significant improvement which indicates accuracy can be increased between 5%-20% for cross domain sentiment analysis. For in-domain, experiments show more than 80% accuracy for all models. Fig. 4 shows the process of getting result for SML approaches.

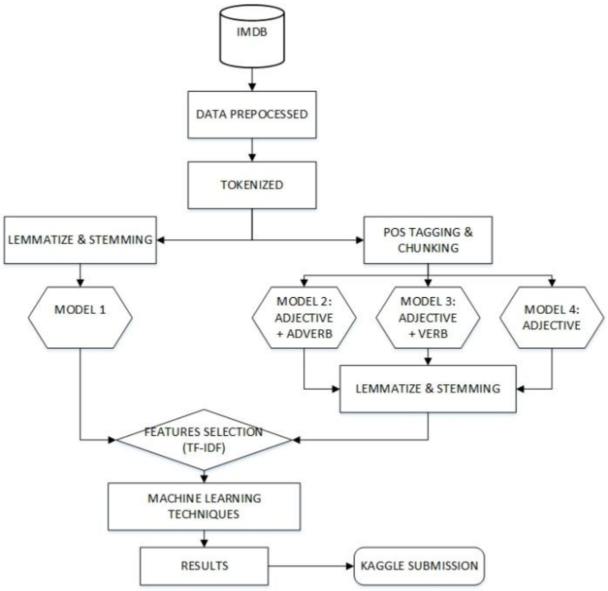

Fig 4: SML process

For a review of the improved performance using TFIDF, the reader is directed to [23].

Terms can be defined as a single word (unigram) or combination of words (bigram, trigram, etc.) derived from a set of documents/sources. The term frequency tf(t,d) is the simplest choice to use as it is the raw count of a term in a document, e.g. the number of times that term t (e.g. word t) occurs in document d. If we denote the raw count by $f_{t,d}$, then the simplest tf scheme is tf(t,d) = $f_{t,d}$. The inverse document frequency (*idf*) is a calculation to know how much information the words provide. It is the process to identify rare words across the documents by dividing the total number of documents by the number of documents containing the term and taking the logarithm of the quotient. The formula below depicts how the calculation is made for *idf*.

$$idf = log \frac{N}{|d \in D: t \in d|} \qquad (\mathbf{6})$$

where N= total number documents in corpus N=|D|, $|d \in D: t \in d|$: number of documents where the term t appeared. Therefore, TFIDF can be calculated as:

$$tfidf = (t.d.D) = tf(t.d).idf(t, d - D).$$

As the results are promising, the selected techniques are selected to be compared with CA. Then, further discussion will be provided to explain the advantages of CA compared against those methods.

IV. RESULTS

The HKT structure that is created from the analysis of fake and real title news is presented in Fig. 5 and 6 respectively.

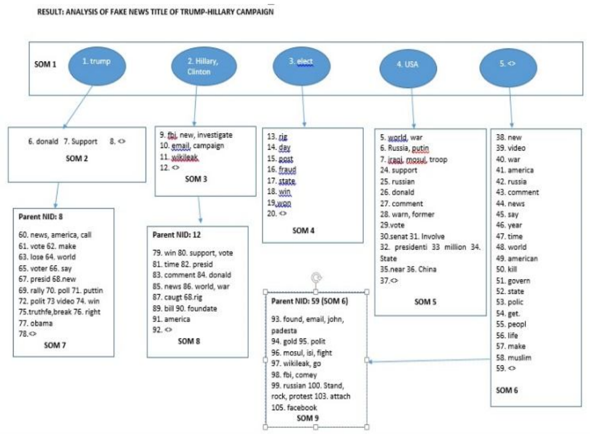

Fig 5. HKT for fake news titles

As can be seen from Fig. 5 and Fig. 6, which show HKT for both fake and real news title analysis, the words in SOM1/seed words create the pillars of the tree. The words indicate the main topics discussed in the news. Both the fake and real news HKT identify *Trump* and *Hillary* as being central to the subject but also having a different context (e.g. they are separate contexts from each other). Other words in the list of seed words include *election* and *USA* for fake and *Obama* for real news. Strong evidence of identical-relation helps to identify words that co-exist together with similar magnitude such as '*Hillary*, *Clinton*', '*World*, *War*', and '*Mosul*, *ISIS*, *fight*'. Interestingly, this also identifies words such as '*Russia*, *Putin*' and '*fbi*, *Comey*' which refer to Russian President, Vladimir Putin and former FBI director, James Comey.

Meanwhile, genealogical-relation and inheritance-relation show a boundary is created between nodes that share the same parent. For example, we find the words in fake news '*email*, *campaign*' and '*fbi*, *investigate*' that share the common parent (*Hillary Clinton*) are separated into different nodes in the child SOM.

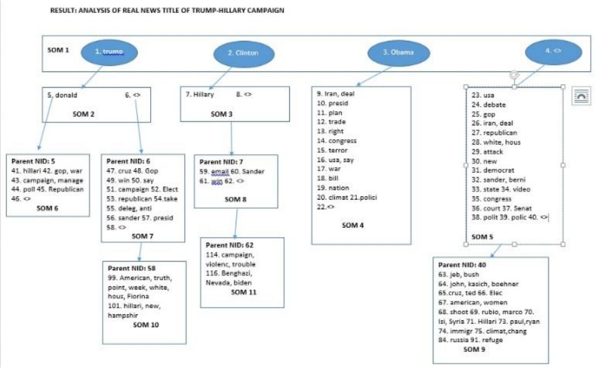

Fig 6. HKT for real news titles

The result can be compared between both trees to find significant differences that exist. There are also different issues discussed in both such as '*climate change*, *immigrant*, *Iran deal*, *policy*' for real news whereas in fake news are FBI investigation of *Hillary*, *wikileak*, *world war and Iraqi Mosul Troop*. The most interesting part is how the same word can appear in a different context. As an example, the identical-relation word in Fig. 5, *Iran deal* appears in *Obama* and also <> node as parent nodes. This will help to understand the clear context where the word is being used. This mechanism can be used to solve one of the major challenges in using NLP techniques which are dealing with semantic words that appear in the sources. Moreover, by classifying the sources based on unique words and their respective sources, will lead to more implicit analysis (drill down) which allows the automatic partitioning of the input data against the words that are present, which in turn can help with classification accuracy. For instance, more detailed analysis (e.g. aspect-based sentiment analysis) can be conducted using others text analytics techniques for the sources that attach to the node with the specific words such as *Trump* or *Clinton*.

Turning now to the experimental evidence of individuals that are found in context, CA identifies several important people apart from Trump, Hillary, Obama, and Putin. They include former director of FBI, James Comey, and John Podesta, chairman of Hillary Clinton's Presidential campaign. Ted Cruz, a candidate for the Republican nomination for US President, Jeb Bush (politician), Paul Ryan, 54th Speaker of the United States House of Representatives, John Boehner, 53th speaker, Bernie Senders (politician) and Marco Rubio (politician) are also found in the HKT for real news.

In order to undertake the prediction of whether new data is fake or real, both trees are combined in order to create a single training tree. Then a new dataset (500 fake titles; 500 real titles) made up of data not used during the training, are mapped against the tree in order to obtain the prediction result. The accuracy of the result is based on general accuracy in ML approaches as depicted in the formula below:

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}\ ; \quad (7)$$

where TP=True Positive, FP=False Positive, TN=True Negative, FN=False Negative.

From 1000 records of the testing dataset, CA can give a prediction for 66% of them, with 34% being unable to be diagnosed by the tree. The fact that the CA approach identifies when it cannot give a prediction is important and also marks a difference between this approach and others. The result for the data that the tree was able to give a prediction for is shown in Table 3, along with the results from the other classification methods.

Table 3: classification performance of ca and ml methods

Model	Class	Accuracy	Average/ Recall
RFC	Fake	0.81	0.79
	Real	0.76	
MNB	Fake	0.81	0.77
	Real	0.72	
SGD	Fake	0.74	0.76
	Real	0.79	
SVM-RGD	Fake	0.0	0.50
	Real	1.00	
SVM-LK	Fake	0.77	0.79
	Real	0.80	
CA	Fake	0.81	0.77
	Real	0.72	

As can be seen from the Table 3, CA method results are similar to those achieved with other established ML methods. The best performance is obtained from RFC and SVM-LK, with CA achieving the second highest performance overall. It is noticeable that the majority of methods performed better on the fake news, with a lower performance for the real news datasets. In that regard, CA matched the best classification performance and achieved the best performance of 0.81 accuracy for fake news.

For existing ML techniques, by using TFIDF features selection, the top 10 ranking of importance words derived from training dataset is shows in Table 4.

Table 4: ten top features selection using tfidf

Word	Score
Trump	81.93
Clinton	60.66
Hillary	51.93
Election	34.36
Usa	34.31
Obama	34.26
Donald	28.59
New	26.58
Gop	23.05
Video	22.91

It is apparent from this table the top feature selection for ML create the top structure of HKT. One of main advantages of CA is it can identify the weight of words whether fake or real news based on sources attach to the nodes whereas feature selection only give possibility of importance words. Therefore, the influential nodes that have been described based on Fig. 3 can be identified and make the decision for prediction on the testing dataset. Examples of influential words for fake news is shown in Table 5.

Table 5: influential words for fake news

Word	Percentage (%)	Total Sources
mystery	100.00	5
Russian	100.00	5
Muslim	100.00	6
comment	100.00	17
wikileak	100.00	13
Kill	100.00	15
election	78.13	64
War	76.19	42

In contrast, Table 6 depict some influential words that appeared in real news.

Table 6: influential words for real news

Word	Percentage (%)	Total Sources
Senat	100.00	13
Cruz	100.00	16
sander	94.44	18
democrat	93.75	16
berni sander	92.00	25
iran	90.48	42
obama	87.50	96
debat	82.50	40
white house	81.48	27

What is interesting in this data is that both tables show the words that indicate real and fake news are based on a classification of the context. Some words have 100% sources that give an indication of whether it is fake or real such as *wikileak* and *kill* for fake or *senat* and *cruz* for real. Furthermore, some words represent high source count like *Obama* (96 sources) for real and *war* (42 sources) for fake news. This is a significant finding as further analysis can be focused to these specific words and their vicinity and context with other words. The word *war* is also important for analysis for more detail as it appears in different context (Fig. 5) for *Trump* and *Obama* as parent nodes.

The further advantage of CA approach is that the words that are used for the prediction are known, and the CA approach also can give a non-result – e.g. will not classify all data. This means that should the words that are being used for real or fake news begin to change, then the CA approach will automatically flag to the user that this is the case as more data will no longer be able to be classified. In addition, it will be possible to identify the new words that are being used – this can lead to a dynamic real time classification process whereby the system flags when classification is not possible, shows the new words being used, with a user giving feedback as to whether they are real news or fake news items.

V. CONCLUSION & FUTURE WORKS

This paper has proposed a novel approach for how to understand text from limited information and perform classification by using news titles dataset. In this investigation, the aim was to assess whether the proposed approach can discover the hidden relationship between the words and sources by using 2000 titles of fake and real news. This study has shown the ability of CA to describe the word context from the three types of relationship generated by HKT (identical-relation; genealogical-relation; inheritance-relation). The relation contributes to new knowledge that can be illustrated by comparing the analysis of fake and real news titles. One of the more significant findings to emerge from this study is the capability of CA for classification for predicting fake or real news with 77% accuracy, an improvement on some other studies. The result surpassed several prominent ML methods and indicates only 2% differences from the best method. Moreover CA captures the relationship of words used to understand issues surrounding the text. In contrast other ML methods treat each word in text as independent words that only will be used for prediction purposes.

An implication of this is that CA presents opportunities to overcome the limitation in existing methods (e.g. supervised ML, NLP approaches) by automatically identifying when the words being used are beginning to change. This would be a very important output since it would also therefore predict when the classification accuracy for all methods would begin to fail. The advantage of the CA method therefore is that it would be possible to continually update the classification tree based on new data and based on human feedback to identify the classification for new data as it appears. This would not be possible with other methods and would require complete retraining of the model.

These findings offer an approach to classify fake news that is becoming the main challenge arising in the new digital era. However, further experiments must be conducted with further datasets and other ML approaches in order to improve the consistency of the result and compare these against other common approaches. Further work needs to be done to establish whether CA techniques can be improved if more information is provided apart from titles. This is important to improve the result and percentage of the dataset that can be classified by CA.

ACKNOWLEDGEMENT

This research work is supported by UniSZA Research Management, Innovation & Commercialization Centre (RMIC).

REFERENCES

[1] Hu, Y., Chen, Y., & Chou, H. (2017). "*Opinion mining from online hotel reviews – A text summarization approach. Information Processing & Management*", 53(2), 436-440. https://doi.org/10.1016/j.ipm.2016.12.002

[2] Aggarwal, C. C., & Zhai, C. (Ed.). (2012). "*Mining text data (1st ed.)*". Boston: Springer.

[3] Tutkan, M., Ganiz, M. C., & Akyokuş, S. (2016). "*Helmholtz principle based supervised and unsupervised feature selection methods for text mining*" doi:https://doi.org/10.1016/j.ipm.2016.03.007

[4] Kohonen, T. (1982). "*Self-organized formation of topologically correct feature maps. Biological Cybernetics*", 43(1), 59-69. https://doi.org/10.1007/BF00337288

[5] Kaggle (2018), "*Fake News: Building a system to identify unrealible news article*". Retrieved from https://www.kaggle.com/c/fake-news

[6] Ghiassi, M., & Lee, S. (2018). "*A domain transferable lexicon set for twitter sentiment analysis using a supervised machine learning approach*". Expert System with Applications, 106, 197-216. https://doi.org/10.1016/j.eswa.2018.04.006

[7] Wang, W., Tan, G., & Wang, H. (2017). "*Cross-domain comparison of algorithm performance in extracting aspect-based opinions from Chinese online reviews*". International Journal of Machine Learning and Cybernetics, 8(3), 1053-1070. https://doi.org/10.1007/s13042-016-0596-x

[8] Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). "*Tweet sentiment analysis with classifier ensembles*". Decision Support Systems, 66, 170-119. https://doi.org/10.1016/j.dss.2014.07.003

[9] Mahalakshmi, S., & Sivasankar, E. (2015). "*Cross domain sentiment analysis using different machine learning techniques*". Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO - 2015), 77-87. https://doi.org/10.1007/978-3-319-27212-2_7

[10] Vani, K. & Gupta, D. (2018). "*Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges*". Information Processing & Management, 54(3), 408-432.https://doi.org/10.1016/j.ipm.2018.01.008

[12] Kotsiantis, S. B. (2007). "*Supervised machine learning: A review of classification techniques*". Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 3-24

[13] Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). "*The diffusion of misinformation on social media: Temporal pattern, message, and source*". Computers in Human Behavior, 83, 278-287. https://doi.org/10.1016/j.chb.2018.02.008

[14] Hyman, J. (2017). "*Addressing fake news: Open standards & easy identification. 2017 IEEE 8th Annual Ubiquitous Computing*", Electronics and Mobile Communication Conference (UEMCON), pp. 63-69. 10.1109/UEMCON.2017.8248986

[15] Allcott, H., & Gentzkow, M. (2017). "*Social media and fake news in 2016 election*". Journal of Economic Perspective, 31(2), 211-236

[16] Brigda, M. & Pratt, W. R. (2017), "*Fake News, The North American Journey Economy & Finances*", 42(, pp. 564-763. https://doi.org/10.1016/j.najef.2017.08.012

[17] Figueira, Á., & Oliveira, L. (2017). "*The current state of fake news: Challenges and opportunities*". Procedia Computer Science, 121, 17-825. https://doi.org/10.1016/j.procs.2017.11.106

[18] Jang, S. M., Geng, T., Queenie Li, J., Xia, R., Huang, C., Kim, H., & Tang, J. (2018). "*A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis*". Computers in Human Behavior, 84, 103-113. https://doi.org/10.1016/j.chb.2018.02.032

[19] Cardoso, E. F., Silva, R. M., & Almeida, T. A. (2018). "*Towards automatic filtering of fake reviews. Neurocomputing*", 309, 106-116. https://doi.org/10.1016/j.neucom.2018.04.074

[20] Gilda, S. (2017). "*Evaluating machine learning algorithms for fake news detection*". 2017 IEEE 15th Student Conference on Research and Development (SCOReD), pp. 110-115. 10.1109/SCORED.2017.8305411

[11] Anitha, N., Anitha, B., & Pradeepa, S. (2013). "*Sentiment classification approaches- A review*". International Journal of Innovations in Engineering and Technology (IJIET), 3(1), 22-31

[21] Granik, M. & Mesyura, V. (2017). "*Fake news detection using naive bayes classifier*". 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 900-903. 10.1109/UKRCON.2017.8100379

[22] Buntain, C., & Golbeck, J. (2017). "*Automatically identifying fake news in popular twitter threads*". 2017 IEEE International Conference on Smart Cloud (SmartCloud), pp. 208-215. 10.1109/SmartCloud.2017.40

[23] Aziz, A.A, Starkey, A. & Bannerman, M., C. (2017). "*Evaluating cross domain sentiment analysis using supervised machine learning techniques*". 2017 Intelligent Systems Conference (IntelliSys), London, 2017, pp. 689-696. 10.1109/IntelliSys.2017.8324369

[24] Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Schasberger, B. (1994). "*The Penn treebank: Annotating predicate argument structure*". Paper presented at the Proceedings of the Workshop on Human Language Technology, Plainsboro, NJ. 114-119. doi:10.3115/1075812.1075835 Retrieved from https://doi.org/10.3115/1075812.1075835

[25] Baba, N.M. & Makhtar, M. & Syed Abdullah, F. & Mohd Khalid, A. (2015). "***Current issues in ensemble methods and its applications***", Journal of Theoretical and Applied Information Technology (JATIT), 81. 266-276.

[26] Aziz, A.A & Starkey, A. (2020). "*Predicting Supervised Machine Learning Performance for Sentiment Analysis using Contextual-Based Approach*", IEEE Access, 8, 17722-7733.

[27] Rosly, R. et al. (2018), "*Analyzing Performance Classifier for Medical Dataset*", International Journal of Engineering and Technology, 7, 136-138.