SUP'COM

Higher School of Communication of Tunis

Fake news detection

Authors:

BOUZAIDI Chaima

RYAHI Ihebeddine

Group:

INDP3 - AIM

Année universitaire : 2019 - 2020

Table of contents

List of figures

Introduction

Social media systems have been dramatically changing the way news is produced, disseminated, and consumed, opening unforeseen opportunities, but also creating complex challenges. A key problem today is that social media has become a place for campaigns of misinformation that affect the credibility of the entire news ecosystem. A unique characteristic of news on social media is that anyone can register as a news publisher without any upfront cost (e.g., anyone can create a Facebook page claiming to be a newspaper or news media organization). Consequently, not only traditional news, corporations are increasingly migrating to social media. Along with this transition, not surprisingly, there are growing concerns about fake news publishers posting fake news stories, and often disseminating them widely using fake followers.

Thus, due to extensive spread of fake news on social and news media it became an emerging research topic nowadays that gained attention. In the news media and social media, the information is spread high-speed but without accuracy and hence detection mechanism should be able to predict news fast enough to tackle the dissemination of fake news. It has the potential for negative impacts on individuals and society.

Also, as the extensive spread of fake news can have a serious negative impact on individuals and society, the lack of scalable fact checking strategies is especially worrisome. Not surprisingly, recent research efforts are devoted not only to better comprehend this phenomenon but also to automatize the detection of fake news.

Therefore, detecting fake news on social media is important and also a technically challenging problem these days. This is the framework of the project described in this report. Henceforth, its aim is the development of an end-to-end solution for classifying fake news. The first chapter describes the general context of our proposed solution. The second one illustrates the organigram of our project whereas the fourth chapter presents tools and achieved results.

Chapter 1: General Context of the project

Introduction

This chapter is devoted to the presentation of the general context of the project. Since the proposed solution revolves around a previously identified problematic, we will begin by developing this problematic. Then, we will situate the project in its own frame and this by fixing its different parts. Finally, we will present the adopted methodology.

1. Problematic

Fake news and the spread of misinformation have dominated the news cycle after US Presidential elections in 2016. Some reports show that Russia has created millions of fake accounts and social bots to spread false stories during the elections. Various motivations are observed for spreading fake news and generating these types of information on social media channels. Some of them are to gain political gains or ruin someone else's reputation or for seeking attention. Fake news is a type of yellow journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media. The importance of fake news can also be judged through below diagram, in figure 1, shows dramatically fake news gained worldwide popularity in 2016 after US presidential elections.

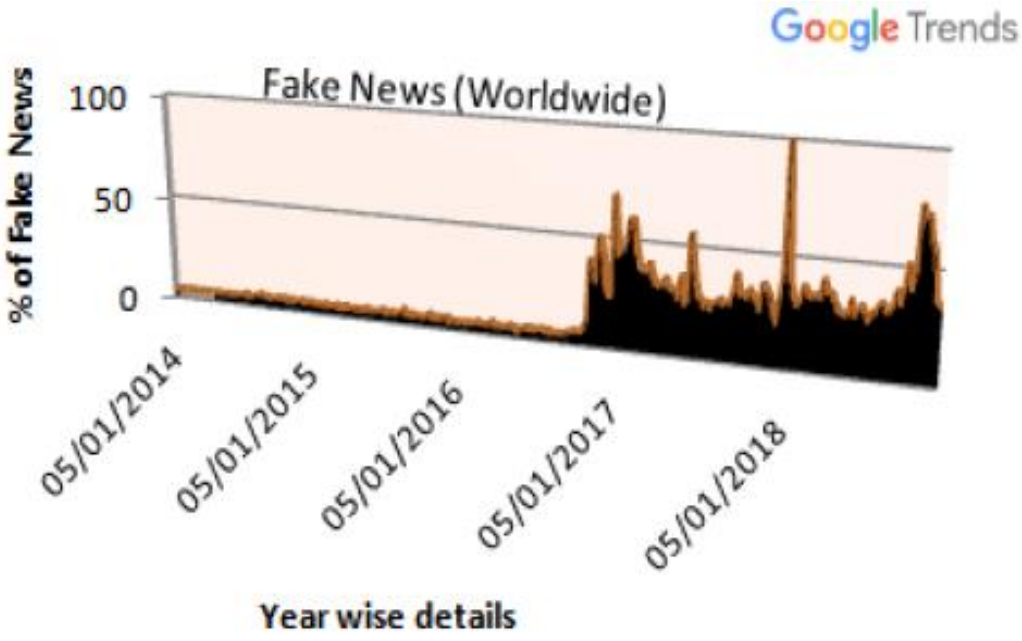


Figure 1: Fake news statistics

The wide-ranging spread of fake news can have a negative impact on society and individuals. Fake news intentionally persuades clients to accept biased or false beliefs. Fake news changes the way people interpret and respond to real news. For example, some fake news was just created to trigger people's distrust and make them confused, impeding their abilities to differentiate what is true and what is not true. There are various approaches for automated fake news detection: Text Classification, Stance Detection, Metadata and Fact Checking.

2. Proposed Solution

Our problem is to define whether or not a certain news article is fake news. Thus, our solution is to build an end-to-end solution for fake news detection using text classification approach.

It consists of three major components:

- **Text preprocessing:** Clean and transform raw text to make it easily understandable.
- **Feature engineering:** Turn each text into a vector to send through a model.
- **Model training:** Fit our dataset into a machine learning model.
- **Web Application:** Provide a user interface for detecting fake news.

3. Methodology

In order to organize and supervise the progress of the different stages, the design and realization of the project were divided into 4 phases:

- **Launch:** Understand the different aspects of the project and develop the problematic.
- **Definition:** Define and delimit the objectives of the proposed solution.
- **Realization:** Analyze feature and define the techniques to use.
- **Validation:** Test and validate the solution as well as identify any problems to correct them.

Conclusion

Throughout this chapter, the objectives of the envisaged solution have been discussed and delimited. At this stage and before introducing the technical specificities of our project, a conceptual study of our project is essential, which will be the subject of the next chapter.

Chapter 2: Design of the project

Introduction

In order to achieve a functional system, it is wise not to get started right away in applying some algorithms and take our time to properly model the different system functionalities. It is in this context that we begin this chapter to present the different functional and non-functional needs as well as the organigram of our fake news classifier.

1. Needs Analysis

1.1. System actors

An actor is an entity, hardware or software, which interacts with the system to perform one or more functions. The only actor interacting with our system is the user that will submit a query image in order to search suitable similar images.

1.2. Functional needs

The functional needs are the different functionalities to be implemented for our fake news classifier.

Indeed, our system must mainly meet the 2 functional needs:

- The user submits a request to predict whether or not an article is fake news.
- The server handles the submitted article, transforms it into an understandable format, predicts if it is a fake or real news using the trained model and returns the prediction output (fake/real) in the UI web application.

1.3. Non-functional needs

Non-functional needs summarize performance options to offer in order to optimize the functioning of our system and ensure better use and better management.

Besides the functional needs, our solution must meet certain requirements:

- **Performance constraint:** Our fake news classifier must have response times nearly equal to the real-time.
- **Ergonomic constraint:** Our fake news classifier must have a simple and user-friendly interface so that the user has easy handling.
- **Scaling constraint:** Our fake news classifier must be scalable.

2. Machine learning solution organigram

Our fake news classifier can be boiled down into 3 distinct steps detailed in the section below.

2.1. Text preprocessing

Our dataset, downloaded from Kaggle, is comprised of 3997 news articles each include a title, text and the target label as REAL/FAKE binary label. As shown in the figure 2, we can check if our dataset target is balanced or not:

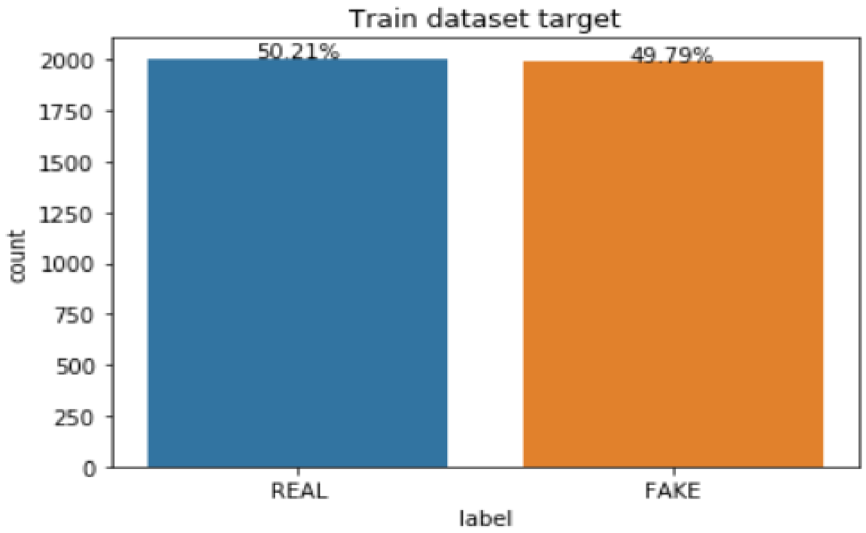


Figure 2: Fake news dataset

We can clearly see that the target is balanced. So, we jump ahead into the preprocessing by doing the following:

- Lowercase the text: This preprocessing is done so words can later be cross checked with the Stop words and pos tagging dictionaries.
- Remove words with just one letter.
- Remove words that contains numbers.
- Tokenize the text and remove punctuation.
- Remove Stop words: Proper analysis of text usually relies on the most recurring words. Stop words including as “the”, “as” and “and” appear a lot in a text, but each word doesn’t really give relevant explanation, so one of we have to remove such words.
- Remove empty tokens: After tokenization, we have to make sure all tokens taken into account contribute to the label prediction.

- Lemmatize the text: In order to normalize the text, we apply lemmatization. In this way, words with the same root are processed equally e.g. when the words “took” or “taken” are found in the text, they are lemmatized to “take”, infinitive of the verb.

2.2. Feature engineering

In this part, we had to vectorize our dataset to make it understandable by the machine learning model. To do, we proceed as follows:

- **Pos-tag Text:** Adding a prefix to each word with its type (Noun, Verb, Adjective, …). e.g: I went to school become PRP-I VBD-went TO-to NN-school
  Also, after lemmatization it will be VB-go NN-school, which indicates the semantics and distinguishes the purpose of the sentence. This will help the classifier differentiate between different types of sentences.
- **TF-IDF weighting:** Calculating the importance of each word based on its recurrence in text.
- **Vectorizing our dataset**

2.3. Training a classifier

We tried to fit our dataset by doing a benchmark between several classifiers and measure accuracy for each classifier to choose the best one. Then, we did a Grid search for different combinations of the past steps along with different models to optimize for the best hyper parameters.

Conclusion

Throughout this chapter, we have carried out the design of our project as well as the overall organigram of the solution. The following chapter will be devoted to the description, on the one hand, the used techniques and tools, and on the other hand, the stages of implementation of the proposed solution and the achieved results.

Chapter 3: Implementation of the project

Introduction

Flexible and easy handling always remains among the most decisive criteria for the success of any project, it is for this reason the choice of techniques, adopted algorithms as well as development tools must be well studied. In this chapter, we will be defining the adopted algorithms and techniques as well as the achieved results.

1. Technical specifications

1.1. Hardware tools

To fit adequately our dataset, some hardware requirements are mandatory to analyze it and apply transformations on it. Thus, to do, we will be using Google Colab with GPU as execution environment type and a RAM of 25GB.

1.2. Software tools

Our system is implementation using Python programming language. The web application is developed using mainly Flask framework on the backend and Reat.js on the frontend.

2. Techniques and algorithms

2.1. NLTK (Natural Language Toolkit)

- **Stop words:** The process of converting data to something a computer can understand is referred to as preprocessing. One of the major forms of preprocessing is to filter out useless data. In natural language processing, useless words (data), are referred to as Stop words. We would not want this words taking up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that we consider to be Stop words. NTLK in Python has a list of Stop words stored in 16 different languages.
- **Wordnet:** WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet’s structure makes it a useful tool for computational linguistics and natural language processing. WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close

proximity to one another in the network are semantically disambiguated. Also, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

- **Part of Speech Tagging (POS Tagging):** Parts of speech Tagging is responsible for reading the text in a language and assigning some specific token (Parts of Speech) to each word. One of the more powerful aspects of the NLTK module is the Part of Speech tagging. Steps involves tokenizing text and applying to it pos tagging.

2.2. TF-IDF weighting

Machine learning algorithms can not work with raw text directly. Rather, the text must be converted into vectors of numbers. In natural language processing, a common technique for extracting features from text is to place all of the words that occur in the text in a bucket. This aproach is called a bag of words model or BoW for short. It's referred to as a *"bag"* of words because any information about the structure of the sentence is lost.

TF-IDF (Term Frequency(TF) - Inverse Dense Frequency(IDF)) is a technique which is used to find meaning of sentences consisting of words and cancels out the incapabilities of Bag of Words technique which is good for text classification or for helping a machine read words in numbers and understand the meaning of the sentence or the document.

Thus, TF-IDF is a weighting factor which is used to get the important features from the documents. It actually tells us how important a word is to a document in a corpus, the importance of a word increases proportionally to the number of times the word appears in the individual document, this is called Term Frequency(TF). It's also a frequent word in all other documents in our corpus so it does not give much meaning so it probably may not be an important feature. To adjust this, we use IDF.

The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents.

2.3. Trigram vectorizer

After pos-tagging, if we take one word as a feature, that's called Unigram. if we take two words at a time as a feature, that's called Bigram. three words at a time as a feature, that's called Trigram

So, Trigram is 3 consecutive words in a sentence.

2.4. Logistic Regression

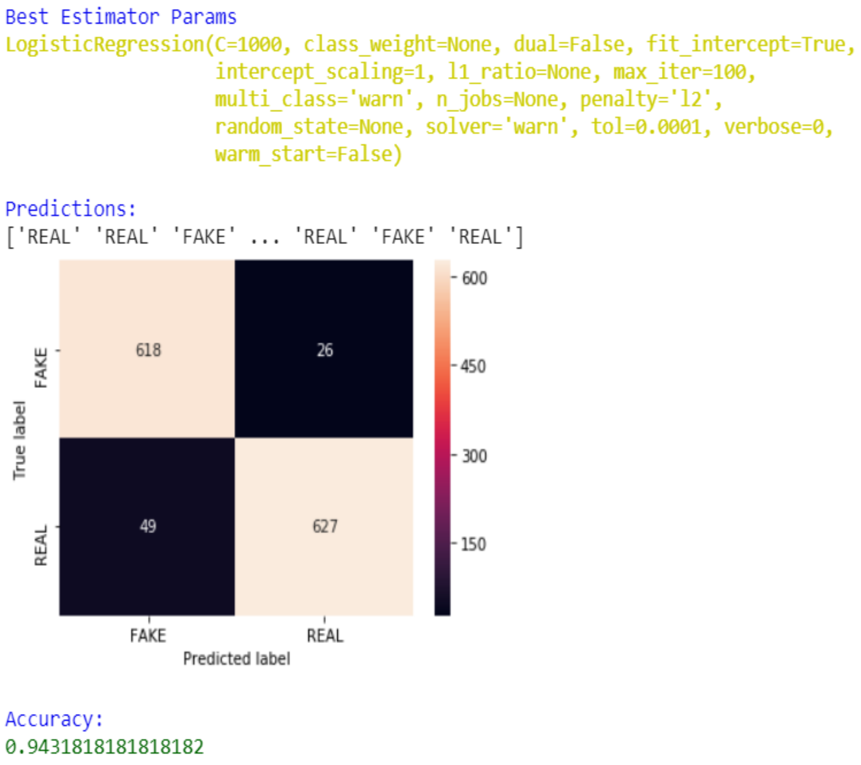
Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

2.5. Web application

The functionalities of our fake news classifier are available to the user via a Single Page Application (SPA) that consumes data via React.js (on the frontend) from an internal, REST-like API via Python/Flask (on the backend).

3. Results

The performances of our classifier are shown in the figure 5.



```
Best Estimator Params
LogisticRegression(C=1000, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)

Predictions:
['REAL' 'REAL' 'FAKE' ... 'REAL' 'FAKE' 'REAL']
```

```
Accuracy:
0.9431818181818182
```

Figure 3: Logistic Regression performances

The user interface of our project is shown in the figure 4:

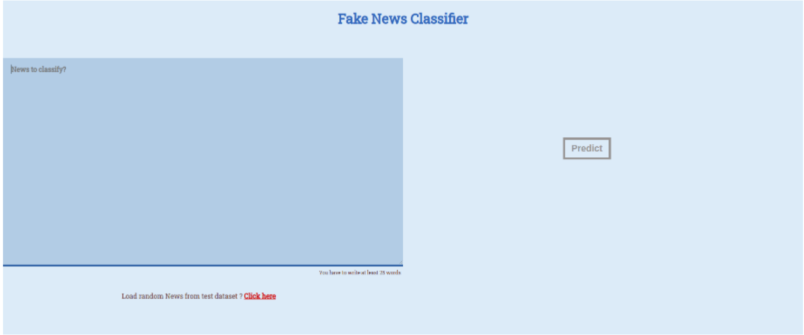
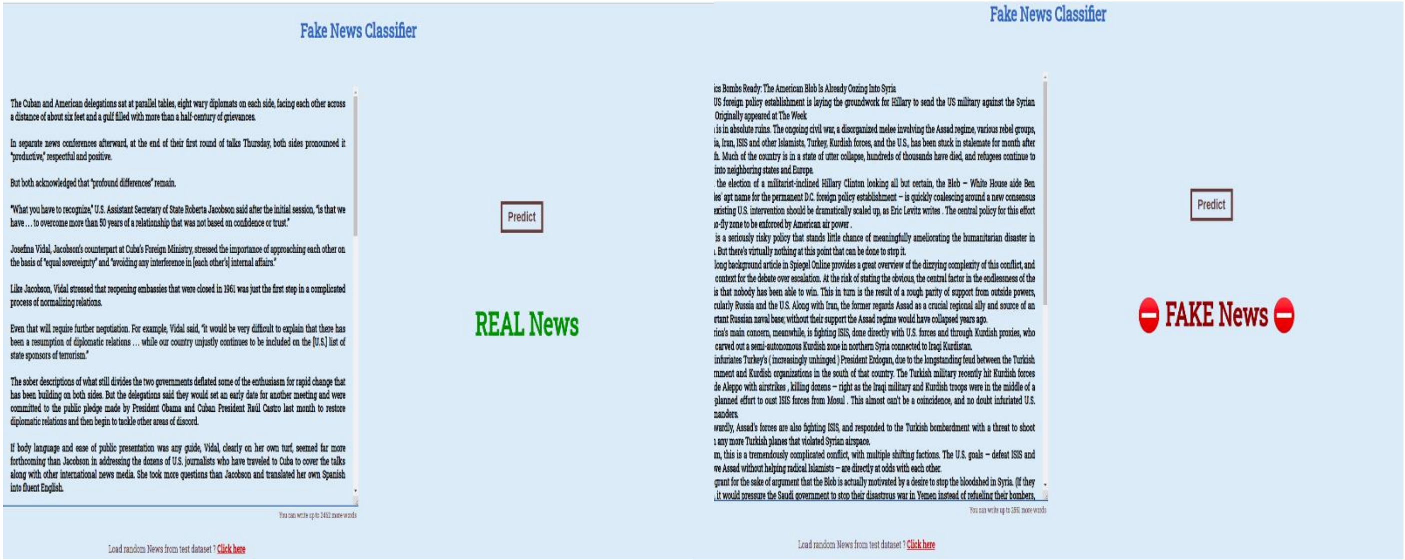

Fake News Classifier
News to classify?
Predict
You have to write at least 20 words
Load random News from test dataset ? Click here

Figure 4: User Interface

When the use pastes an article, and clicks on predict, the backend server handle that query and do the following steps in order to return prediction (Real/Fake):

- Preprocess the article
- Vectorize the article
- Predict if it is REAL or FAKE, as the illustrated example in figure 5.


Fake News Classifier
Predict
REAL News
You can write up to 2452 more words
Load random News from test dataset ? Click here
Fake News Classifier
Predict
FAKE News
You can write up to 2551 more words
Load random News from test dataset ? Click here

The Cuban and American delegations sat at parallel tables, eight wary diplomats on each side, facing each other across a distance of about six feet and a gulf filled with more than a half-century of grievances.

In separate news conferences afterward, at the end of their first round of talks Thursday, both sides pronounced it "productive," respectful and positive.

But both acknowledged that "profound differences" remain.

"What you have to recognize," U.S. Assistant Secretary of State Roberta Jacobson said after the initial session, "is that we have … to overcome more than 50 years of a relationship that was not based on confidence or trust."

Josefina Vidal, Jacobson's counterpart at Cuba's Foreign Ministry, stressed the importance of approaching each other on the basis of "equal sovereignty" and "avoiding any interference in [each other's] internal affairs."

Like Jacobson, Vidal stressed that reopening embassies that were closed in 1961 was just the first step in a complicated process of normalizing relations.

Even that will require further negotiation. For example, Vidal said, "it would be very difficult to explain that there has been a resumption of diplomatic relations … while our country unjustly continues to be included on the [U.S.] list of state sponsors of terrorism."

The sober descriptions of what still divides the two governments deflated some of the enthusiasm for rapid change that has been building on both sides. But the delegations said they would set an early date for another meeting and were committed to the public pledge made by President Obama and Cuban President Raúl Castro last month to restore diplomatic relations and then begin to tackle other areas of discord.

If body language and ease of public presentation was any guide, Vidal, clearly on her own turf, seemed far more forthcoming than Jacobson in addressing the dozens of U.S. journalists who have traveled to Cuba to cover the talks along with other international news media. She took more questions than Jacobson and translated her own Spanish into fluent English.

ics Bombs Ready: The American Blob Is Already Oozing Into Syria
US foreign policy establishment is laying the groundwork for Hillary to send the US military against the Syrian
Originally appeared at The Week
is in absolute ruins. The ongoing civil war, a disorganized melee involving the Assad regime, various rebel groups,
ia, Iran, ISIS and other Islamists, Turkey, Kurdish forces, and the U.S., has been stuck in stalemate for month after
th. Much of the country is in a state of utter collapse, hundreds of thousands have died, and refugees continue to
into neighboring states and Europe.
the election of a militarist-inclined Hillary Clinton looking all but certain, the Blob – White House aide Ben
les' apt name for the permanent D.C. foreign policy establishment – is quickly coalescing around a new consensus
existing U.S. intervention should be dramatically scaled up, as Eric Levitz writes . The central policy for this effort
no-fly zone to be enforced by American air power .
is a seriously risky policy that stands little chance of meaningfully ameliorating the humanitarian disaster in
. But there's virtually nothing at this point that can be done to stop it.
long background article in Spiegel Online provides a great overview of the dizzying complexity of this conflict, and
context for the debate over escalation. At the risk of stating the obvious, the central factor in the endlessness of the
is that nobody has been able to win. This in turn is the result of a rough parity of support from outside powers,
cularly Russia and the U.S. Along with Iran, the former regards Assad as a crucial regional ally and source of an
rtant Russian naval base; without their support the Assad regime would have collapsed years ago.
rica's main concern, meanwhile, is fighting ISIS, done directly with U.S. forces and through Kurdish proxies, who
carved out a semi-autonomous Kurdish zone in northern Syria connected to Iraqi Kurdistan.
infuriates Turkey's (increasingly unhinged) President Erdogan, due to the longstanding feud between the Turkish
rnment and Kurdish organizations in the south of that country. The Turkish military recently hit Kurdish forces
de Aleppo with airstrikes , killing dozens – right as the Iraqi military and Kurdish troops were in the middle of a
planned effort to oust ISIS forces from Mosul . This almost can't be a coincidence, and no doubt infuriated U.S.
nanders.
wardly, Assad's forces are also fighting ISIS, and responded to the Turkish bombardment with a threat to shoot
any more Turkish planes that violated Syrian airspace.
m, this is a tremendously complicated conflict, with multiple shifting factions. The U.S. goals – defeat ISIS and
ve Assad without helping radical Islamists – are directly at odds with each other.
grant for the sake of argument that the Blob is actually motivated by a desire to stop the bloodshed in Syria. (If they
, it would pressure the Saudi government to stop their disastrous war in Yemen instead of refueling their bombers,

Figure 5: Fake news classifier

Conclusion

This chapter described the adopted techniques and algorithms. Then, we focused on the implementation as well as the achieved results of our fake news classifier