IDENTIFYING FAKE NEWS ON TWITTER USING NAÏVE BAYES, SVM AND RANDOM FOREST DISTRIBUTED ALGORITHMS

CIPRIAN-GABRIEL CUȘMALIUC, LUCIA-GEORGIANA COCA, ADRIAN IFTENE

"Alexandru Ioan Cuza" University of Iasi, Faculty of Computer Science
{cusmuliuc.ciprian.gabriel, coca.lucia.georgiana, adiftene}@info.uaic.ro

Abstract

In our days using social networks like Facebook or Twitter in order to find news is something usual. Users from these networks create a lot of content: posts (with text, images, videos, and links), comments, likes, but also redistribution of information with retweet option or with simple copy/paste operations. Nowadays, a common problem is the detection of fake users and of fake news and it is often difficult to distinguish the credibility of news that appears in a social networking. Sometimes users are in the position of believing the certain news at their first appearance, only because the information is written in the network and they do not check to see if it's true or not. The present paper aims to analyze different techniques of detecting fake news, their performance and how we can do fine tuning in order to improve the actual results.

Key words — Fake news, Naïve Bayes, Random Forest, SVM, Twitter.

1. Introduction

The popularity of social networks has increased significantly in recent years, and reading news on these social networks has become a natural activity for all users. The news is instantly transmitted to these networks, which are read quickly, marked with opinions (see Facebook), retransmitted (retweet on Twitter, share on Facebook) without having to check many times whether they are true or false news.

Over the years there have been different approaches to the classification of fake news. From *linguistic approaches* (with machine learning) to *social network approaches* researchers apply different techniques to identify fake news (Conroy et al., 2015). In *linguistic approaches*, statistics on n-grams were created and analyzed to identify fake information (Hadeer et al., 2017), or sentences were transformed into more advanced forms of information representation (such as parsing trees), which then analyze probabilities attached to identify anomalies (Perez-Rosas et al., 2018), or it analyses semantically the contents of a user's statements, constructs pairs of the attribute form: descriptor and calculates compatibility scores (Shu et al., 2018), or relations between the linguistic elements are built, which help determine the proximity to the centers of truth or deception (Popoola, 2017; Rubin and Lukaianova, 2018), or use neural networks that identify fast the fake news (Sneha et al., 2018), or similar to our approach, SVM classifiers or Naïve Bayesian-type classifiers are used to predict future clutter-based fraud and distances in (Rubin et al., 2016; Singh et al., 2018). In *social networking approaches*, knowledge networks are exploited to identify the lie (Idehen, 2017), or the fact that users

are forced to authenticate when using the social network, provides increased confidence to the data that appears here (Shu et al., 2018; Wu and Liu, 2018).

Our proposal is similar with work presented in (Strehl et al., 2000), but we increase the data set and the quality of it. The authors used SVM and they created a dataset of 345 valid news articles, dataset including an equal number of news reports from three well known and largely respected news agencies: *National Public Radio, New York Times*, and *Public Broadcasting Corporation*. The most important features used by their SVM algorithm are the following: *author, published, title, text, language, crawled, site_url, country, thread_title, spam_score, main_img*. Their experiments obtained an overall accuracy of 0.87. Additional to increasing the data set, in our experiments we implement more algorithms, like Random Forest, SVM and Naïve Bayes in order to see which is best for this problem.

In Chapter 2 we will see details about our data and the implemented algorithms. Chapter 3 presents the system architecture and results obtained with the algorithms. Chapter 4 presents relevant use cases and analysis performed on tweets with correct classification and those with incorrect classification. The last chapter contains the conclusions of this paper and future work.

2. Proposed solution

2.1. Data set

In order to apply our classification techniques, we built a dataset that was annotated manually with the help of human annotators. All the data was collected from Twitter from a period of one month, in August 2011. The data was not doubly annotated, the criteria used by the humans in order to label a tweet were related to the type of tweet, if the tweet was **normal chatter** and the user was real (e.g.: had followers, real name and some past tweets) they would label it as not-fake, if the tweet was a **news segment** or an **ad** a thorough investigation of the tweet would have been done in order to decide if the tweet information was real (the user analysis was also relevant, if it seemed like a fake account there were high chances of labeling the tweet as false). In some cases, one annotator would ask another for opinion in order to best classify a tweet; the dataset used makes no exception from subjective decisions. In the end, there were about 10.000 tweets, that were split into two categories, training and test, they were manually labelled into fake or not and they will be the basis of all the classifications that will be done from here on.

The tweet content is of multiple types: (1) **normal chatter** that can be found on twitter (e.g.: *Good morning friends! Manila looks like London again! Bring out the wellies! Stay dry!*), (2) **news** (e.g.: *London Riots #riots: send in water cannon to clear streets_ Theresa May told: Theresa May_ the Home Secretary_ was where... bio link @ http://bit.ly/qRNBkO*), (3) **other ads** (e.g.: *#jJOB: Expert Witness Director (Forensic Delay background) in London_ West End_ bio link @ :P :P http://www.rengineeringjobs.com/career/ 311991*) and (4) **fake news** (e.g.: *We are at #LondonRiotsZoo Attacking the animal at Zoo @a_fisho @blckBss @LondonDude NICE ONE_ Like we for wedge coolly for GH. Man come dey chase #LondonRiotszoo bus!!!!!*). This variety of tweets, that include personal touches, but also impersonal speech and fake

news, creates a diversity that is very characteristic of the Twitter platform and is very important for the classification algorithms, in order to be able to learn from all possible inputs and scenarios.

In order to not be biased for the algorithm, toward users with high favorites and retweets count (i.e.: *only the very popular on Twitter say the truth*) the dataset does not contain outliers; the highest retweet count is around 200 and the favorites are around 100, very natural numbers for average twitter users (from various sources, the average being 278 retweets). Also, users that use many hashtags and reference other users in their tweets are more likely to tell the truth, as what they say is based either on the hashtag (that might be, for example, a trend) or in response to another user. Like email spam filters, the algorithm calculates how many tweets contain spam word, in order to increase the elimination rate.

2.2. Cleaning and pre-processing of data

However various modifications had to be done to these datasets in order to have consistent data. Such modifications included: discard of tweets that: were not in English, do not have all the attributes necessary (were incomplete), were malformed, had invalid characters, were clearly outliers that would have strongly affected any algorithm or would have led to overfitting. From initial 15.468 raw tweets considered using (Gupta and Kumaraguru, 2012), the processing resulted in about 10.000 tweets that have been splint 70-30, meaning for 7.000 training and 3.000 for test. This pre-processing was done manually, by eliminating the unwanted tweets.

The pre-processing technique uses NLP in order to create a metric of the tweet. The algorithm considers *how many favorites the tweet has, retweets, hashtags, mentions of other users, if it has any URL, if the user swears in the tweet and how many words could be spam* (that are based on statistical evidence, they could be spam, e.g.: "F R E E").

2.3. Tweets features

The main tweet features were: *Date, Tweet_Text, Tweet_Id, User_Id, User_Name, User_Screen_ Name, Retweets, Favorites, Class*.

An example of such a tweet is the following: (Date: *2011-08-06 19:59:59*, which is in standard date format, Tweet_Text: "*RT @joeashtonsing New Guidelines for Product Branding at London 2012 Olympics - Bike Europe http://bit.ly/ pFRMqG*", which is the tweet content, Tweet_Id: *98900000000000000*, User_Id: *67898811*, User_Name: *The London bike bot*, User_Screen_Name: *67898811*, Retweets: *2*, Favorites: *3*, Class: *1*, which represent the fact that the tweet was manual classified, and it represents a fake tweet).

2.4. Algorithms

Testing the model involves selecting appropriate algorithm for the task at hand. The chosen algorithms are: Random Forest, Support Vector Machines and Naïve Bayes, all of which will be tested using different parameters in order to measure the accuracy and performance.

The classification algorithms are backed by Spark[1], a parallel environment, using different techniques provided by the platform and multiple machines communicating in the network, a distributed execution is possible, training and classifying data at a much faster pace than traditional execution environments.

3. *System architecture*

The system architecture is composed of: the algorithmic part, that is implemented in Python using PySpark, a Flume[2] agent that collects data from twitter and inserts it into a database (see Figure 1) and a Java web service that users can make request on, in order to collect and classify random tweets from the social platform. The proposed solution follows a Service Oriented Architecture, as we want to have a decoupled system in place.

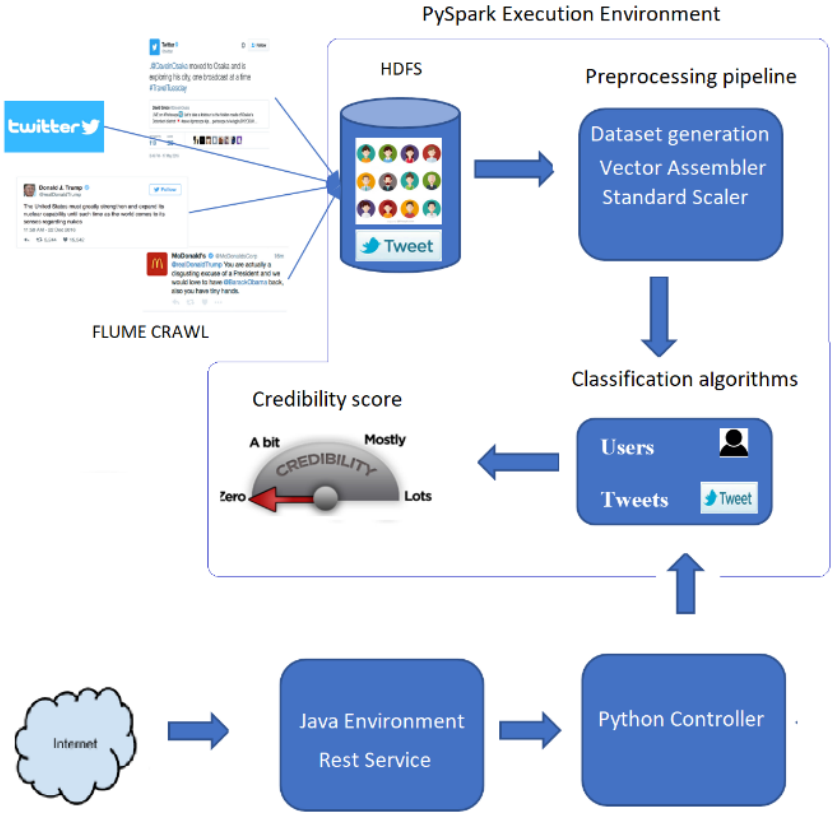

Figure 1: System architecture

3.1. *Naïve Bayes*

Naïve Bayes by default uses the "multinomial Naïve Bayes" internally, as it classifies data based on every input feature (*Date, Tweet_Text, Tweet_Id, User_Id, User_Name, User_Screen_ Name, Retweets, and Favorites*). The Naïve Bayes classifier considers each of these features to contribute independently to the probability. The smoothing was

[1] https://spark.apache.org/
[2] http://flume.apache.org/

set to a default value of 1. Naïve Bayes resulted in an accuracy of 92.43% and it is very fast, on our data with 10.000 tweets (split into 70% for training and 30% for testing). The method is very fast, the execution time being around 2 seconds. The results are presented in Figure 2 (left), where with *red*: we mark Non-Fake tweets from training data set, with *green*: Fake tweets from training data set, with *black*: we mark Non-Fake tweets from test data set and with blue: we mark Fake tweets from test data set.

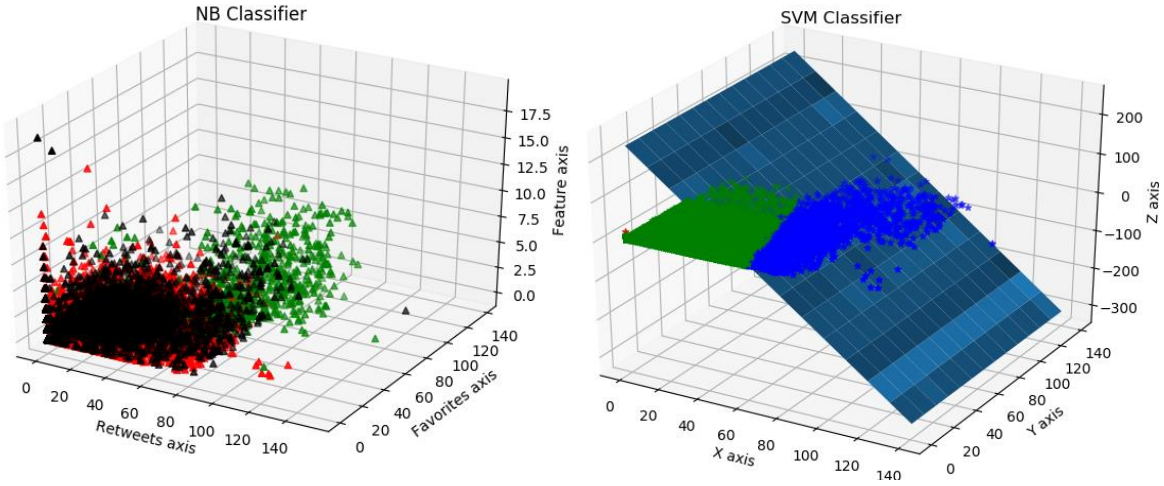

Figure 2: Results obtained with Naïve Bayes classifier (left) and with SVM classifier (right)

The confusion matrix for Naïve Bayes classifier on about 3 thousand test data is presented below in Table 1.

Table 1: The confusion matrix for Naïve Bayes classifier

Naïve Bayes	Predicted No	Predicted Yes
Actual No	135	133
Actual Yes	94	2638

3.2. Support Vector Machine

The Linear SVM Classifier uses a hyperplane to segregate the classes as Fake or Non-Fake in the given training data, i.e., given the labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes the test data as Fake or Non-Fake. The SVM is using a linear kernel, the decision function shape is set to "one vs rest" and the error threshold was set to 2^(-5) in order for the algorithm to perform faster, as a higher error was considered acceptable as long as the execution time was very good. The accuracy of our SVM classifier 95% and the execution time is 8.98 seconds. In Figure 2 (right), we mark with *green*: Non-Fake tweets from test data set, and we mark with *blue*: Fake tweets from test data set. The confusion matrix for SVM algorithm on about 3 thousand test data is in Table 2.

Table 2: The confusion matrix for SVM classifier

SVM	Predicted No	Predicted Yes
Actual No	148	71
Actual Yes	79	2702

3.3. Random Forest

The Random Forest classifier operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random decision forests correct for decision trees habit of overfitting to their training set.

The Random Forest used was set to have a number of 10 decision trees, the impurity was set to gini[3] index and the maximum depth was left to a default of 5. The execution time is 3.5 seconds and the accuracy is 95.93%. In Figure 3, we have the results, where with *red*: we mark Non-Fake tweets from training data set, with *green*: Fake tweets from training data set, with *black*: we mark Non-Fake tweets from test data set and with blue: we mark Fake tweets from test data set. The confusion matrix on about 3 thousand test data is in Table 3.

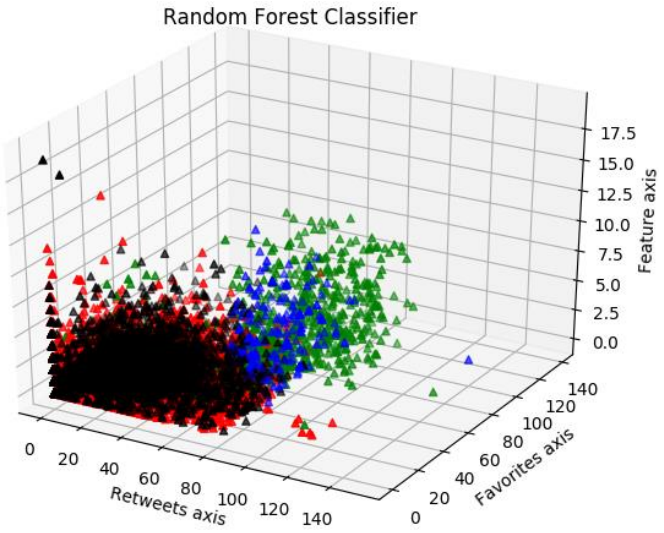

Figure 3: Results obtained with Random Forest classifier

Table 3: The confusion matrix for Random Forest classifier

Random Forest	Predicted No	Predicted Yes
Actual No	167	60
Actual Yes	62	2711

[3] https://www.investopedia.com/terms/g/gini-index.asp

3.4. Results

Overall, the accuracy for all three methods is presented in Table 4. We can see how the best result was provided by Random Forest algorithm, followed by SVM algorithm.

Table 4: Overall accuracy

Algorithm	**Naïve Bayes**	**SVM**	**Random Forest**
Accuracy	92.43%	95%	95.93%

3.5. Similar solutions

One of the main similar applications available is the Thomson Reuters' Reuters News Tracer[4], which is a proprietary algorithm which looks at over 700 features in order to determine whether a tweet is truthful or not. The claimed accuracy is 78 percent for tweets, hence slightly lower than our own algorithm. Unfortunately, the algorithm behind the Reuters News Tracer is proprietary, hence cannot be compared on the same dataset to view key differences.

Most solutions, such as widely popular B.S. Detector[5] rely on curated datasets instead of machine learning, hence classifying the site and not the content itself, which may be an issue depending on how the classification may be awarded. Some look for indicators of sensationalism in the titles in order to detect whether it might be a title meant for user to follow the link to the website. This is done using natural language processing, and is limited with the matter currently analyzed, with no context regarding any trends it may be a part of.

The approach closest to the one we propose is the use of Deep Neural Networks (Bajaj, 2017). These allow classifications by looking at multiple features. These types of applications seem to have better results than our solution (for example, in (Bajaj, 2017) authors report a precision of 97%).

4. Use Cases

In next use cases, we will see when our algorithms classify correctly the tweets in fake or non-fake news, and cases when our algorithms incorrectly classify them. We also explain the reasons for which our algorithms decide the type of the tweet: *real* or *fake*.

4.1. Cases with correct classification

Tweets with real information

For tweet with details: Tweet_Text: "*Moments ago, President @realDonaldTrump answered reporter questions about DACA as a government shutdown looms. http://fxn.ws/2FB0WFa*", Date: "*2018-01-15 3:00:21*", User_Screen_Name: "*foxnews*", Retweets: "*101*", Favorites: "*260*" our system decide that the information is real. The

[4] https://blogs.thomsonreuters.com/answerson/ making-reuters-news-tracer/

[5] http://bsdetector.tech/

tweet is correctly labelled because it refers to an actual news, fact that the US president stated some information about a popular topic, DACA. The fact that Fox News is one of the biggest news stations in US weighs a lot in the algorithm final decision (i.e. over 17.1 million of followers).

The tweet with details: Tweet_Text: "'*We thank President Trump,'say #IranProtesters, who praise the American public for their support. Here\'s my exclusive with #Iranprotest activists who vow to bring freedom to #Iran and look toward the #USA and @POTUS @realDonaldTrump for help: Watch:*", Date: "*2018-01-15 3:03:16*", User_Screen_Name: "*foxnews*", Retweets: "*81*", Favorites: "*216*" is correctly labelled, since it refers to a pretty trending subject at the time being. However, the algorithm decision was also biased by the fact that the "*foxnews*" user is pretty popular amongst the social environment in cause, which is a genuine fact, at a first glance upon the user's number of followers or the posts' number of retweets and favorites.

For tweet with details: Tweet_Text: "*Dear @POTUS: If you want to honor our troops, here are things more useful than an expensive #militaryparade: - Have a coherent N Korea strategy, - Nominate US Amb to S Korea, - Have a coherent Afghanistan strategy (we've been in 16 years), - Have a coherent Syria strategy, - Deter Russia, https://t.co/Y2NdFaYdjw*", Date: "*2018-02-17 3:12:21*", User_Screen_Name: "*tedlieu*", Retweets: "*4345*", Favorites: "*10726*", due to the high number of retweets and favourites the algorithm classifies it as real. Also, the writing is very correct and probably the term "*#militaryparade*" was very popular at the moment. The user "*tedlieu*" may also not be reported as a spammer in our dataset.

For tweet with details Tweet_Text: "*There are tens of thousands of homeless veterans on our streets, rather than waste millions of dollars on a #militaryparade use that money to help the women & men that fought for us.*", Date: "*2018-02-07 3:00:00*", User_Screen_Name: "*sahluwal*", Retweets: "*472*", Favorites: "*956*", our algorithms correct identify the fact that it is very popular and also it's associated with a very popular hashtag. The writing is correct probably there are other non-fake tweets in the database that it can match.

For tweet with details: Tweet_Text: "*A message from Paul Simon: February 5, 2018 https://t.co/ kdNRIgKswR https://t.co/EFq3Ry4cUp*", Date: "*2018-02-05 3:00:00*", User_Screen_Name: "*gavinandersonn*", Retweets: "*50*", Favorites: "*0*", the tweet portrayed above is retweet. The original tweet does not have any exclamations of signs of it being a baiting message (keywords suggesting something extraordinary). Looking at the original message, the author of the tweet seems to be posting about himself. As such, users would not generally lie about themselves. The author seems to have a rather large following, having approximately 20 thousand followers, therefore being trusted by the same amount of people. The retweet does not add anything to the original message and the user retweeting has a small number of followers: roughly 300. Comparing to the tweet related to *Trayvon Martin* (from next section), the use case seems rather similar: One regular user with a small/average number of followers picks up and distributes to his followers the message of a rather popular account, both having a link attached. This being said, the contents of the two messages are rather far apart: this one has a neutral tone, labelling the attachment, while the next one had a sad tone and presented a keyword which

would definitely pick up attention: king. Therefore, the link in the aforementioned tweet is considered to be a bait link to a fake news story, while the one in this tweet is not.

Tweets with fake information

The tweet with details: Tweet_Text: "*A cannabis business in Canada is now accepting @StellarOrg $XLM as a form of payment.*", Date: "*2018-01-14 14:05:33*", User_Screen_Name: "*StellarRumors*", Retweets: "*7*", Favorites: "*14*" has all the trademarks of a fake message. The user doesn't have a lot of followers (i.e. 89) and the tweet has only 7 retweets. The previous pieces of information influence, in a great measure, the label that will be provided by the classification algorithm. The fact that the tweet doesn't contains any link that can be used to verify the authenticity of the information along with the lack of other Twitter accounts references, weighs in the algorithm decision.

The tweet with details: Tweet_Text: "*(Audrey) I champion #Moms #Women #Girls the #Homeless becoming #financially self sufficient under 2018 #Business #Tax laws & regulations! #workfromhome #Resist*", Date: "*2018-01-31 3:00:00*", User_Screen_Name: "*especiallyme50*", Retweets: "*0*", Favorites: "*0*" has the lack of sense and also the lack of interest of the public and suggest the fact that the tweet is shady. Also, the very often use of hashtags makes the algorithm think this is a spam and mark the tweet as fake.

The tweet with details Tweet_Text: "*Must read! #WednesdayWisdom #Military #Budget2018 https://t.co/E0wKI1KMsq*", Date: "*2018-02-07 3:45:21*", User_Screen_Name: "*momofmonday*", Retweets: "*3*", Favorites: "*3*" has a very low number of retweets and of favorites. The tweet is also very short, isn't comprehensive, only some exclamation marks and some hashtags and a link, the algorithm classifies it as clickbait. It is possible that the user is also known in the database as a frequent spammer.

The tweet with details Tweet_Text: "*AMAZING to watch #Democrats & other #Liberals REJOICING over the drop in The Dow Jones Industrial Average. https://t.co/udyQibJnoD*", Date: "*2018-02-05 20:38:36*", User_Screen_ Name: "*T_W_Haines*", Retweets: "*0*", Favorites: "*0*" has all the trademarks of a fake message. The usage of capital letters while referring to data. The user has a rather large number of followers, which means that the classification is focused on the message, not on the user himself. The tweet does have no retweets, asserting its controversial status. Also, it refers to a rather large population doing a certain activity, which is rather unverifiable. Although the post is satirical, it is definitely misleading to a user not familiar to the author. Given the negative connotation of the message and the fact that is talks about a large category of individuals, whose actions cannot be properly and profoundly verified, this leads to the conclusion that the message is most likely a fake news. The most identifying features of the messages are the capitalized keywords, which the author uses in order to express amazement, while it could also be interpreted as an extraordinary situation under which the subjects of the tweet did not act as expected of them.

The tweet with details: Tweet_Text: "*Mysterious 'black snow' in Kazakhstan sparks investigation (PHOTOS, VIDEOS) https://www.rt. com/news/415781-kazakhstan - pollution-black-snow/ ... #News #Bibleprophecy #Truth #Knowledge #Wisdom*

#Economist #Endtimes", Date: "*2018-01-14 16:21:29*", User_Screen_Name: "*clintonkowach*", Retweets: "*5*", Favorites: "*5*" has the same type of problems like previous fake tweets: low number of retweets and favorites, low number of user's followers makes this tweet marked as fake. Since these are the features that mostly affect our decision, it's trustworthy to believe that this label is genuine.

4.2. Cases with incorrect classification

Tweets with real information

For the tweet with details: Tweet_Text: *"#NEWS: Exoplanets found to orbit iron-rich stars tightly: http://tinyurl.com/y8p2k2dy*", Date: "*2018-01-14 14:23:23*", User_Screen_Name: "*spaceanswers*", Retweets: "*2*", Favorites: "*1*", the incorrect tag regarding the falsehood of the news' veracity marks the above case as negative labelled, since the exposed subject is not popular/appealing among users. This situation is a typical event, where trends and popularity are a big factor of influence with respect to the news' genuineness verdict. The very low number of retweets and favorites, along with the user's inconspicuousness, makes this piece of news (originally published by NASA) a perfect, common example of incorrect labelling cause.

The tweet with text: Tweet_Text: "*The citizens of #Hodeidah #Yemen suffer greatly bcz they lost their #jobs due to #Saudi #US #UAE blockade, which closed the port of Hodeidah and made them suffer #news #media #press #1000DaysOfWarOnYemen #UN #UNCIF #Trump #FightForRights #FixTrumpIn5Words*", Date: "2018-01-14 10:18:52", User_Screen_Name: "_0YemenPrincess", Retweets: "*17*", Favorites: "*5*" is also incorrectly labelled. It contains accurate information about a current crisis that occurred in Yemen. Even if the text contains a lot of hashtags (e.g. *#1000DaysOfWarOnYemen, #FightFor Rights*) that increase the tweet credibility, the algorithm decision is strongly biased by the reduced number of retweets and followers.

For the tweet with details: Tweet_Text: "*Happy #NationalSigningDay y'all. Friendly reminder: These are high school kids choosing a college. DO NOT threaten or tweet insults at teenagers because they didn't choose the school you, an adult on twitter, wanted. With that said, #GoVols #NSD18*", Date: "*2018-02-07 17:00:22*", User_Screen_ Name: "*dianneg*", Retweets: "*58*", Favorites: "*234*", one thing that pops is the excessive use of hashtags, the algorithm might have been biased to thing lots of hashtags are associated with fake news. Also the writing is a bit odd, the user uses "*DO NOT*" and "*y'all*". The user may be a bit unknown and the tweet doesn't have a lot of retweets or favorites, its average.

For the tweet with details: Tweet_Text: "*추운_겨울의_봄이_되어준_우현아_생일축하해*", Date: "*2018-02-07 3:00:00*", User_ Screen_Name: "*wooohyunz*", Retweets: "*12*", Favorites: "*136*", the algorithm doesn't know Korean and this is the main reason for its classification as fake. Second of all the post has a small number of retweets and favorites, and when the algorithm classifies, it matches the tweet with a fake one which is in its database.

The tweet with details: Tweet_Text: "*RT @HipHopDX: Happy Birthday to the young king #TrayvonMartin. Trayvon would have been 23 years old today. Our thoughts and*

prayers are with…", Date: "*2018-02-05 20:38:27*", User_ Screen_Name: "*7daysToxic*", Retweets: "*45*", Favorites: "*0*", was classified as fake tweet. Looking at the text of the tweet, we notice that the tone is rather toned down. There are no exclamations, thus not suggesting something sensational, neither is the formulation suggestive by any means. Also, the account the original tweet came from is an account with over 500 thousand followers, suggesting it is not an account made for click baiting users. The message does not appear to be any news story. The only word that seems out of place is *king*. As such, the algorithm probably considers the tweet to be related to some monarch, deeming it something which would catch attention rapidly. Diving deeper, this is a tweet offering condolences. Correlating with the keyword aforementioned (*king*) and with the negative sentiment surrounding the tweet (due to it being a condolences message), the machine learning algorithm considers the tweet as something related to the *king's* death. The user retweeting the message does not have a large number of followers and, consequently, the algorithm considers the tweet as a fake message.

Tweets with fake information

The tweet with details: Tweet_Text: *"@6BillionPeople @PACcoinOfficial @PACcoinMan\n50k PAC giveaway!! Today ONLY. Contests ends at 12:01 am January 15!! \n2 winners 25k each!! Let's welcome the launch of PACFYLE!!! \n Share, retweet and put your wallet # in comments.\n #BlockChain #giveaway #bitcoin\n#PacCoin #News\n#PAC*", Date: "*2018-01-14 11:15:42*", User_Screen_Name: "*PACcoinCZAR*", Retweets: "*94*", Favorites: "*55*" is incorrectly labelled as real message. This message has all the trademarks of a real message. It contains information about an alleged contest, the required steps that need to be fulfilled to take part and the prize of the contest. Because the user has 3.000 followers along with the fact that the tweet has only 94 retweets in a short time period, biased the algorithm to provide the given label.

The tweet with details: Tweet_Text: *"#Iran #News: urgent, For The world to see,\n This is the reality of Islamic Republic of #Khamenei \n Down with dictator #FreeIran #Iranprotests #Regime Change #UNGA #HumanRights @Asma_ Jahangir @realDonaldTrump @nikkihaley*", Date: "*2018-01-15 3:08:52*", User_Screen_Name: "*IranNewsUpdate1*", Retweets: "*220*", Favorites: "*153*" has all the odds of being marked as genuine (numbers are high on any feature our algorithm analyses). The fact that the tweet is short and contains lots of hashtags, may produce a misfit (as in current example), which is error prone.

The tweet with details: Tweet_Text: "*That dark-skinned #Briton story is more complex than it appears at first glance. Sure the scientists have concluded that Cheddar Man had black skin, but I've just seen a comment from GaryBrexit88 saying it's bollocks so I don't know who to believe.*", Date: "*2018-02-07 3:32:00*", User_Screen_Name: "*somegreybloke*", Retweets: "*1033*", Favorites: "*2491*" has the hashtag *#Briton*, which suggests that, as said in the tweet Cheddar Man had black skin. The hashtag is very popular but this tweet seems to contradict everybody. The algorithm assumes it is ok because of high public interest and the very well written tweet, also the *#Briton* have influenced it.

The tweet with details: Tweet_Text: "*DONT FUCK UP YOUR LIFE #WednesdayWisdom*", Date: "*2018-02-07 3:45:21*", User_Screen_Name: "*the_*

ironsheik", Retweets: "*246*", Favorites: "*461*" uses profanity and the fact that nothing was truly said makes us label it as a fake news. But the algorithm matched the high retweet count and the favorites to think the tweet is not fake. The user seems rather popular so the algorithm might have associated his popularity with trust.

The tweet with details Tweet_Text: "*RT @BetoORourke: BREAKING: The Dallas Morning News calls Beto the no-brainer choice in this race! Your donation to Beto's campaign https://t.co/tGhtIvpHZg*", Date: "*2018-02-05 20:38:24*", User_Screen_Name: "*detrice0704*", Retweets: "*1111*", Favorites: "*0*" portrays a message of endorsement regarding an upcoming election. The message is marked as not fake, although it should be considered fake. The author of the tweet asserts that he was promoted as the best choice in the upcoming race by a media outlet. The message does not point to a source and the author is talking about himself, promoting himself with information unverifiable within the message. The message also contains keywords such as "*BREAKING*", written with capitalized words, suggesting an extraordinary event or situation. This is meant to attract attention to the tweet in order for it to gather more reads and, possibly, retweets. That being said, the original author of the tweet has a rather large following, count 97 thousand followers, with the post under inquiry having been shared roughly 1.1 thousand times, hence pointing that the message has garnered quite the attention from users. This can only mean that the algorithm weighed the traction of the tweet more than the contents of it.

4.3. Error analysis

After gathering the results from the aforementioned examples, close examinations were considered to ascertain the causes of incorrect labelling.

The most conspicuous reason for inaccurate classifications is the user's feedbacks, which include retweets and the number of people that favorited the article concerned. Since the greatest trust is in the social application's users, the detection system is heavily influenced on this side, even though the tweet describing the article body of a trustworthy and genuine topic is an actual reliable subject.

Other examples, which were also mentioned in this paper include tweets posted in a different language than English that were marked as fake, since the Machine Learning model built behind the whole system only contains training data consisting of tweets posted in English.

5. Conclusions

With the comparison of the results among all the classifiers like Support Vector Machine, Random Forest, Naïve Bayes, we have found that Random Forest has the maximum accuracy followed by SVM. Our research work is an attempt to solve a problem commonly encountered in society, namely the ambiguity of understanding information and manipulating citizens.

We plan to expand the work by adding neural network algorithms such as Long Short-Term Memory and other algorithms. The goal is to make continuous improvements to this topic with better precision. Also, in the future, we plan to refine and engaging the algorithms we already use through a more rigorous classification. We plan to create a

user-friendly interface that allows easy and intuitive access to fake news: by entering the news title the user will find out whether this is credible or not.

References

Bajaj, S. (2017). *"The Pope Has a New Baby!" Fake News Detection Using Deep Learning*. Stanford University, CS 224N - Winter 2017.

Conroy, N. J., Rubin, V. L., Chen, Y. (2015). Automatic Deception Detection: Methods for Finding Fake News. In *ASIST 2015*, November 6-10, St. Louis, MO, USA.

Hadeer, A., Issa, T., Sherif, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In *Proceedings of International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments ISDDC 2017*, 127-138.

Gupta, A., Kumaraguru, P. (2012). Credibility Ranking of Tweets during High Impact Events. In *PSOSM '12 Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, 2, 7 pages.

Idehen, K. U. (2017). *Exploitation of a Semantic Web of Linked Data, for Publishers*. Open Link Virtuoso Universal Server, https://medium.com/virtuoso-blog/ exploitation-of-a-semantic-web-of-linked-data-for-publishers-295f16ee8525

Perez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R. (2017). *Automatic Detection of Fake News*. Cornell University Library, https://arxiv.org/abs/1708.07104.

Popoola, O. (2017). Using Rhetorical Structure Theory for Detection of Fake Online Reviews. In *Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms*, 58−63, Santiago de Compostela, Spain, September 4 2017. ACL.

Rubin, V. L., Lukoianova, T. (2014). Truth and deception at the rhetorical structure level. In *Journal of the American Society for Information Science and Technology*, 66 (5), DOI: 10.1002/ asi.23216.

Rubin, V. L., Conroy, N. J., Chen, Y., Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Workshop on Computational Approaches to Deception Detection at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-CADD2016), San Diego, California, June 17, 7-17.

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. In *ACM SIGKDD Explorations Newsletter*. 19 (1): 22-36, DOI: https://doi.org/10.1145/3137597.3137600.

Singh, V., Dasgupta, R., Sonagra, D., Raman, K., Ghosh, I. (2017). Automated Fake News Detection Using Linguistic Analysis and Machine Learning. In *Conference SBP-BRiMS*, 1-3, DOI: 10.13140/RG.2.2.16825.67687.

Sneha, S., Nigel, F., Shrisha, R. (2017). 3HAN: A Deep Neural Network for Fake News Detection. In *24th International Conference on Neural Information Processing (ICONIP 2017)*, Springer International Publishing AG 2017, Part II, LNCS 10635, 1–10, https://doi.org/10.1007/ 978-3-319-70096-0_59.

Wu, L., Liu, H. (2018). Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *WSDM 2018, Proceedings of the Eleventh*

ACM International Conference on Web Search and Data Mining, 637-645, DOI: 10.1145/3159652.3159677.