Decision-making in the age of algorithm

Comparing Random Forest Classifier with human evaluation on fake-news detection.

Oumaima Hajri
STUDENT NUMBER: 24047322

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Supervisor: Dr. Eva O. J. Vanmassenhove
Second Reader: Dr. Eric Postma
External supervisor: Mr. Nadia Benaissa

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
July 2021

Preface

Primarily, I would like to sincerely thank my supervisor Dr. E.O.J. Vanmassenhove for her excellent guidance, patience, critical questions and most importantly warm spirit during this process. It was a pleasure to write my thesis under her supervision. I would also like to thank my internship supervisor, Mr. Nadia Benaissa, for her great supervision during my internship at Bits of Freedom. She has truly inspired me to further develop my interest and passion for Responsible AI and has helped me with ideas, constructive feedback and positive encouragement.

My deepest praises to Allah and His blessings for the completion of this thesis. And my deepest gratitude goes to my mother Souad Abouliaqin and my father Rachid Hajri. It would not be possible to write this thesis without their loving and never-ending support.

Decision-making in the age of algorithm

Comparing Random Forest Classifier with human evaluation on fake-news detection.

Oumaima Hajri

Fueled by ever-growing amounts of digital data and advances in artificial intelligence, we have been shifting from human to algorithmic decision-making. This has been accompanied by profound possibilities in several sectors, but also with far-reaching effects on citizens' lives due to algorithmic complexity, non-transparency and non-accountability. This study aims to research whether we should continue with the dehumanisation of decision-making. The central question investigates the ability of a supervised machine learning algorithm, Random Forest Classifier (RFC), and human annotators (N = 6) in correctly fact-checking news articles. Two vectorisation techniques were used for our RFC, TF-IDF and Word2Vec, to see whether a particular technique affects our model's performance. The aim is to compare algorithmic and human performance and, given the outcome, to see whether a hybrid approach of both algorithms and humans towards decision-making can be recommended. Results show that the algorithm (with a combination of RFC and TF-IDF being the best) significantly outperforms our human annotators. Nevertheless, our human annotators provide us with more explainable and interpretable decisions. We conclude that considering that we have already turned our world over to artificial intelligence, there is not only an urgent need to bring explainability and interpretability earlier to the design table; but also for algorithmic accountability, indicating the obligation to justify actions based on algorithmic decision-making.

Keywords: algorithmic decision-making, human decision-making, fake news, Random Forest Classifier, human annotators, explainability, interpretability, accountability.

1. Introduction

1.1 Research motivation

Machine learning (ML) algorithms are everywhere and seem to solve most problems by taking over crucial decision-making (Zuiderveen Borgesius et al. 2016). We have been increasingly deferring decisions to algorithms since machines' reputations of outperforming human decision-making processes have been boosted (Amnesty, 2021). From employee hiring to national security management, algorithms impact many decision processes on this planet (Amnesty, 2021). In regard of this trend, one would almost believe that algorithms can do everything better than humans and that human judgement does not hold a crucial value anymore. An example of this development is within the field of jurisprudence, wherein software programs, such as COMPAS, are used to calculate the probability of recidivism among convicts, which may alter their sentence (Washington 2018). However, research by ProPublica in 2016 exposed the negative impact of the algorithm behind COMPAS showing racial bias. The algorithm incorrectly

assigned a high-risk score to black people, whilst white people poorly received a lower score more often, without the algorithm providing any explanation (Washington 2018).

As algorithms expand into public decision-making processes, the potential benefits of algorithmic decision-making are myriad and clear since they are less costly and time-efficient. Yet, it is of utmost importance to focus on the considerable risks associated with their deployment, considering that when algorithms fail they can have far-reaching effects on people. A recent example in the Netherlands has been the controversial 'risk classification algorithm' used by government bodies for detecting social welfare fraud, resulting in the so-called 'Toeslagenaffaire' (Chavannes, Verhulst, and Strijbos 2012). In its combat against fraud, Dutch government bodies have been pooling a broad range of personal data from citizens in various databases since 2013 (Chavannes, Verhulst, and Strijbos 2012). This algorithm has been used to detect possible fraud in the 'kinderopvangtoeslag' (an allowance paid to parents by the governments to compensate for day-care expenses) of several parents. It was designed to flag individuals linked to 'suspicious' data patterns, which, in its turn, triggered further investigation entailing the analysis of these 'positive hits'. The bottom line of this scandal was that the algorithm, being uninterpretable, could not justify its decisions but, still, its outcomes were implemented unquestioningly. The resigned cabinet responded that it is the system's fault and not the people involved within this decision-making process. *"We didn't intend it that way"* is out of place in a situation resulting in destroying the lives of more than 26000 innocent parents (Chavannes, Verhulst, and Strijbos 2012). This is why algorithms need managers, too.

1.2 Fake news

Another example wherein algorithmic decision-making is used is the assessment of truthfulness of claims in the news, namely fact-checking. According to Groshek and Koc-Michalska (2017), the growing relationship between the rise of populism and extensive social media usage exacerbate polarisation. Consequently, users' growing polarisation and existing confirmation bias play a crucial role in spreading fake news (Vicario et al. 2019). A critical aspect is that one given message can reach a vast target audience and consequently have a giant impact (Vicario et al. 2019)). Furthermore, another essential part is that it is more complex for individual users to nose out the authenticity of a piece of information, since information is distilled without necessarily knowing its source (Groshek and Koc-Michalska 2017).

The massive spread of fake news, which became a widespread phenomenon during the 2016 U.S. presidential election, stresses the importance of this problem (Gu, Kropotov, and Yarochkin 2017). According to Cambridge Dictionary, fake news can be operationalised as: *"false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke"*. Another definition is operationalised by Gu, Kropotov, and Yarochkin (2017): *"...the promotion and propagation of news articles via social media. These articles are promoted so that they appear to be spread by other users instead of being paid-for advertising. The news stories distributed are designed to influence or manipulate users' opinions on a certain topic towards certain objectives"*. One of the tools applied to combat this problem is automated fact-checking. Vlachos and Riedel (2014) state that fact-checking is *"...the assignment of evaluating the authenticity and truthfulness of specific claims"*.

The traditional solution to this task is to ask experts, such as journalists, to check claims against evidence-based on previously spoken or written facts (Oshikawa, Qian, and Wang 2018). Algorithms have already been assisting fact-checkers and could be put

to further use particularly in monitoring, claim matching and managing communities. However, there are limits to their effectiveness, as one of the most crucial challenges experienced by fact-checkers is the task of having to explore multiple facets before classifying an article based on its truthfulness (Ahmad et al. 2020). According to Nucci, Boi, and Magaldi, principal researchers on the Fandango project that aims to combat misinformation, fake news is *"not a mathematical question of algorithms and data, but a very philosophical question of how we deal with the truth"*. Thus, it is crucial that we must first answer a very delicate question related to the classification of news as fake or true, namely: what is truth? The latter epistemological question builds up whether it is feasible to fact-check an article objectively by labelling its language based on its factual correctness. Interpretation, being ubiquitous and discretionary, makes the process of fact-checking challenging. According to a 2020 report by Fullfact, fact-checkers are sceptical since they believe that this process is wrongly being automated whilst it requires human judgement (Arnold 2020). If interpretation is even ubiquitous and discretionary for humans, can we then let algortihms deal with problems, and eventually decisions, that we as humans cannot easily explain and/or interpret?

Nevertheless, considering the rapid spread of fake news and the overwhelming amount of data generated in today's society, solutions for automated fake-news detection must exist and leave no other option since it is hard to keep up with the pace without using automated tools (Shu et al. 2018). Automation and scalability are benefits brought with automated detection systems and are reached with various techniques and approaches. The two most commonly used approaches for detecting fake news are Natural Language Processing (NLP) and ML. NLP refers to enabling computers to understand human language and respond appropriately - and ML refers to computers learning without explicitly being programmed. Unfortunately, both these approaches, consisting of millions of calculations, can't be described in a human interpretable way and will, therefore, be considered black boxes. Thus, explainable and interpretable AI must be a research priority to make the results of these algorithms explainable as well as interpretable for their developers, their users, and the people they affect too.

1.3 Research questions

The main goal of this research is to investigate whether algorithmic decision-making produces better results than human decision-making. Our study will compare manual fact-checking by human annotators with automated fact-checking by an algorithm. Overarching thoughts and questions leading to this research are as follows: is the dehumanisation of decision-making, by deploying algorithms instead, a development we should further pave the way in? Or should algorithms not replace human decision-makers, since we cannot solely rely on their outputs? Therefore, the central research question of this thesis is as follows: ***"How does the performance of the Random Forest Classifier (RFC) compare to human annotators, in fact-checking news articles?"***.

This primary research question is answered by addressing the following sub-questions that each account for a specific part of the analysis:

1. To what extent can the RFC, when either used Word2Vec or TF-IDF as a vectorisation technique, correctly classify an article as 'fake' or 'real'?
2. How does the performance of our RFC compare to the judgements of our human annotators in terms of accuracy, precision, recall and F1-score?

3. Given the outcome of sub-question (1) and (2), can a hybrid approach of both algorithms and humans towards decision-making be recommended? Why/Why not?

2. Related Work

Before diving further into the research question of our study, it is crucial to set out the discussion of the most recent academic insights beforehand. We will do so by firstly going through automated techniques used for fact-checking. Secondly, vectorisation techniques will be elaborated. And, finally, existing literature regarding algorithmic decision-making vs human decision-making will be discussed.

2.1 Automated fact-checking techniques

As mentioned earlier in section 1, the spread of fake news has a tremendous social-political impact on society and exacerbates polarisation (Oshikawa, Qian, and Wang 2018). Luckily, automated detection of fake news has been researched in ML, NLP, and Deep Learning (DL). According to (Shu et al. 2018), formulating fake news detection can best be done by approaching it as a binary classification problem, wherein supervised learning is used. The latter entails training an algorithm on input data that has been labelled beforehand for a particular output. Another approach used is to present fake news detection as a regression problem, wherein the output is a numeric score of truthfulness (Nakashole and Mitchell 2014)). However, since the datasets used for fake news detection have discrete scores, the challenge remains in converting discrete labels to numeric scores. Thus, most automatic fake-news detection studies employ supervised learning and approach it as a classification problem. Most of the existing research uses supervised methods, while semi-supervised or unsupervised methods are less commonly used. Numerous techniques have been deployed to evaluate ML algorithms in text classification. More specifically, in a survey conducted by Oshikawa, Qian, and Wang (2018), various supervised learning ML models are presented for fake news detection. They are differentiated between neural and non-neural network models. The aim is to train a mathematical model sufficiently in one of the two categories, in such a way that it can predict examples of, e.g. fake news, based on numeric clustering and distances (Shu et al. 2018). Looking at literature examining text classification, Logistic Regression (LR) appears to be most commonly used as a baseline algorithm due to its robustness, low processing time and high performance (Mahir et al. 2019); (Shu et al. 2018); (Jurafsky and Martin 2018).

Looking at benchmark models, it is evident that the state-of-the-art literature suggests using neural networks such as Recurrent Neural Network (RNN), which is a prevalent method in NLP. The idea behind an RNN is to use sequential information to predict the next element in the sequence (Zellers et al. 2019). Specifically, Long Short-Term Memory (LSTM), a special kind of RNN, is frequently used for text classification since it entails a feedback loop containing short-term and long-term memory. It reaches great performance for anything with a sequence and performs well on text classification since the meaning of a word depends on the word before and after (Zellers et al. 2019). In a study conducted by Vijayaraghavan et al. (2020), LSTM reached a mean test score of 94.88% when modelled with CountVectorizer as a vectorisation technique. However, it is good to note that for, any neural network, the training phase is a very resource-intensive task and, thus, needs more computational power as well as more input data.

Furthermore, looking at non-neural network models, Support Vector Machines (SVM) and decision trees such as Random Forest Classifiers (RFC) are used occasionally for text-classification tasks due to their good performance (Shu et al. 2018). For example, the SVM model, as proposed in (Ahmad et al. 2020); (Looijenga 2018); (Pérez-Rosas et al. 2017), has frequently been used for text classification considering its high performance and robustness (Hmeidi et al. 2015). Furthermore, due to SVM's having different clustering methods and distance functions, occasionally good results are achieved. For example, in a study conducted by Ahmed et al. (2021), wherein a comparison in detecting fake news between five well-known ML algorithms is conducted, SVM showed outperformance with an F1-score of 0.94% and a recall score of 1.0. The latter entails the measure of a model's ability to correctly predicting true positives. Finally, according to Joachims (1998), SVM's work well for text classification due to: their ability to handle large feature spaces; their ability to find linear separators which work well with text categorization due to their linear separability; and their implementation of feature selection by avoiding irrelevant features. However, it is essential to note that the latter can lead to aggressive feature selection, resulting in a loss of information (Joachims 1998).

As mentioned earlier, RFC's achieve good performance as well. In another study by Hassan et al. (2017), wherein a Random Forest classifier is used to flag non-factual sentences, the model reaches an F1-score of 97%. It is interesting to note that the fundamental reasons to use RFC's for text classification tasks are that they can deal with high dimensional noisy data (Islam et al. 2019). Additionally, an RFC combines several predictions of many trees into one single model, with the logic that reaching a majority vote will be better than using one single model. This results in RFC's being less prone to overfitting (Hassan et al. 2017). When comparing RFC's to neural networks, it can be stated that RFC's are less computationally expensive and do not require a GPU to finish training. Furthermore, neural networks need much more data than one can have on hand for the model to be effective.

2.2 Text vectorisation techniques

According to Uysal and Gunal (2014), text vectorisation is the activity of converting text data into a representation that is typically in the form of a numerical representation (vector) that encodes the meaning of a particular word (Uysal and Gunal 2014). Some state-of-the-art models for text vectorisation techniques are TF-IDF (Jones 1972); (Luhn 1957) and Word2Vec (Mikolov et al. 2013).

Groundbreaking advances were made with the advent of the latter, introduced by a Google research team. An intuitive explanation of Word2Vec is that it uses neural networks to learn word embeddings. These word embeddings capture information about the words surrounding the word to be vectorised (Mikolov et al. 2013). This was a remarkable improvement in terms of identifying semantic features of words. However, a drawback of Word2Vec is the inability to handle unknown words that it has not encountered before, which results in the usage of a random vector (Di Gennaro, Buonanno, and Palmieri 2021). An example is within the domain of Twitter which consists of a lot of sparse and noisy data, leading to a large corpus with words not often occurring.

Moreover, TF-IDF evaluates how specific relevant terms are to a document in a collection of documents. This statistical measure takes two metrics into account: how many times a term appears in a document (TF) and its inverse document frequency across a set of documents (IDF). This process makes this model easy to compute.

However, according to Zhang, Yoshida, and Tang (2011), TF-IDF is only helpful as a linguistic level feature since it does not capture the position in the text, co-occurrences, or semantics and, thus, makes it context-independent. Nevertheless, Truşcă (2019) states that TF-IDF outperforms Word2Vec when it is considered that data should meet the requirement of linear separability - which is the case in fake-news detection.

Particularly within fake-news detection, not a lot of research has been conducted in comparing these different techniques in the task of achieving good classification. However, in 2020, a study conducted by Vijayaraghavan et al. (2020) investigated the effectiveness of CountVectorizer, Word2Vec and TF-IDF implementation in different supervised models to solve the problem of fake-news detection. Among the three techniques, CountVectorizer achieved the best performance, with TF-IDF being second best. Nevertheless, according to Vijayaraghavan et al. (2020), this good performance is due to TF-IDF and CountVectorizer being heavily keyword-driven and therefore penalising the contextual meaning of the words.

2.3 Algorithmic decision-making vs human decision-making

The discussion of whether humans or algorithms make the best decision-makers has been an interesting field of research for many scholars. This has led to a field of interest wherein the performance of algorithms is compared with that of humans within several academic disciplines. Examples are: "A Comparison of Human and Computer Information Processing" (Whitworth and Ryu 2009); "Comparison of human and computer performance across face recognition experiments" (Phillips and O'toole 2014); "Evaluating human versus ML performance in classifying research abstracts" (Goh et al. 2020)) and "Human vs supervised ML: Who learns patterns faster?" (Kühl et al. 2020). The results vary from algorithms being able to learn new patterns faster and human performance not improving anymore due to cognitive overload; to humans being able to outperform algorithms because machines learn slower and need more instances to train compared to humans to obtain the same results (Kühl et al. 2020). Furthermore, many emphases have been put on the quality of decision-making, with concerns of algorithms indeed being able to learn faster - but bringing with them the consequences of opaque, biased and unaccountable decisions (De Laat 2018).

To date, specifically within the field of automatic fake-news detection compared to human evaluation, not a lot of research has been conducted. This is mainly due to two critical factors. On the one hand, the shortage of datasets regarding fake news; and the available datasets only depending on one domain of interest, namely politics (Pérez-Rosas et al. 2017). And on the other hand, building fake-news datasets that require human expertise to verify the news articles - which is considered a time-consuming task. This consequently complicates the acquirement of datasets representing different domains that could eventually result in - to some extent - generalisable datasets (Pérez-Rosas et al. 2017). Furthermore, it is not straightforward to differentiate the truth of published news timely, due to the pace of how fast information, in all its diversity and subjectivity, is being shared on the internet (Ko et al. 2019).

Moreover, related to our study, research is done on the automatic detection of fake news, wherein results are only put in perspective by comparing performance with an empirical human baseline accuracy (Pérez-Rosas et al. 2017). Another study conducted by Pasquetto et al. (2020) at the Harvard Kennedy School of Misinformation review, concluded that algorithms could indeed speed up time-consuming processes when fact-checking. However, Pasquetto et al. (2020) states that it is crucial to note that algorithms cannot draw final decisions, since many concepts related to fake news remain nuanced

and subjective to humans. Furthermore, Pérez-Rosas et al. (2017) adds that classifying news as fake or true is often difficult since there are instances where news can be either partially true or fake (Pérez-Rosas et al. 2017).

Ahmad et al. (2020) propose an interesting approach to deal with this problem by adding additional classes such as 'mostly fake' or 'mostly true'. However, within this research, we will only focus on binary classification since our dataset is already labelled in this way. Furthermore, it is crucial to consider that most literature within the field of fake news detection focuses on specific datasets, predominantly within politics, which is the case with our dataset as well. Thus, the algorithms trained in this research will perform best on datasets of the same type since they will be trained on a unique textual structure. As Shu et al. (2018) state, it is difficult to have a generic algorithm trained and working on all kinds of datasets and, therefore, conclusions in this research should be drawn carefully.

3. Methodology

In this section, our methodology considered in this research will be discussed. Firstly, our dataset will be presented followed by a brief explanation of our algorithms. Secondly, our vector representation techniques will be elaborated. Finally, the task of our human annotators will be defined.

Our literature review has shown that most studies within automatic fake-news detection conduct supervised learning for this task and approach it as a binary classification problem. Referring to section 2, RFC will be used as a benchmark model and LR as a starting point for our baseline measures. Furthermore, Vijayaraghavan et al. (2020) inspired us to implement a similar methodology wherein different vector representation techniques implemented within algorithms are compared. For our research, TF-IDF and Word2Vec will be used for vector representation to obtain state-of-the-art results. The aim of using both these techniques will be to see whether algorithms' performances depend on specific text vectorisation techniques. Based on the literature review, we hypothesise that our model will outperform our human annotators. However, the human annotators will be able to provide us with more interpretable and/or explainable results.

3.1 Algorithms

For the purpose of our research, as mentioned above, two RFC algorithms will be implemented with Word2Vec and TF-IDF. The crucial aspect of this research is to compare algorithmic performance with that of human annotators. Therefore, although our proposed algorithms have been discussed briefly in section 2, we will be discussing them further in the paragraphs below.

3.1.1 Baseline model. LR, being a straightforward model, is used to predict the probability of a categorical dependent variable. The objective of an LR is, given independent variables, to train a classifier that can make a binary decision about the class of a new input observation (Jurafsky and Martin 2018). Here we introduce the sigmoid function that helps the LR make this decision. The term 'sigmoid' means S-shaped and is also known as a 'squashing' function, since it maps real-valued numbers into a range of [0,1]. In our study, 0 refers to a fake news article, and 1 to a true news article. Thus,

we would want to know the probability of P(y = 1 | x) of an article being true and the probability of P(y = 0 | x) of an article being fake (Jurafsky and Martin 2018).

Our LR makes a classification by comparing the weighted sum of our preprocessed input values to a threshold that is set before. Whenever an observation is greater than the threshold, the model will classify this observation with 1. Whenever an observation is smaller than the threshold, the model will classify this observation with 0 (Jurafsky and Martin 2018). The higher a threshold is, the more restrictive the LR will be before classifying an article as 1, and, thus, more false negative errors will be made (stating that an article is fake whilst it is true). The lower a threshold is, the less restrictive the LR will be before classifying an article as 0. Thus, more False Positive errors will be made (stating that an article is true whilst it is fake) (Jurafsky and Martin 2018). The default threshold for an LR is 0.50 and will be used for our study as well. Figure 1 below is an illustration of the sigmoid function.

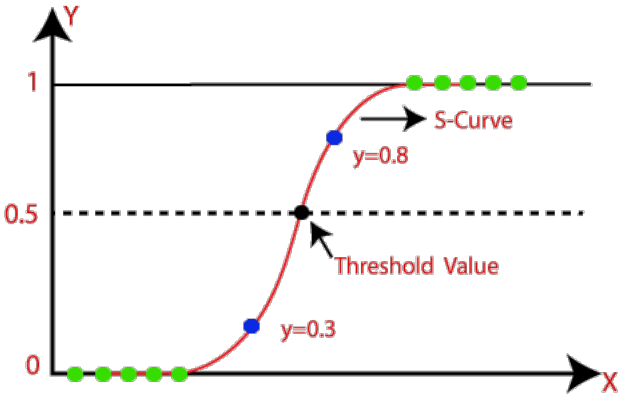


Figure 1: The sigmoid function with an S-shaped curve. The threshold used is 0.5. Considering the explanation above, the data points will be classified as follows: y = 0,3 → 0 and y = 0,8 → 1. This image is retrieved from a blog by Antony Christopher at medium.com (Christopher, 2021)

3.1.2 Benchmark model. An RFC is an ensemble algorithm that merges multiple decision trees to obtain a prediction (Gregorutti, Michel, and Saint-Pierre 2017). Since we are dealing with a categorical target variable (whether our article is fake or true), the decision trees in our RFC are of the type of categorical variable decision trees. Before we dive deep into the working principle of the RFC, we must understand the architecture related to the decision tree.

- Decision tree

As the name goes, a decision tree uses a tree-like model of decision. It is constructed using two kinds of significant elements: the nodes and the branches (Cutler, Cutler, and Stevens 2012). Typically, a decision tree starts from the decision itself called a 'node' with each 'branch' representing a possible outcome. The furthest branches on the tree represent the end result and are called the 'leaves' (Cutler, Cutler, and Stevens 2012). This decision pathway follows a set of if-else conditions, which gives this algorithm a treelike shape. These if-else conditions depend on the so-called feature importance,

which describes which features are relevant to a particular conclusion. The higher the value, the more critical the feature. Figure 2 below shows a simple decision tree with a categorical target variable for the question "Is a person fit?".

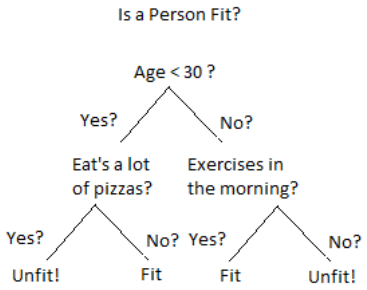


Figure 2: This figure illustrates a decision tree for the question "Is a person fit?". The root node presents the questions of whether the person's age is smaller than 30. The branches suggest' Yes?' or 'No?' and the terminal nodes are presented with the outcomes 'Fit' and 'Unfit!'. Source: xoriant.com

However, with more than one feature participating in the decision-making process, it is necessary to decide their relevance to the final decision. Thus, the most relevant feature is placed at the root node traversing the tree down by splitting the nodes (Cutler, Cutler, and Stevens 2012). By calculating the Attribute Selection Measure (ASM), like the Gini Index or Information Gain (IG), it can be decided which feature is placed at the root node and which features are placed whilst traversing the tree (Devi and Nirmala 2013). An ASM measures the best splitting criterion separating a given data partition, D, labelled into classes (Devi and Nirmala 2013). It performs best when each split results in a 'pure' split, which entails the features falling into the class they belong to (Devi and Nirmala 2013). Thus, as we traverse further down the tree, the level of impurity and, hence, uncertainty decreases, leading us to better classifications.

For our study, the Gini Index will be used as an ASM since it is computationally inexpensive and produces the same results as IG for text classification tasks, as stated by Tangirala (2020). The Gini Index calculates the chance of a feature being wrongly classified when the model randomly chooses it. The degree of a Gini Index varies between 0 and 1, where 0 refers to all features being classified to the class they belong to, and 1 to the features being randomly distributed among both classes (Tangirala 2020). Within our research, the words within our articles will serve as features that will decide whether an article will be classified as fake or true. It will be interesting to see which words will serve as the best splitting criterion and will be decisive for classifying a particular article.

- RFC

For fake news detection, an RFC is used in our study since it can handle large datasets - and produces a low error in text classification tasks due to low correlation between the decision trees (Gregorutti, Michel, and Saint-Pierre 2017). Another significant incentive for deploying an RFC for our study is the fact that it is an explainable algorithm due to its intuitive nature, which means that its algorithmic parameters and mechanics can be explained in a way understandable for humans (Sagi and Rokach 2020). This will make it easier to perspective our results when comparing them to our human annotators since our subquestion 3 concerns an eventual hybrid approach.

Furthermore, the choice to deploy an RFC instead of one single decision tree for our task is because bundling different decision trees will result in more stable predictions (Gregorutti, Michel, and Saint-Pierre 2017). This is due to the fact that an RFC does not rely on the feature importance given by one single decision tree; instead, it chooses its features randomly during the training process by trying out different combinations. Therefore, an RFC would be able to generalise better over our data, making it more accurate than a single decision tree.

3.2 Text vectorisation techniques

To analyse and model text after preprocessing, it must first be converted into features. Below the methods used in our study are presented.

3.2.1 TF-IDF. TF-IDF, short for Term Frequency - Inverse Document, is a statistical score that calculates how important a word is for a document. The term frequency (TF) increases proportionally with the number of times it appears in a document but is offset by its frequency in the overall corpus (IDF) (Erra et al. 2015). Its main goal is to scale down the impact of words that occur very frequently and are, hence, empirically less informative (Erra et al. 2015). While TF-IDF is a good primary metric for converting text into a numerical representation understandable for our algorithm, a downside is that it does not consider a word's semantic meaning (Erra et al. 2015). It is good to note that since our RFC is only sensitive to feature importance, the term frequency of particular words will not significantly impact the decisions.

3.2.2 Word2Vec. Word2Vec is a two-layer neural network trained to produce low-dimensional vector representations of words, i.e. word embeddings, that capture essential information surrounding the word to be vectorised (Mikolov et al. 2013). In other words: it is a way to convert input text data in a numerical format that our algorithm can read while maintaining the original between words in a corpus. An example of this process is portrayed in Figure 3 below.

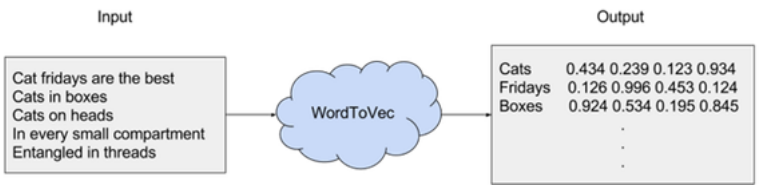


Figure 3: This figure illustrates a diagram of Word2Vec with a text corpus as an input and a vector representation as an output. This image is retrieved from a blog by Zafar Ali at medium.com (Ali, 2019).

As an input, Word2Vec takes a substantial, large corpus of words and, eventually, every word vector is positioned in the vector space in proximity based on their similarities (Mikolov et al. 2013). For example, words like 'dad' and 'mom' will thus be closer in semantic space than the words 'dad' and 'cat'.

3.3 Human annotators

Participants in this research, also referred to as human annotators, will be operating in politics, journalism or academia. An essential condition for these experts to be selected is to have affinity with reliable sources and be able to distinguish them from unreliable ones. Furthermore, having general know-how of political subjects would help since this is the overarching theme in the dataset. However, this is not a hard requirement. The main task is to conduct manual fact-checking by reading the articles and classifying them as 'fake' or 'true'. Manual fact-checking will be conducted by, for example, going to the sources to verify specific claims or by checking government resources. Briefly, the application of verification (a scientific-like approach to getting the facts and the correct facts) will enable fact-checking and lay a foundation for the manual evaluation process (DataJournalism.com, 2021).

Considering that the human annotators can't evaluate the whole dataset, a small randomly selected subset will be sampled from fake and true news articles for the manual human evaluation. Furthermore, to guide the human annotators in this task, they will be provided with an instruction document containing guidelines. These are derived from the news literacy education tool by Facebook called "Tips to spot false news" and the journalistic principles set up by the Dutch Association of Editors-in-Chief. The instruction document can be found in Appendix A. It has been deliberately chosen to provide the participants with a selected amount of information beforehand, to not guide them too much in this process with the chance of steering their way of thinking.

4. Research setup

4.1 Dataset description.

The 'ISOT Fake News' dataset is a compilation of 44.898 fake and true news stories compiled by (Ahmed, Traore, and Saad 2017). An essential challenge for automated fake news detection is the availability and quality of datasets with enough true and fake labels (Oshikawa, Qian, and Wang 2018)). Therefore, to achieve good performances, a requirement is to have enough solid ground labelled data from fact-checking sites (Oshikawa, Qian, and Wang 2018). Thus, this justifies the choice to use this dataset for our proposed research.

4.2 Exploratory Data Analysis (EDA)

The dataset comprises of two CSV documents ("True.csv" and "Fake.csv") that are concatenated into one CSV document for our data analysis. The features included in each row of the data are "title", "text", "subject", "date". However, the features used as an input for our algorithms are sole "title" and "text". To distinguish between the fake and true articles, an additional feature named "class" is added, with 0 referring to an article being fake and 1 referring to an article being true. Furthermore, during our EDA, duplicate values are identified and removed. There were no missing values in our dataset. Table 1 and 2 below give a breakdown of the distribution of articles and their corresponding classes, both before and after EDA.

Bearing in mind the distribution of our classes, we notice a 17,6% difference between the fake and true articles, with the fake articles being over-represented. However, considering this slight difference, we assume that our dataset is almost balanced and

Table 1: Distribution of the articles in the ISOT Fake News Dataset before EDA.

Total number of articles	Fake articles	true articles
44898	23481	21417

Table 2: Distribution of the articles in the ISOT Fake News Dataset after EDA.

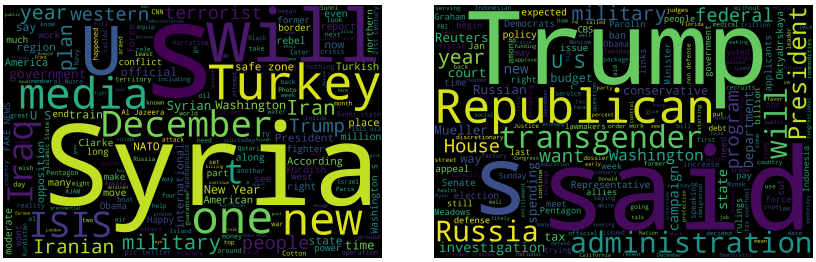
Total number of articles	Fake articles	true articles
38647	21192	17455

should not significantly impact our model performance. Furthermore, our dataset was divided into two subsets after shuffling it randomly. The distribution is as follows: 70 percent of our data is used as a training set and 30 percent as a test set to obtain an unbiased estimation of the performance of our algorithms. Furthermore, the dataset is stratified during splitting to ensure our classes are evenly balanced in both splits. Considering that our test set is not large enough, it will not be split up into a validation set. Table 3 below provides more information about the distribution of the classes within the train and test split.

Table 3: Distribution of our labels within our training and test set.

	Training set (70%)	Test set (30%)
Fake articles number	12232	5223
True articles number	14820	6372

As mentioned earlier, our RFC's will be considering feature importance. This makes it interesting for our study to obtain a first-hand insight into the words present in our datasets. Hence we have created word clouds after our pre-processing. The word clouds presented in Figure 4 below are not used for analytical purposes. Instead, they are used for quick visualisation to depict specific words' frequency and, thus, importance.



(a) Wordcloud fake news **(b)** Wordcloud true news

Figure 4: Word clouds generated from our fake and true news articles. The more often a specific word appears in our dataset, the bigger and bolder it appears.

4.3 Fact-checking process Politifact

According to Ahmed, Traore, and Saad (2017), the articles in the dataset are acquired from reliable (such as Reuters) as well as unreliable websites - and subsequently flagged

by Wikipedia and Politifact. Politifact is one of the most prominent fact-checking independent and non-profit organisations in the United States (Nieminen and Sankari 2021). In PolitiFact, journalists and domain experts (referred to as reporters) review news articles and provide fact-checking evaluation results to claim news articles as fake or true (Nieminen and Sankari 2021). Furthermore, they claim to work according to their 'ethics policy', which requires journalists to set their opinions aside and to not being led by an agenda or biases. Their process of fact-checking goes as follows: going through reviews of what other fact-checkers have found previously; conducting an in-depth Google search; conducting an in-depth search of online databases; consultation with a variety of experts and reviewing their publications; and finally having an overall review of the available evidence. Another important note to add is how Politifact classifies the news articles to reflect relative accuracy. This is achieved with their so-called 'Truth-O-Meter':

- **True**: the statement is accurate, and there's nothing significant missing.
- **Mostly true:** the statement is accurate but needs clarification or additional information.
- **Half true:** the statement is partially accurate but leaves out important details or takes things out of context.
- **Mostly false:** the statement contains an element of truth but ignores critical facts that would give a different impression
- **False:** the statement is not accurate.
- **Pants on fire:** the statement is not accurate and makes a ridiculous claim.

The reporter who researches a particular article suggests a rating and, subsequently, reviews this assigned rating with an editor by providing an additional explanation for his rating choice. After agreeing on the rating, the rated fact-check is, once again, reviewed by two additional editors. Final voting is conducted between the two additional editors and the initial reporter, and the report is then published.

Nevertheless, a drawback of the ISOT Fake News dataset is that it is unclear whether all the articles in the dataset have followed this process of fact-checking. Furthermore, not all labels above are present in our dataset, which does not ensure whether there were any 'mostly true', 'mostly false' or 'half true' articles and, if yes: how they have been classified. It is solely stated by Ahmed, Traore, and Saad (2017) that the articles are retrieved from Wikipedia and Politifact and no further information is given.

4.4 Pre-processing

As the performance and the efficiency of the model depend on the quality of the data, it is of utmost importance to pre-process our data neatly before feeding it into our model. This process is divided into the following stages: data cleaning, data integration, data reduction, and data transformation. The next steps below are taken to preprocess the dataset:

- The input text needs to be converted into a suitable format. This process includes: making the entire dataset lowercase and deleting all

non-alphabetic characters (including punctuation, scientific notations and numerical characters).

- Furthermore, the English Language Toolkit (NLTK) library is used to perform different functions. Contracted words are expanded (don't → do not) to retain essential information of the words, and stopwords are deleted to eliminate insignificant words in our corpus.
- Tokenisation will be performed to split the input text into smaller lines/sentences since we are dealing with entire articles. Lastly, lemmatisation will return the root of valid words by calling their dictionary form.

4.5 Algorithms

In the paragraphs below, the implementation of our algorithms will be elaborated. Firstly, the techniques for vectorisation will be discussed. Subsequently, the experimental setup of both our baseline and benchmark models will be presented.

4.5.1 Vector representation. The first step is to convert our input text into numerical features. Then, our final preprocessed dataset is **duplicated** so that both TF-IDF and Word2Vec are used on these datasets separately. Starting with TF-IDF, the first step is to create matrices of TF and IDF by importing *TfidfVectorizer*. After retrieving a matrix of a TF-IDF representation, *TfidfTransformer* is imported to normalise this representation. Finally, the impact of words that occur frequently should be reduced.

Furthermore, Word2Vec is used to construct word embeddings for each word in our dataset. The first step is to load and organise our input text into sentences. Subsequently, this list of sentences is provided with a constructor of a new Word2Vec instance. The following parameters are used to configure this construction:

- **vector_size** (the number of dimensions of the embedding),
- **window** (the maximum distance between a target word and words around the target word),
- **min_count** (the minimum count of words to consider when training the model) and,
- **workers** (the number of threads to use while training).

For all these values, the default values are being used. Finally, after our model is trained, it is saved to a binary format, loaded, and summarised. Typically, the quality of word vectors increases significantly with the amount of data used to train them. An example could be to use pre-trained vectors trained on the Google News dataset (containing about 3 million words and phrases). However, because this process is computationally expensive, we have only trained the constructor on the corpus of our dataset. After training this constructor on our dataset, the model should predict target words given a particular word for a specific context. Finally, based on the assumption that similar words will have similar embeddings due to their appearances with similar words, they should cluster together.

4.5.2 Baseline model. To retrieve baseline results, a simple LR baseline model is constructed with both TF-IDF and Word2Vec separately. The simplest model we can think of is our model randomly guessing the labels of our articles, with a model accuracy of 55% since our test-set has a 45/55 split. In any case, our benchmark model should at least beat our baseline accuracy level. Finally, the class label is assigned according to the probability given to a particular article by our LR. If the probability score outputs < 0.50, then the class assigned = 0 (fake) and if the probability score outputs > 0.50, then the class assigned = 1 (true).

4.5.3 Benchmark model. Two RFC models have been developed with TF-IDF and Word2Vec separately to retrieve benchmark results. Our algorithms will first be trained with the default values of the hyperparameters. The objective is to retrieve a first approximation of our algorithms' performance on the test set. The next step is to tune our hyperparameters so that performance is optimised. The best method to determine the optimal setting of our hyperparameters is to try different combinations and evaluate each combination. However, evaluating each model only on the training set can lead to overfitting, wherein bias is introduced to the model since it is too closely related to the dataset and would not generalise well on new data.

Usually, when split into a test and validation set, the latter is used to assess the different models with their hyperparameters and select the best performing one (Ren, Li, and Han 2019). However, since we do not have a validation set in our case, Gridsearch cross-validation (GridsearchCV) will be utilised for hyperparameter optimisation. GridsearchCV is an exhaustive tuning technique attempting to compute our algorithms' hyperparameters' optimum values. It does so by splitting our training set into K number of subsets called K-folds, and out of these K-folds, $K - 1$ are used for training, and the remaining for testing. Finally, our K-fold cross-validation results are the average of the results obtained on each set. Figure 5 below illustrates the architecture of K-fold cross-validation.

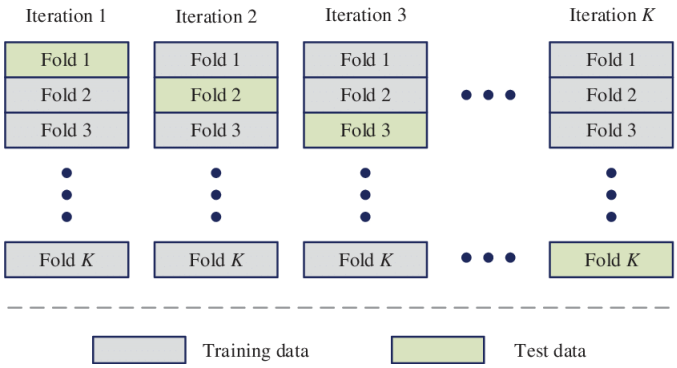


Figure 5: This figure displays the K-fold cross-validation method. The dataset is divided into $K = 3$ subsets, and during every iteration, one subset is used as a test set. This image is retrieved from (Ren, Li, and Han 2019)

Our RFC model is trained on $K = 3$ with the following parameters and their corresponding values:

- **Max_depth** (the maximum depth of each decision tree) [50, 100, 150, 200]
- **N_estimators** (number of trees in the forest) [100, 200, 300, 400]

4.6 Human annotators

In our research, the total numbers of annotators are six. Based on gender, the division among our participants is 50 percent female and 50 percent male and consisted of: a former diplomat, an Editor-in-Chief, a PhD student in Sociology, a Postdoctoral researcher in fake news, a campaigner and a developer of Voting Advice Applications. The total number of evaluated articles is 18, distributed over three groups containing two human annotators each. This way, every article is evaluated by two annotators and will, thus, provide us with the inter-annotator reliability within each group - which is the level of percental agreement between the annotators. Furthermore, the total number of articles are randomly selected and consist of 10 true articles and 8 fake articles.

This research was conducted online, whereby participants were sent a form via email (an example can be found in Appendix B) with the corresponding instruction document for the task. As stated in section 4.3, Politifact's manual evaluation is extensive in terms of checks and balances and does not rely solely on two ground labels ('fake' or 'true) – and, thus, leaves room for interpretation by providing more labels. However, in our research, only the labels mentioned earlier are provided to see whether the participants come across an article that can't be labelled within the two classes – and what their response and/or conclusion will be. Giving them the option of labelling the articles with classes such as 'most likely truth' or 'most like fake' beforehand will steer them in a certain way of thinking.

4.7 Performance evaluation

4.7.1 Algorithmic evaluation. To evaluate model performance for fake news detection, our test set will be used. Accuracy, precision, recall and F1-score are used as test scores to analyse how well the model can classify our articles into the corresponding categories. We will use the weighted test scores since the distribution of our classes is uneven and, thus, we will be able to account for the number of instances in each class. This way we will retrieve the overall performance of the models on both fake and true articles.

4.7.2 Human annotators vs algorithms. Another aspect of our study is to compare the empirical results of our algorithms with the manual evaluation of our human annotators. It is important to note that our model will be generalising over 38647 articles in total (with 11595 articles in the test set) - which is significantly more than the 18 articles that our annotators will evaluate. Thus, one (mis)classification by a participant could affect the performance severely, and, therefore, we should carefully draw our conclusions when comparing. Subsequently, we will run our algorithms on the same 18 articles as evaluated by the annotators to perspective our results. The same evaluation metrics as mentioned in the paragraph above will be used to retrieve the test scores. Furthermore, the inter-annotator reliability is presented to perceive the percental agreement within the duo's.

4.8 Implementation

This proposed research uses the Python programming language with jupyter as its Integrated Development Environment (IDE). For data analysis, the widely used Pandas (McKinney et al. 2010) and Numpy (Harris et al. 2020) are implemented. For pre-processing, the NLTK package (Bird, Klein, and Loper 2009) is used to determine the list of stop words, to lemmatise our corpus, and for text cleaning. Lastly, the Gensim v4.0.0 (Rehurek and Sojka 2011) library and T-SNE tool (Van der Maaten and Hinton 2008) are used for our Word2Vec model (Mikolov et al. 2013), and sklearn (Pedregosa et al. 2011) is used for our TF-IDF and RFC.

Below, in Figure 6, a schematic representation of our proposed methodology is presented.

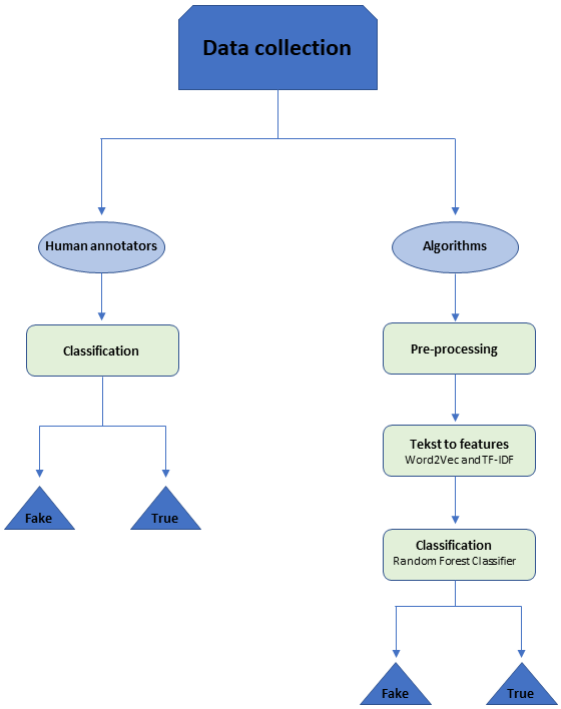


Figure 6: Schematic representation of our proposed methodology.

5. Results

This section is dedicated to presenting the results of our described research setup. The results are displayed as follows. Firstly, visualisations of both TF-IDF and Word2Vec are displayed; secondly, the results of our LR's and RFC's are presented and, finally, the results of our human annotators are presented.

5.1 Word vectorisation

5.1.1 TF-IDF. Using TF-IDF, a matrix of $N = 120322$ features was generated, converting our text data into a numerical representation understandable for our RFC. As mentioned earlier, feature importance describes which features are relevant and takes the

average over all trees in the forest. Figure 7 below shows the top 50 features for our RFC when modelled on RFC, calculated by the Gini Index.

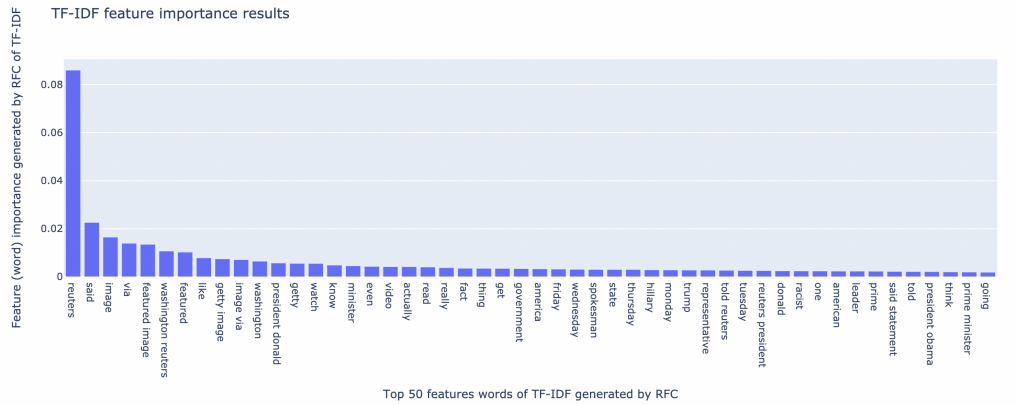


Figure 7: Results of TF-IDF feature importance generated by RFC.

It is interesting to note that the feature 'Reuters' scores the highest, taking into account it being one of the world's most reliable news outlets. This way, we assume that our true articles contain in large quantity the word 'Reuters' and, thus, our RFC places this word at the first node since it is a critical feature in classifying an article as true or fake. Nevertheless, worth mentioning is that the word 'Reuters' does not only appear in true articles, but also in fake articles.

5.1.2 Word2Vec. In our study, 50 Word2Vec features (i.e. 50 dimensions) were produced. Considering these multi-dimensional word embeddings, it was not possible to visualise the feature importance bar chart as we did for our TF-IDF above. However, to understand how Word2Vec works and how the relations between vectors captured from our data can be interpreted, we have visualised our Word2Vec using T-SNE. This is an algorithm for data visualisation based on non-linear dimensionality reduction with a basic idea to map multi-dimensional data (as in our case) to two or more dimensions - where points that were initially far from each other are also located far away, and close points are also converted to close ones. Figure 8 portrays a snapshot of our interactive plot of the first Word2Vec feature, wherein words with similar semantic meaning are clustered together.

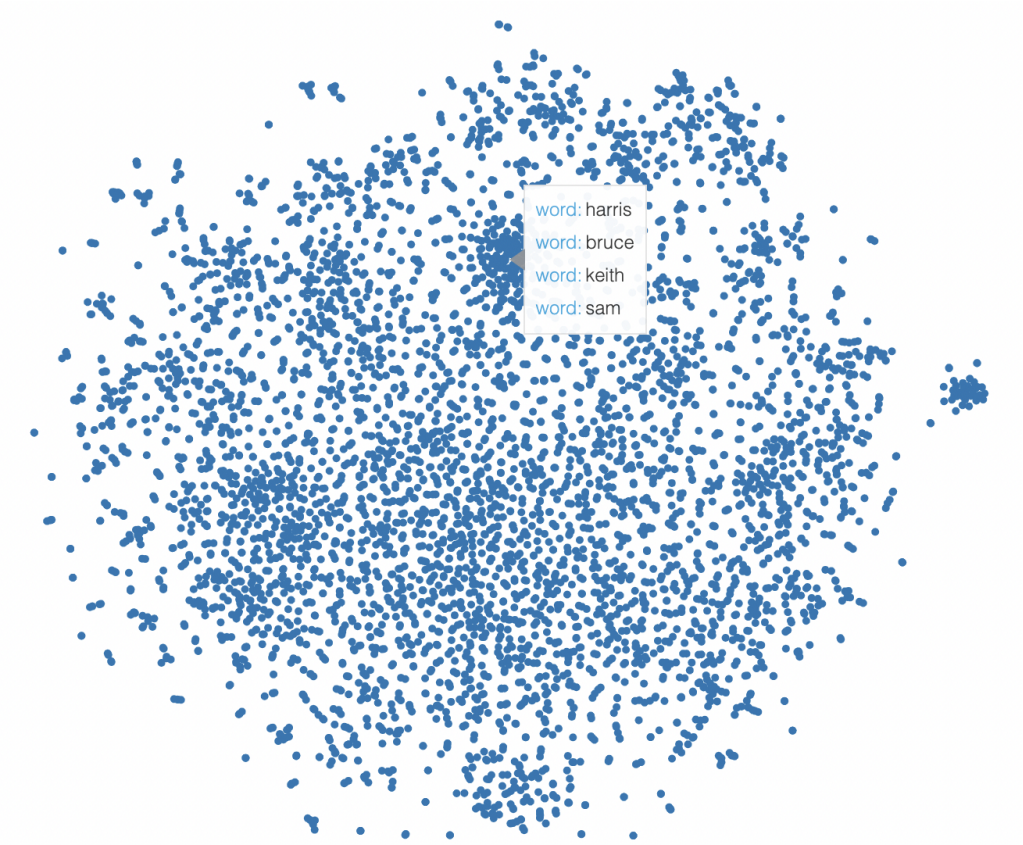


Figure 8: T-SNE Word2Vec visualisation. This figure illustrates that the names 'harris', 'bruce', 'keith' and 'sam' are clustered together due to their similar semantic meaning.

5.2 Baseline model

We adhere to the probability of classifying the articles in our corpus correctly (fake/true) for our baseline. Table 4 below provide us with the **weighted** baseline results of our LR model, both modelled with TF-IDF and Word2Vec. We note that our baseline model achieves good performance and transcends our baseline measure. This applies to both TF-IDF and Word2Vec, with TF-IDF exceeding Word2Vec and almost reaching a 100% F1-score. Given the results above, we can assume that a model with TF-IDF will perform significantly better than Word2Vec.

Table 4: Weighted baseline test scores of our LR's modelled on Word2Vec and TF-IDF.

	Accuracy	**Precision**	**Recall**	**F1-score**
LR(Word2Vec)	0.93	0.93	0.93	0.94
LR(TF-IDF)	0.99	0.99	0.99	0.99

5.3 Benchmark model

Furthermore, two RFC's have been applied with TF-IDF and Word2Vec as vectorisation techniques. Table 5 provides us with the weighted results obtained before hyperparameter tuning used on the model's default parameter values. It is evident that our model performs better on TF-IDF when compared to Word2Vec, with TF-IDF still sticking to a 99,9% performance as shown in our baseline model.

Table 5: Weighted test scores of our RFC's modelled on Word2Vec and TF-IDF before GridsearchCV.

	Accuracy	**Precision**	**Recall**	**F1-score**
RFC(Word2Vec)	0.93	0.93	0.93	0.93
RFC(TF-IDF)	0.99	0.99	0.99	0.99

Table 6 provides us with the results of our RFC model after hypertuning and cross-validation. GridsearchCV (K= 3) computed the optimum values of our hyperparameters and obtained the best set of parameters which are as follows: max_depth: 150 and n_estimators: 400 for both RFC(Word2Vec) and RFC(TF-IDF).

Table 6: Weighted test scores of our RFC's modelled on Word2Vec and TF-IDF after GridsearchCV.

	Accuracy	**Precision**	**Recall**	**F1-score**
RFC(Word2Vec)	0.94	0.94	0.94	0.94
RFC(TF-IDF)	0.99	0.99	0.99	0.99

Considering the results above, we can conclude that our RFC does not remarkably improve in performance after hyper tuning. We only perceive an insignificant improvement of 0.01 for our RFC(Word2Vec) test scores after GridsearchCV. Moreover, Figure 9 below portrays the visualisation of a receiver operating characteristic (ROC) curve on RFC(TF-IDF) and RFC(Word2Vec). The ROC curve illustrates the diagnostic capacity of a binary classifier system. It is a graphical representation that plots two parameters: the true positive rate (samples correctly classified as positive) and the false positive rate (samples incorrectly classified as positive). Classifiers that produce curves closer to the top-left corner indicate better performance. Considering this, it is evident that both our RFC's produced remarkably good performance, with RFC(TF-IDF) being the best combination.

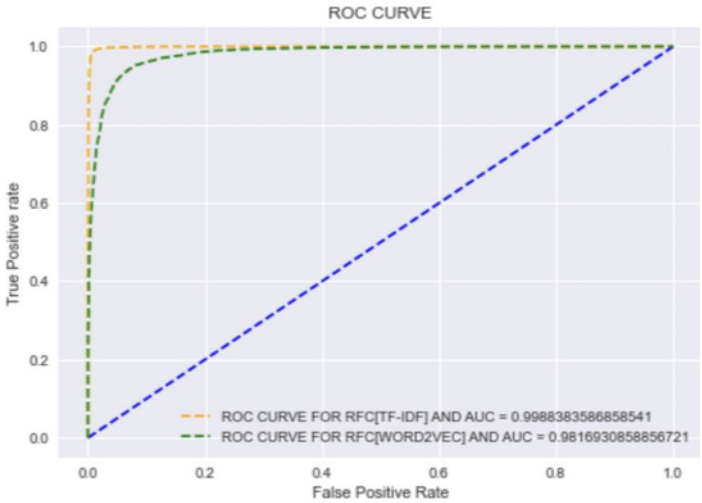


Figure 9: Combined ROC curve of RFC on both Word2Vec and TF-IDF with $K = 3$.

5.4 Human annotators

Within our research, six human annotators divided into duos classified six articles each. Every duo has been randomly handed over the same set of articles, providing us with 36 classifications in total. Interesting to note is that not all human annotators have stuck to the classification task of labelling the articles into one of the two provided classes. Namely, some articles have been labelled with both 'fake' as well as 'true' and provided with accompanying comments to justify this specific choice. Since these labels were not considered at the beginning of our research, these will be clustered together as 'undecided' and will not be taken into account for our test scores. Table 7 below provides us with the test scores of the manual classification task by our human annotators - wherein it is crucial to note that the test scores are calculated **without** the articles that been classified as 'undecided'.

Table 7: Test scores of the manual classifications by our human annotators ($N = 34$) without the articles that have been classified as 'undecided'.

	Accuracy	**Precision**	**Recall**	**F1-score**
Test scores human annotators	0.72	0.82	0.53	0.65

Furthermore, table 8 portrays the precise results of the manual classification task with the actual (y) class of the articles and the predicted ($\hat{y}$) class, accompanied by the inter-annotator reliabilities of the duo's.

Table 8: Results of predicted labels ŷ (N = 36) by our annotators of the selected articles (N = 18). Articles that were classified as 'undecided' are left blank.

		N = 36 classifications		
		y	Annotator 1 ŷ	Annotator 2 ŷ	inter-annotator reliability
Duo 1	Article 1	1	0	1	16.7%
	Article 2	0	1	0	
	Article 3	0	0	1	
	Article 4	1	1	0	
	Article 5	0	-	0	
	Article 6	1	**1**	**1**	
Duo 2	Article 1	1	**1**	**1**	83.3%
	Article 2	0	1	0	
	Article 3	1	**1**	**1**	
	Article 4	1	**1**	**1**	
	Article 5	0	**1**	**1**	
	Article 6	1	**1**	**1**	
Duo 3	Article 1	0	**0**	**0**	66.7%
	Article 2	0	0	1	
	Article 3	0	-	0	
	Article 4	1	**1**	**1**	
	Article 5	0	**1**	**1**	
	Article 6	1	**1**	**1**	

Considering the results produced by our duo's in table 8, it is crucial to state that the falsely classified articles are not identical. I.e. the human annotators do not perceive the same articles as fake/true. This also applies to the articles that are classified as undecided (both true and false). The inter-annotator reliability rates suggest an overall good agreement. However, it is interesting to note that Duo 1 only agreed on one article. These results confirm the diversity in knowledge that people, and even experts, can have.

To put the results into perspective, it is important to see how our algorithms performed on the same selection of articles. We do this by looking at the weighted test scores. Table 9 below provides us with the test scores of our algorithms on the 18 articles.

Table 9: Weighted test scores of the classification on the selected articles (N = 18) by our RFC's.

	Accuracy	**Precision**	**Recall**	**F1-score**
RFC(Word2Vec)	0.83	0.75	1	0.85
RFC(TF-IDF)	0.83	0.72	1	0.84

Furthermore, it is also desirable for our research to look at the probabilities of the predictions, since these probabilities go beyond the numeric expression of uncertainty between 0 and 1; and state how certain a particular decision is. Table 10 below provides us with a detailed description of our models' probabilities in classifying these articles.

Table 10: Results of predicted labels $\hat{y}$ by our RFC's on the selected articles (N = 18) with the corresponding probabilities.

	N = 18 articles				
	y	RFC(Word2Vec) $\hat{y}$	probability	RFC(TF-IDF) $\hat{y}$	probability
Article 1	1	1	0.81	1	0.95
Article 2	**0**	**1**	**0.76**	**0**	**0.88**
Article 3	0	0	0.95	0	0.96
Article 4	1	1	0.92	1	0.96
Article 5	**0**	**0**	**0.58**	**0**	**0.87**
Article 6	1	1	0.93	1	0.93
Article 1	1	1	0.90	1	0.90
Article 2	0	0	0.86	0	0.92
Article 3	1	1	0.90	1	0.99
Article 4	1	1	0.95	1	0.98
Article 5	**0**	**1**	**0.56**	**1**	**0.62**
Article 6	1	1	0.71	1	0.94
Article 1	**0**	**0**	**0.57**	**0**	**0.85**
Article 2	0	0	0.57	0	0.94
Article 3	**0**	**1**	**0.76**	**1**	**0.71**
Article 4	1	1	0.71	1	0.96
Article 5	**0**	**1**	**0.54**	**0**	**0.86**
Article 6	1	1	0.91	1	0.97

Considering the probabilities corresponding to these predictions, it is interesting to note that the ones trying to flag fake articles are overall significantly lower than the ones for true articles. Crucial to the task of predicting fake articles is to obtain a relatively small number of false negatives, since they cause more harm when spread. However, it is not explainable for us why our RFC's produce low confidence when classifying these articles.

Lastly, in an attempt to try understand the results produced by our RFC's, we have aimed to provide a complete visualisation. However, considering the extensive sizes for both RFC(Word2Vec) and RFC(TF-IDF) with **max_depth: 150** and **n_estimators: 400**, it was computationally too expensive to execute this task. Hence, we have visualised one decision tree out of RFC(Word2Vec) for illustrative purposes. Furthermore, it is noteworthy to state that the architecture of the decision tree is still explainable from the visualisation when zooming in, by encountering how the calculations are made based on feature importance. Nevertheless, it is not possible to precisely understand how specific articles are classified. This would also not change if it was possible for us to visualise the whole RFC, since the articles as a whole become redundant - considering that our RFC's are being modelled on TF-IDF and Word2Vec features. Thus, the decision pathway of single articles cannot be distilled from the visualisation and, therefore, results cannot be interpreted.

Figure 10 below illustrates the first decision tree (out of 400) from our RFC modelled with Word2Vec.

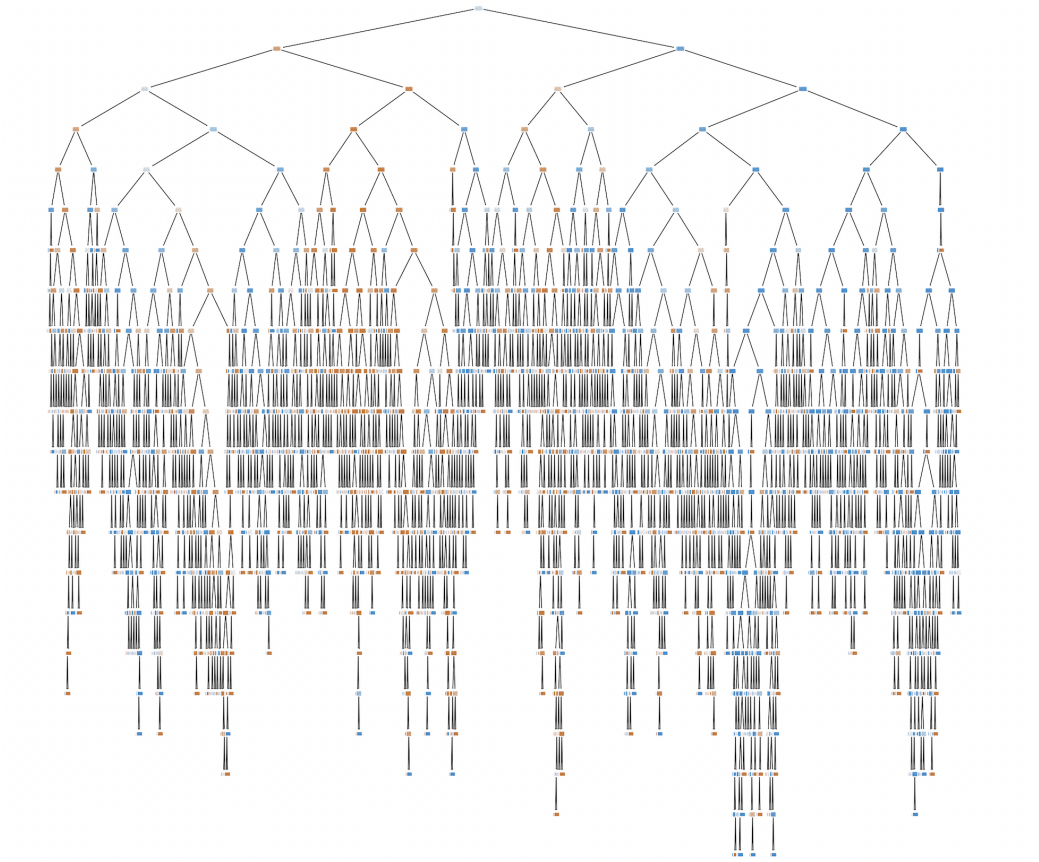


Figure 10: Illustration of one decision tree (out of 400) produced by our RFC(Word2Vec) model.

6. Discussion

The goal of this study was to explore (1) to what extent our RFC can correctly classify our articles, (2) how this compares to the performance of our human annotators and (3) whether a hybrid human-algorithm approach for decision-making can be recommended. Section 6.1 will be devoted to discussing the first and second subquestion, section 6.2 will elaborate more about explainability and interpretability, section 6.3 will present recommendations and finally, section 6.4 will draw some limitations of our study.

6.1 Performance comparison: algorithms vs annotators

Both our models performed above our accuracy baseline. The tables presented in section 5 denote that a combination of RFC with TF-IDF is the best among our models, with a weighted F1- score of 99%. The F1-scores of our model, both when used with TF-IDF

and Word2Vec, did not remarkably improve after hypertuning. It is interesting to note that TF-IDF performed outstandingly on our RFC. Thus it leads us to conclude that the choice for a particular vectorisation technique can affect a model's performance. Furthermore, TF-IDF's outstanding performance can be explained by its main idea, which assumes that the count of different words provides independent evidence of similarity without using semantic similarities between words. The extra inverse document frequency related to TF-IDF masks the contextual meaning within words that appear more frequently across other documents (Erra et al. 2015). Even though a log function smooths this penalisation, our results imply that this penalisation process may be too high. Additionally, this conclusion is in line with the study of Vijayaraghavan et al. (2020), who also concluded that TF-IDF performs better. Moreover, when compared to TF-IDF, Word2Vec does not perform well. A reason for this could be that we have built our own Word2Vec instead of using the pre-trained embeddings available online from massive corpora, such as word2vec-google-news-300. This indicates that Word2Vec needs to be built from a more extensive as well as more diverse dataset. Nevertheless, since Word2Vec captures the essential information of words and, thus, does not mask their contextual meaning, we conclude that it is a more favourable technique to use for such classification tasks.

Furthermore, our human annotators' performance shows that they perform above our accuracy baseline with an accuracy of 72% and an F1-score of 65%. However, these percentages lead to tentative conclusions since two articles are left out because they fall under the 'undecided' category. Moreover, the test set provided to our algorithms, when compared to the total set of articles provided to our annotators, is significantly larger. Thus, we should consider that one misclassification by our human annotators can heavily penalise our annotators' performance. Finally, the evaluation of our algorithms on the same selection of articles evaluated by our human annotators shows a significant outperformance. However, we must consider that the algorithms have already been trained on 27052 articles enhancing their predictive power beforehand. Hence, comparatively our human annotators perform well.

Additionally, the probability scores obtained from our algorithms in table 10 provide interesting insights related to false negatives. As stated in section 5, it is crucial to this task of fact-checking to obtain a relatively small number of false negatives since they cause more harm when spread. For example, an article spread by a person whilst thinking that it is true news will make a person become a 'walking distribution centre' of fake news - without him/her realising this. As mentioned in section 1, the spread of fake news comes with real harm to people's lives and eventually even to our democracy. Referring to the probabilities, it is evident that our algorithms produce a higher probability (and thus confidence) when classifying true articles when compared to fake articles. Moreover, it is not explainable why particular articles are classified correctly whilst still having a low confidence score (and vice versa). Evidently, ML is all about drawing predictions from uncertain data and, thus, predictions will never be perfect.

6.2 Explainability vs interpretability

Within the context of ML and AI, explainability and interpretability are often used interchangeably because they are very closely related. However, it is essential to state the differences before discussing possible solutions and perspective our results.

According to Rudin (2019), explainability is the extent to which algorithmic parameters and mechanics of an algorithm can be explained in a way understandable

for humans. The less explainable an algorithm is, the more it tends to be a black box. Rudin (2019) further states that interpretability refers to how easy it is for humans to understand an algorithm's processes to arrive at a particular outcome. In other words, the extent to which a cause and effect relationship can be perceived within a model given a change in input or parameters (Rudin 2019). For example, Deep Neural Network (DNN) algorithms, which are frequently used for extensive tasks such as fake news detection, are highly recursive and consist of millions of calculations. This inclines them to be black boxes due to their algebraic complexity (which makes them to some extent unexplainable) and thus provides little information regarding their decision-making processes (which causes them to be, to some extent, uninterpretable). But how much are these predictions worth if we don't understand the reasoning behind them?

When examining the algorithm used in our study, it is evident that an RFC is to a certain extent an explainable algorithm because RFC's, consisting of multiple decision trees, use tree representations to solve classification problems (Cutler, Cutler, and Stevens 2012). As mentioned in section 3, a significant advantage of decision trees is their transparent nature which facilitates, for example, tracing each conclusion. This makes the model very intuitive and relatively simple to interpret when visualised, providing more insight into how the decision-making process occurs. Nevertheless, it is essential to note that RFC's are interpretable as long as they are short and, thus, should not consist of hundreds of decision trees. Considering the number of trees ($N = 400$) and the depth of each tree ($N = 150$) used for our RFC's, it was evident from section 5 that a visualisation did not provide us with interpretable information considering its extensive size . Even the option of taking one decision tree out to break down the decision-making process of one article (for example the ones with low confidence score) was impossible, given the structure of the algorithm and its total size.

Hence, referring to table 10 with the probabilities, it is impossible to determine why and how our RFC came to certain confidence scores when classifying fake articles. On the contrary, when compared with our human annotators, it was possible to retrieve more information about their decision-making process. During a short post-survey chat with some of the participants, we were able to understand the rationale behind their classification of certain articles. Some of the comments are listed below:

"I found article (x) the most difficult. The content is correct, but the first sentence isn't. Can the article then directly be regarded as a fake article? Dilemma..." – Human annotator

"I found article (x) quite difficult. The first part reads as a clear piece of opinion from a blog post or something. I wasn't sure whether to judge the piece itself or the story since it seems correct. So, I just referred to it as real news because can opinions ever be fake? Unless they contain false facts, of course..." – Human annotator

The need for explainability and interpretability originates from the bias, mistakes, and lack of trust that humans have in AI systems. The 'Toeslagenaffaire', as mentioned in section 1, is an example of a recent event causing this mistrust. As Noto La Diega (2018) argued, the crux is that as long as algorithms remain unexplainable and interpretable, we should not let algorithms make decisions that can significantly impact citizens. Additionally, as stated in Wang et al. (2019) and when looking at our research, the crux in fake news and other unreliable content does not lay in the fact that there is a certain level of misinformation; instead, it is due to the underlying intent of whether it is feasible at all to label language based on its factual correctness. The difference between humans and algorithms is that humans are remarkably divergent in their

thinking, which is extremely limited within algorithms (Wang et al. 2019). For example: when giving an algorithm the task to pursue a binary classification, it will not reason about specific data points not fitting in one of the two classes and perhaps create a new class for the undecided data points. Instead, it will include the data point to the label according to the threshold provided, which does not leave room for subjectivity. This is because humans possess tacit and implicit knowledge of the world, which is difficult to express, extract, or codify in algorithms (Ellis 2008).

Furthermore, it is also interesting to note that not all articles are classified similarly within the duo's by our human annotators, considering our inter-annotator reliability rates. This emphasises the diversity in knowledge that humans have and, thus, how it can affect decision-making. A downside is that this diversity could lead to indecisiveness which is automatically tackled by algorithms. However, this argument can be disproved by setting out what comes with greater importance: being indecisive but having a thoughtful decision at the end vs speeding up decision-making without the chance of decent explainability and/or interpretability?

An important aspect to note is that even if algorithms were to be explainable and interpretable, they would still not be held accountable with respect to the decisions they make. Considering how, for example, judiciary decision-making takes place, it is evident that it is provided with an appropriate legal, ethical and value-based framework (Brand 2020). Within the rule of law, public values and interests serve as a normative framework for identifying opportunities, risks and assessing whether the legal frameworks are still future proof (Kulk et al. 2020). Examples of public values and interests that can be at stake when using algorithmic decision-making are the right to data protection, non-discrimination and legal protection (Kulk et al. 2020). As a matter of fact, Art 22 (1) of the General Data Protection Regulation (GDPR) states that automated decision-making (ADM) can only proceed under the condition that it does not produce legal effects concerning the individual. Art 22(1) goes as follows: *"any data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling"*.

However, according to Dreyer and Schulz (2019), Art 22 (1) demonstrates room for manoeuvre in the area of ADM systems since it only sets conditions for 'fully automated' decision-making and implies that systems that only prepare the basis for human decisions can be used without regulation (Dreyer and Schulz 2019). Furthermore, within the rule of law, rules and regulations are considered static and only come into life when applied to real-life problems determined by facts and circumstances (Kulk et al. 2020). However, when applied to algorithms: it seems like their 'set of rules' that have been drawn up in advance are fixed and do not require assessments for future proof. Zerilli et al. (2019) stated that this is especially unfortunate since full automation may extend to humans unconsciously implementing decisions unquestioningly, due to humans increasingly overestimating algorithms. In the end, we should stop anthropomorphising algorithms by mapping them into the same scale of intelligence as that of humans and, on top of that, stop letting algorithms make decisions that we cannot control.

Finally, our research turned out that our algorithms achieved a significant high performance when classifying the articles. Green and Chen (2019) propose an 'algorithm-in-the-loop' paradigm, which argues that algorithmic aid should enhance human decision-making, which contrasts with a 'human-in-the-loop paradigm'. However, although the 'algorithm-in-the-loop' paradigm emphasises developing systems integrated into a socio-technical context, it is not feasible given that, for example, the accelerated spread of fake news and solely fact-checking by experts being time-consuming and expensive (Oshikawa, Qian, and Wang 2018). Nevertheless, humans should never

just blindly hand over the reins to algorithms as our study portrayed that our algorithms could not make fine-grained distinctions. Needless to say, our human annotators have produced false classifications as well; however, in the case of misclassification, they would be able to justify their way of thinking/argumentation resulting in the concerned choice. Naturally, the argument can be made that algorithms are more efficient, take the least amount of execution time and memory usage possible and, hence, should be used to improve the quality of our decision-making processes. Nevertheless, although algorithmic decision-making offers us opportunities, we need to think more carefully about ethical frameworks, explainability and transparency before we, as a society, commit entirely to algorithms. Only then, as questioned in subquestion (3), a hybrid approach of both algorithms and humans can be pursued.

6.3 Recommendations

As Mittelstadt et al. (2016) argued, if 'good quality' and 'efficient' decisions produced by algorithms will later require profound damage control (such as in the case of the 'Toeslagenaffaire') - then to what extent can it be considered a good quality decision? Therefore, as argued in the study of Sakaguchi et al. (2020), an excellent focus point that should be regarded in this grand pursuit of generalisable algorithms is: to what extent can algorithms really understand what they read?

For now, we have already turned our world over to ML and algorithms, and, therefore, the next step should be to think of ways to understand better and, most importantly, manage what we have built and done. The same way we do not accept arbitrary decisions that we do not understand by people or entities, we should not accept that for algorithms. The search for explainability and interpretability within AI has been an essential field of study lately, with the European Commission's proposed regulation on AI being the most recent development in this field. Although this regulation claims to set up a human-centric framework that puts people first, which is a major step in the right direction, it still entails principal gasps regarding accountability. Therefore, the following recommendations are proposed:

- As AI systems are involved in more and more decisions that could impact citizens' lives; explainable and interpretable AI should become the norm. Therefore, research in the field of explainable and interpretable AI should continue to flourish so that we can move from a 'black box' to a constructive 'white box' approach.
- We should move towards a system of algorithmic checks and balances wherein a human-in-the-loop approach provides human interaction in every step of the virtuous ML cycle. This recommendation is twofold: 1) the humans in the loop should consist of a diverse pool considering the diversity in knowledge, views and experiences; and 2) algorithms should stop being the arbiters of final decisions to be made, considering their epistemic and normative concerns.
- Finally, and most importantly, accountability should become a critical desideratum put forward in the field of algorithmic decision-making. We see accountability here as an overarching principle indicating the obligation to justify actions produced by algorithmic decision-making. This can be put forward by establishing a national and/or European (domain-specific) legal framework, providing the parameters within

which a balance must be found between the needs, opportunities and risks that apply for algorithmic decision-making.

6.4 Limitations

In this section, we would like to point out some of the limitations of our research. One of these limitations is that due to the relatively small size of our dataset, the test set could not be split up into validation and test set. Furthermore, our dataset was relatively not large enough for generalisation, which could lead to our model excessively adjusting to our training data – and, thus, we could be fooled by overfitting. Additionally, our model was not diverse in terms of article subjects (it was politics and world-news centred), making it hard to generalise on fake news that is not necessarily centred around that subject. Moreover, looking at our human annotators, having a larger pool of annotators would produce more reliable results and interesting findings. However, considering this task being time-expensive – it was tough to find enough annotators with significant expertise. Finally, since it was not clear whether all the articles in the dataset were annotated in the way suggested by PolitiFact, it made it harder to compare the results of our annotators and the original annotators of the articles.

7. Conclusion

In conclusion, this research has aimed to investigate whether algorithmic decision-making produces better results than human decision-making. We have compared manual fact-checking by human annotators with automated fact-checking by two RFC's, modelled with TF-IDF and Word2Vec. Our results suggest that a combination of RFC and TF-IDF is the best among the two models. However, we suggest utilising Word2Vec since it captures the essential information surrounding the word to be vectorised and, thus, semantics are not lost. Furthermore, our human annotators performed above our baseline accuracy measures yet still did not perform as good as our algorithms. However, a crucial finding was that our human annotators provided us with more interpretable results by making fine-grained distinctions between the articles - and by being able to justify their way of thinking/argumentation. In contrast to our human annotators, our algorithms could not provide us with the same, and specific articles could not be investigated on their decision-making's path due to our algorithm's extensive size.

Our findings have steered our discussion into the need for more explainability and interpretability regarding algorithms and their deployment. Ideally, an 'algorithm-in-the-loop paradigm' would be recommended wherein algorithms serve as aid and should only be employed to enhance human decision making. However, considering that we have already turned our world over to ML and algorithms, this is an unthinkable reality. Finally, recommendations were presented wherein accountability was emphasised as the most crucial factor. For further research, it may be valuable to compare the performance of existing explainable algorithms with human performance to see whether responsibility can be devolved more towards explainable-algorithms regarding decision-making. In the meantime, considering that algorithms can produce significant results, we should bring the requirement of explainability and interpretability **much earlier** to the table so that the architectural design can incorporate it. This way, the callous decision to use more straightforward but explainable models with low performance - or complex black-box models with no explanations will not be at stake anymore.

References

Ahmad, Iftikhar, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. 2020. Fake news detection using machine learning ensemble methods. *Complexity*, 2020.

Ahmed, Alim Al Ayub, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting fake news using machine learning: A systematic literature review. *arXiv preprint arXiv:2102.04458*.

Ahmed, Hadeer, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138, Springer.

Arnold, Phoebe. 2020. The challenges of online fact checking. Technical report, Technical Report. Full Fact, London, UK. https://fullfact. org/media/uploads ....

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Brand, Dirk. 2020. Algorithmic decision-making and the law. *JeDEM-eJournal of eDemocracy and Open Government*, 12(1):115–132.

Chavannes, Remy, Dorien Verhulst, and Anke Strijbos. 2012. Kroniek technologie en recht.

Cutler, Adele, D Richard Cutler, and John R Stevens. 2012. Random forests. In *Ensemble machine learning*. Springer, pages 157–175.

De Laat, Paul B. 2018. Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & technology*, 31(4):525–541.

Devi, RA and K Nirmala. 2013. Construction of decision tree: Attribute selection measures. *International Journal of Advancements in Research & Technology*, 2(4):343–347.

Di Gennaro, Giovanni, Amedeo Buonanno, and Francesco AN Palmieri. 2021. Considerations about learning word2vec. *The Journal of Supercomputing*, pages 1–16.

Dreyer, Stephan and Wolfgang Schulz. 2019. The general data protection regulation and automated decision-making: Will it deliver. *Potentials And Limitations In Ensuring The Rights And Freedoms Of Individuals, Groups And Society As A Whole< https://ethicsofalgorithms. org/wp-content/uploads/sites/10/2019/01/GDPR_withoutCover-1. pdf> sitesinden*, 29:2019.

Ellis, Nick C. 2008. Implicit and explicit knowledge about language. *Encyclopedia of language and education*, 6:1–13.

Erra, Ugo, Sabrina Senatore, Fernando Minnella, and Giuseppe Caggianese. 2015. Approximate tf–idf based on topic extraction from massive message stream using the gpu. *Information Sciences*, 292:143–161.

Goh, Yeow Chong, Xin Qing Cai, Walter Theseira, Giovanni Ko, and Khiam Aik Khor. 2020. Evaluating human versus machine learning performance in classifying research abstracts. *Scientometrics*, 125(2):1197–1212.

Green, Ben and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre. 2017. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678.

Groshek, Jacob and Karolina Koc-Michalska. 2017. Helping populism win? social media use, filter bubbles, and support for populist presidential candidates in the 2016 us election campaign. *Information, Communication & Society*, 20(9):1389–1407.

Gu, Lion, Vladimir Kropotov, and Fyodor Yarochkin. 2017. The fake news machine. *How Propagandists Abuse the Internet and Manipulate the Public. Pobrane*, 25.

Harris, CR, KJ Millman, SJ van der Walt, R Gommers, P Virtanen, D Cournapeau, E Wieser, J Taylor, and S Berg. 2020. Smith 474 nj. *Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'ıo JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. Nature*, 585(7825):357–362.

Hassan, Naeemul, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812.

Hmeidi, Ismail, Mahmoud Al-Ayyoub, Nawaf A Abdulla, Abdalrahman A Almodawar, Raddad Abooraig, and Nizar A Mahyoub. 2015. Automatic arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, 41(1):114–124.

Islam, Md Zahidul, Jixue Liu, Jiuyong Li, Lin Liu, and Wei Kang. 2019. A semantics aware random forest for text classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1061–1070.

Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142, Springer.
Jones, Karen Sparck. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.
Jurafsky, Daniel and James H Martin. 2018. Speech and language processing (draft). *Chapter A: Hidden Markov Models (Draft of September 11, 2018). Retrieved March*, 19:2019.
Ko, Hoon, Jong Youl Hong, Sangheon Kim, Libor Mesicek, and In Seop Na. 2019. Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system. *Cognitive Systems Research*, 55:77–81.
Kühl, Niklas, Marc Goutier, Lucas Baier, Clemens Wolff, and Dominik Martin. 2020. Human vs. supervised machine learning: Who learns patterns faster? *arXiv preprint arXiv:2012.03661*.
Kulk, S, S van Deursen, Th Snijders, V Breemen, A Wouters, S Philipsen, M Boekema, and S Heeger. 2020. Juridische aspecten van algoritmen die besluiten nemen. Technical report, Universiteit Utrecht-Montaigne Centrum voor Rechtsstaat en Rechtspleging.
Looijenga, Maarten S. 2018. The detection of fake messages using machine learning. B.S. thesis, University of Twente.
Luhn, Hans Peter. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.
Van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
Mahir, Ehesas Mia, Saima Akhter, Mohammad Rezwanul Huq, et al. 2019. Detecting fake news using machine learning and deep learning algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5, IEEE.
McKinney, Wes et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, Austin, TX.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.
Nakashole, Ndapandula and Tom Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019.
Nieminen, Sakari and Valtteri Sankari. 2021. Checking politifact's fact-checks. *Journalism Studies*, 22(3):358–378.
Noto La Diega, Guido. 2018. Against the dehumanisation of decision-making–algorithmic decisions at the crossroads of intellectual property, data protection, and freedom of information. *Against the Dehumanisation of Decision-Making–Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information (May 31, 2018)*, 9.
Nucci, Francesco Saverio, Silvia Boi, and Massimo Magaldi. Artificial intelligence against disinformation: the fandango practical case.
Oshikawa, Ray, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.
Pasquetto, Irene V, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.
Phillips, P Jonathon and Alice J O'toole. 2014. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1):74–85.
Rehurek, Radim and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).
Ren, Qiubing, Mingchao Li, and Shuai Han. 2019. Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. *Big Earth Data*, 3(1):8–25.

Rudin, Cynthia. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sagi, Omer and Lior Rokach. 2020. Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*, 61:124–138.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Tangirala, Suryakanthi. 2020. Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm. *Int. J. Adv. Comput. Sci. Appl*, 11:612–619.

Truşcă, Maria Mihaela. 2019. Efficiency of svm classifier with word2vec and doc2vec models. In *Proceedings of the International Conference on Applied Statistics*, volume 1, pages 496–503, Sciendo.

Uysal, Alper Kursat and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.

Vicario, Michela Del, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22.

Vijayaraghavan, Sairamvinay, Ye Wang, Zhiyuan Guo, John Voong, Wenda Xu, Armand Nasseri, Jiaru Cai, Linda Li, Kevin Vuong, and Eshan Wadhwa. 2020. Fake news detection with different models. *arXiv preprint arXiv:2003.04978*.

Vlachos, Andreas and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Wang, Liqiang, Yafang Wang, Gerard de Melo, and Gerhard Weikum. 2019. Understanding archetypes of fake news via fine-grained classification. *Social Network Analysis and Mining*, 9(1):1–17.

Washington, Anne L. 2018. How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ*, 17:131.

Whitworth, Brian and Hokyoung Ryu. 2009. A comparison of human and computer information processing. In *Encyclopedia of Multimedia Technology and Networking, Second Edition*. IGI Global, pages 230–239.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Algorithmic decision-making and the control problem. *Minds and Machines*, 29(4):555–578.

Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.

Zuiderveen Borgesius, FJ, D Trilling, J Möller, S Eskens, B Bodó, CH de Vreese, N Helberger, et al. 2016. Algoritmische verzuiling en filter bubbles: een bedreiging voor de democratie? *Computerrecht*, 2016(173).

Appendices

A. Instruction document

Dear participant,

Thank you very much for agreeing to take part in this research.

My name is Oumaima Hajri, and I am currently writing my thesis for my MSc Data Science & Society about algorithmic decision-making.

This research will compare human evaluation with algorithmic evaluation, using fact-checking news as a case study.

This task concerns fact-checking a small, randomly selected selection of articles.

Please answer as openly and truthfully as you can - there are no right or wrong answers! It will take you **less than 45 minutes** to complete this task.

After finishing this task, please save it as follows: **[LASTNAME].pdf**. The deadline for this task is the **9th of May at 21:00**. Do not hesitate to send it back earlier than the proposed date and time.

Thank you again for your time and effort. Your contribution will be of the most outstanding value.

If you have any questions or comments, please feel free to contact me at:
oumaima@bitsoffreedom.nl or +316 48 27 93 30

Please take time to read the instructions on this page carefully.

Reliable and multiform journalism is of the utmost importance for a democratic society, which cannot function properly without informed citizens and a free exchange of ideas. Here are some tips on what to look out for:

1. Watch out for unusual formatting. Many false news sites have misspellings or awkward layouts. Read carefully and pay attention if you see these signs.

2. Inspect the dates. False news stories may contain timelines that make no sense or event dates that are altered.

3. Investigate the source. Ensure that the story is written by a source that you trust with a reputation for accuracy. However, don't always fall for the source since many fake news stories use accurate sources to appear 'legitimate'.

4. Check the author's sources to confirm that they are accurate. Lack of evidence or reliance on unnamed experts/statistics may indicate a false news story.

5. Is the story a joke? Sometimes, false news stories can be hard to distinguish from humour or satire. Check if the story's details and tone suggest that it may be just for fun.

B. Example of classification sheet (here, the right answers are provided)

Tilburg University

BITS OF FREEDOM
Voor jouw internetvrijheid

Please check the box relevant to your answer

Article 1
DAKAR/BAMAKO (Reuters) - Five soldiers from Niger and three U.S. Army Special Forces troops were killed and two wounded in an ambush on a joint patrol in southwest Niger on Wednesday, according to Nigerien and U.S. officials. The five Green Berets were attacked while on a routine patrol in an area known to have a presence of insurgents, including from al Qaeda in Islamic Maghreb (AQIM) and Islamic State, a U.S. official told Reuters. It was unclear who fired on the U.S. and U.S.-backed forces, the official said. Those forces were not patrolling the area with any specific objective, such as a high-value target or rescuing a hostage, the official added. A spokesman for U.S. Africa Command confirmed the attack after Radio France International (RFI) reported a lethal ambush near the Niger/Mali border. "We can confirm reports that a joint U.S. and Nigerien patrol came under hostile fire in southwest Niger", said the spokesman. Namatta Abubacar, an official for the region of Tillaberi in Niger, said five Nigerien soldiers were among the dead. A Niger diplomatic source said the attackers had come from Mali and had killed several soldiers, without saying whether any of the U.S. troops stationed in the West African country were among the victims. U.S. President Donald Trump was briefed by telephone on the attack by White House Chief of Staff John Kelly while Trump flew back on-Air Force One from Las Vegas, where he had been visiting victims and first responders affected by Sunday s mass shooting. RFI said earlier on Wednesday a counterattack was underway. African security forces backed by Western troops are stepping up efforts to counter jihadist groups forming part of a growing regional insurgency in the poor, sparsely populated deserts of the Sahel. A relatively new militant group called Islamic State in the Greater Sahara has claimed some of the attacks. Geoff D. Porter, head of North Africa Risk Consulting, said that any confirmation of Islamic State's role in Wednesday's strike would lead to a strategic shift from Libya toward the Sahel band, stretching eastwards from Senegal to Chad. The U.S. Africa Command has hundreds of soldiers deployed across the region, including at an air facility in Agadez, and offers training and support to Niger's army in aspects such as intelligence gathering and surveillance.
FAKE NEWS ARTICLE ☐
TRUE NEWS ARTICLE ☒

Please check the box relevant to your answer

Article 2
Australian Senator Larissa Waters recently went viral after she made history by becoming the first woman to breastfeed in the country s Parliament. But Waters resigned today after it was revealed that she has dual citizenship in Australia and Canada, a breach of Australia's constitution for sitting senators. Waters, a member of Australia's Green Party, became an international sensation in May when she breastfed her two-month old daughter in Australia's parliament. The country had only just legalized the practice in 2016, paving the way for a more family-friendly environment in Australian politics. Waters was born in Winnipeg in 1977 to two Australian parents and moved down under when she was just 11 months old. Australia's constitution forbids anyone with dual citizenship from serving as a senator. Waters, who says she was unaware that she held citizenship in Canada, was elected in 2010. "I was devastated to learn that because of 70-year-old Canadian laws I had been a dual citizen from birth, and that Canadian law changed a week after I was born and required me to have actively renounced Canadian citizenship", Waters said at a press conference today. "It is with a heavy heart that I am forced to resign as senator for Queensland and co-deputy leader of the Australian Greens, effective today", she continued. Waters is the second Australian senator who's had to resign in as many weeks.
FAKE NEWS ARTICLE ☒
TRUE NEWS ARTICLE ☐

Please check the box relevant to your answer

Article 3
Wikileaks released an email from Center for American Progress President Neera Tanden, coaching Hillary on how to best gain the trust of the Black community by going directly to the victims being held up by Black Lives Matter movement as heroes, and faking empathy with the parents. She also mentions it would be a good idea to use the idea that their grief should be magnified because it happened at the hands of law enforcement officers (the state). The email was forwarded by Jen Palmieri who is the same person who was busted criticizing Catholics. Karen Finney, Hillary's Senior Advisor for Communications and Political Outreach, & Senior Spokesperson for Hillary for America was also copied on the suggestion about how to fake concern for parents of Black Lives Matter victims. Here is an excerpt from the Wikileaks email: *Hillary clearly took her advice, as she can been seen milking her phony support for mothers of Black Lives Matter victims, by giving them a chance to appear at the DNC as her special guests to chant:* ***"Ever the actress, here's Hillary appealing to the black community with her newfound black dialect."***
FAKE NEWS ARTICLE ☒
TRUE NEWS ARTICLE ☐

Please check the box relevant to your answer

Article 4
ISTANBUL (Reuters) - "Decisions made at the approaching meeting of the Organisation of Islamic Cooperation (OIC) will show that U.S. recognition of Jerusalem as the Israeli capital will not be easy to implement", Turkish President Tayyip Erdogan said on Sunday. A spokesman for Erdogan on Wednesday had announced that the OIC would a hold an urgent meeting in Turkey on the 13th of December to coordinate a response to the decision by the United States. The OIC, established in 1969, consists of 57 member states with a Muslim majority or a large Muslim population. "We explained to all our interlocutors that the United States' decision does not comply with international law, diplomacy or humanity", Erdogan said at a Justice and Development Party (AKP) assembly in Turkey's central province of Sivas, referring to phone calls he made to leaders including French President Emmanuel Macron and the Pope. "With the roadmap we will create during the OIC meeting, we will show that the decision will not be easy to implement", he said, adding that Turkey considers U.S. President Trump's Jerusalem announcement void. The Arab League, in a statement issued after an emergency session in Cairo on Saturday, called the announcement a dangerous violation of international law and said it would seek a U.N. Security Council resolution rejecting the U.S. move. The Arab League, which consists of Arabic-speaking nations, currently has 22 active member states. Trump's announcement has also upset U.S. allies in the West. At the United Nations, France, Italy, Germany, Britain and Sweden called on the United States to bring forward detailed proposals for an Israeli-Palestinian settlement. Palestinians took to the streets after the U.S. announcement. Demonstrations also took place in Iran, Jordan, Lebanon, Tunisia, Somalia, Yemen, Malaysia and Indonesia, as well as outside the U.S. Embassy in Berlin.
FAKE NEWS ARTICLE ☐
TRUE NEWS ARTICLE ☒

Please check the box relevant to your answer

Article 5
Iran may have received an additional $33.6 billion in secret cash and gold payments facilitated by the Obama administration between 2014 and 2016, according to testimony provided before Congress by an expert on last summer's nuclear agreement with Iran. Between January 2014 and July 2015, when the Obama administration was hammering out the final details of the nuclear accord, Iran was paid $700 million every month from funds that had previously been frozen by U.S. sanctions. A total of $11.9 billion was ultimately paid to Iran, but the details surrounding these payments remain shrouded in mystery, according to Mark Dubowitz, executive director at the Foundation for Defense of Democracies. In total, Iran may have received as much as $33.6 billion in cash or in gold and other precious metals, Dubowitz disclosed. New questions about these payments are emerging following confirmation from top Obama administration officials on Thursday that it was forced to pay Iran $1.7 billion in cash prior to the release of several U.S. hostages earlier this year. The administration insisted that cash had to be used for this payment.
FAKE NEWS ARTICLE ☒
TRUE NEWS ARTICLE ☐

Please check the box relevant to your answer

Article 6
A Hezbollah member of the Lebanese parliament said on Thursday that proposed new U.S. sanctions against the powerful Iran-backed group aimed to provoke unrest in Lebanon. The U.S. House of Representatives on Wednesday endorsed new sanctions on the Shiite Hezbollah militia, part of an effort to increase pressure on Iran. The new sanctions have not yet become law. "The sanctions law (...) is a blatant interference in Lebanese internal affairs, a violation of its national sovereignty and an unacceptable targeting of the Lebanese people", Hezbollah parliamentarian Hassan Fadlallah said in a televised statement. America aims, through this aggressive behavior in legislation, to subjugate Lebanon, to stir unrest and deprive its people of development, Fadlallah said. Hezbollah is in Lebanon's delicate, national unity government and fights alongside Syrian President Bashar al-Assad in Syria s more than six-year long conflict. It is classified as a terrorist group by Washington and on Wednesday the House of Representatives passed a resolution urging the European Union to do the same. One of the measures passed by the House of Representatives was an amendment strengthening the 2015 U.S. Hezbollah International Financing Prevention Act (HIFPA) which aimed to sever the group's global funding networks. When HIFPA was introduced it caused alarm in Beirut where the government feared major damage to the banking sector that underpins Lebanon s economy. But Lebanon's central bank Governor Riad Salameh told Reuters on Tuesday Lebanon had mechanisms already in place to deal with any new sanctions. Salameh also said this week the American Treasury appeared content with how Lebanon was applying sanctions regulations. They consider the measures which Lebanon s central bank has put in place to be sufficient, he said after a visit to the U.S., in a statement distributed by Lebanon's presidential media office.
FAKE NEWS ARTICLE ☐
TRUE NEWS ARTICLE ☒