Fake news detection using topic and author agnostic approach supplemented by Twitter data

Manmeet Kumar Chaudhuri
SRH Hochschule Heidelberg
manmeetkumar.chaudhuri@stud.hochschule-heidelberg.de

Rishabh Garg
SRH Hochschule Heidelberg
rishabh.garg@stud.hochschule-heidelberg.de

Rohit Bewoor
SRH Hochschule Heidelberg
rohitkeshav.bewoor@stud.hochschule-heidelberg.de

Sanika Medankar
SRH Hochschule Heidelberg
Sanika.Medankar@stud.hochschule-heidelberg.de

Shekhar Singh
SRH Hochschule Heidelberg
Shekhar.Singh@stud.hocschule-heidelberg.de

Siddarth Venkateswaran
SRH Hochschule Heidelberg
siddarth.venkateswaran@stud.hochschule-heidelberg.de

ABSTRACT

In current scenario, online social media plays an important role during real-world events. It can be used by authorities for effective disaster management or by malicious entities to spread rumours and fake news. Fake news detection, especially on websites and on social media is an active area of research. Several entities like social media giants, journalists and news outlets need to continuously do fact checking. In this paper, we present our system built with the aim of identifying if a website has published fake news - where the website URL has been tweeted by Twitter users. The approach does not use the topic of the article or the Twitter user characteristics; therefore this is a topic **[1]** and author agnostic approach.

CCS CONCEPTS

Computing methodologies, Supervised learning by classification; Feature selection, Twitter scraping, Text mining

1. INTRODUCTION

Penetration of internet has been increasing steadily and many people now consume news via online websites and social media more than ever before. As per the report published by the Pew Research Centre on 2nd October 2019, 55% of US adults get their news from social media either "often" or "sometimes", and that's a 9% jump over last year. Social media for news consumption is like a double-edged sword **[2]**. While it provides a low cost, easy to access, medium allowing rapid dissemination of information, on the other hand, it enables fake news or false information to be communicated widely and rapidly. Deliberately misleading articles, websites and social media posts can come about for lots of different reasons: they might be trying to influence elections or policies; they might represent a form of cyberwarfare between states; they might be aimed at raising someone's profile and influence or discrediting their opponents. Or they might simply be about making money, relying on the attention-grabbing nature of outrageous lies to generate ad revenue; case in point the "digital gold rush" that saw a small Macedonian town register more than 150 pro-Trump websites during the 2016 presidential race **[3].**

Previous approaches to fake news identification have largely focused on using the content of news websites to determine the veracity of the news. However, using the content has important limitations. Notably, due to the dynamic nature of news, there is a continuous change in topics and discourse. Therefore, a classifier trained using content from articles published at a given time is likely to become ineffective in the future.

While there is no full-proof means of telling fact from fake yet, there are certain features of the content of the website that one may notice that should put us on guard. Is the writing style more informal than one would expect? Does it contain a lot of superlatives and emphatic language? Does it make subjective judgments or read more like a narrative than reportage? Ultimately, we may have to rely on artificial intelligence to do the heavy lifting for us – and it should be able to tell us whether the identified linguistic patterns are seen in large datasets of fake news, invisible to the "naked eye", are present.

In this paper, we study the problem of detecting fake news published on websites. Using a set of URLs news content, and that the URL has been tweeted on Twitter at least once, our goal is to determine whether the website is likely to contain fake news or not. We propose a new approach that is topic agnostic and author profile agnostic. We shall not be analyzing the topic using approaches like bag of words. Instead, we use topic agnostic features and tweet specific features in our classification models.

2. LITERATURE REVIEW

Social media provides a new paradigm of information creation and consumption for users. The information seeking and consumption process are changing from a mediated form (e.g., by journalists) to a more disinter-mediated way **[7]**. Having a reliable way of identifying fake news is burning issue presently. The fundamental problem is that it mimics reliable reporting – and people can't always tell the difference. That is why, for the past few years, researchers

have been trying to work out what the linguistic characteristics of fake news are. Computers that are fed material already classified as misleading can identify patterns in the language used. They are then able to apply that knowledge to the new material presented to them and flag it as potentially dubious based on these previously identified patterns

One such project, led by Fatemeh Torabi Asr, at Simon Fraser University in Canada, recently found that "on average, fake news articles use more words related to sex, death and anxiety" **[4]**. "Overly emotional" language is often deployed. In contrast, "Genuine news contains a larger proportion of words related to work (business) and money (economy)."

Another group of researchers analyzed the relationship of various grammatical categories to fake news. They concluded that words which can be used to exaggerate are all found more often in deliberately misleading sources. These included superlatives, like "most" and "worst", and so-called subjective, like "brilliant" and "terrible". They noted that propaganda tends to use abstract generalities like "truth" and "freedom", and intriguingly showed that use of the second-person pronoun "you" was closely linked to fake news **[5]**.

Some of these approaches have their problems. Jack Grieve, at the University of Birmingham, cautions that scholars don't always control for the genre – so the differences in language as seen above might just come down to the difference between a more formal news article, and a more casual Facebook post **[2]**. To get around this problem, Grieve's team compared 40 retracted and 41 non-retracted articles by Jayson Blair, who resigned from the New York Times in disgrace in 2003. Though these articles were produced in a single genre - national newspaper writing - they still displayed subtle differences in register related to the different communicative purposes they served (on one hand to inform, on the other to deceive). Even though he was trying to pass his work off as factual, there were subtle hints that only become evident when the data is crunched. For example, there were more emphatic words, like "really" and "most", in Blair's retracted articles. He used shorter words and his language was less "informationally dense". The present tense cropped up more often and he relied on the third person pronouns "he" and "she" rather than full names – something that's typical of fiction.

Sonia Castelo and team in May 2019 **[1]** presented a new approach to detect fake news websites which uses topic-agnostic features. Through a detailed experimental evaluation, they showed that their approach accurately classifies not only political news as topics evolve over time, but also news from different domains, outperforming content-based approaches while using significantly fewer features and requiring no frequent re-training. The results suggest that topic-agnostic features are effective for distinguishing between fake and real news, and that robust classifiers can be constructed that enable the timely discovery of fake news articles. They used a corpus of over 14,000 websites (unique URLs) drawn from 137 sites and spanning 6 years. This was a first of its kind in terms of size, focus on the Web, and inclusion of HTML mark-up information.

3. METHODOLOGY

Building off the approach used by Ms. Castelo et al., our approach can be broken down into the following steps:

Stage 1: Data Collection

- The URLs and their content are taken from a GitHub[1] repository that already had a corpus of URLs tagged under various categories like Fake, Conspiracy, Neutral, Reliable, etc. We selected URLs tagged as "Fake" and "Reliable". The corpus had around 2 million URLs tagged as Reliable and 0.9 million as Fake.
- Next, Twitter data was extracted for the URLs tagged as Fake and Reliable. We selected 5500 URLs from each category making sure that the content for the URL had only English language text. Thus our corpus had 11,000 URLs with an equal split for Fake and Reliable types.

Figure 1: Web pages from reliable and fake news sites.

Stage 2: Feature Extraction

The GitHub repository files already had the headline and content extracted from the URLs. This data was cleaned during pre-processing and subsequently various features were extracted.

Figure 1 shows some examples of fake and reliable news pages. Fake news pages not only have a larger number of ads and polluted layouts but also have a distinctive style to their headlines, often in the form of a sensationalist slant. Additionally, besides attempting to describe the article, these headlines often contain terms designed to catch the readers' attention. These observations motivated us to investigate linguistic-based features (morphological, psychological and readability-related). The features are listed in ***Table 2*** and we summarize them below.

[1] github.com/several27/FakeNewsCorpus

For each of the selected URLs, we identified certain topic agnostic and user agnostic features - these were thus independent of the category/ type of the news and the author. There are four groups of features: "Morphological", "Psychological", "Readability" and "Propagation and Social Sharing". The first three are extracted using linguistic analysis method as per the approach cited in the paper earlier. For the last one, we extracted tweets. For brevity, we will refer to the fourth category as "Twitter Features" hereafter.

a) Morphological features:

The branch of linguistics concerned with formation of words and phrases along with the interaction of words among themselves is termed as Morphology. Morphological analysis gives us detailed information of the text which can be used to generalize the text, e.g., by replacing words by their parts of speech, which allows us to identify constructs such as (preposition, noun) instead of combinations of specific words. We obtain these patterns through part-of-speech tagging.

We used the spaCy[2] library part-of-speech tagger to compute Morphological Features.

b) Psychological features:

Psychological features capture the percentage of total semantic words in texts. By drawing on massive amounts of text, we can begin to link everyday language use with behavioural and self-reported measures of personality, social behaviour, and cognitive styles. The basic idea is to explore the links between word usage and basic social and personality characteristics.

For Psychological Features, Linguistic Inquiry and Word Count software **[8]** (LIWC[3], Version 1.3.1 2015) was used.

c) Readability features:

With the aim of conveying important information to most readers, an evaluation of readability helps text writers to adjust their content to their target audience's level. We obtain these features through readability scores and counting of character, words, and sentences usage.

To compute Readability features, we used the TextSTAT[4] and NLTK[5] libraries.

d) Tweet features:

These are derived from Twitter[6] data for each of the URLs. With the resources available at our disposal, we only extracted tweets which contained the URLs of interest in the text entered by a Twitter user. Any retweets, replies or retweet-with-comments, on the aforementioned set of tweets, were not extracted.

A fake news cascade can be represented in terms of the number of steps (i.e., hops) fake news has travelled (i.e., hop-based fake news cascade) or the times it was posted (i.e. time-based fake news cascade) **[6]**. Thus, from a network propagation effect, while we are focussing only on the first hop, the necessity of collecting huge amounts of data for the subsequent hops is obviated.

Stage 3: Creation of URL level features data and application of Classification Models

- Different features were extracted belonging to each of the four groups mentioned above. These were then combined at a URL level to produce a final dataset.
- Using this dataset, different classification techniques (KNN, Logistic Regression, Decision Trees and others) were used to generate a set of models.

A workflow for the steps just described is in ***Figure 2***.

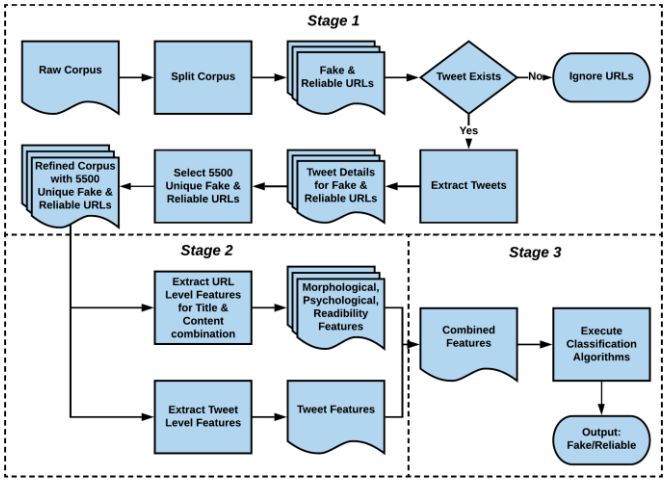

Figure 2: Workflow of the implemented process.

Python[7] Programming Language has been used to implement above mentioned methodology and relevant codes can be found in this GitHub[8] link.

4. EVALUATION & RESULT

We evaluated various models trained with different combinations of feature groups (separately and jointly).

Feature Selection for the models:

For each of the URLs, we started out with a total of 183 features across the four groups. Psychological group had the maximum number (93) while Readability had the lowest (18). Feature selection was performed using Random Forest Decision Tree using Scikit-learn[9] package with the default parameters.

From the breakdown below we see that: a) Morphological and Psychological groups had an equal number of features; and b) the other two groups had much lower but similar number important features.

[2] https://spacy.io/usage/spacy-101
[3] http://liwc.wpengine.com
[4] https://pypi.org/project/textstat/
[5] https://www.nltk.org
[6] https://twitter.com
[7] https://www.python.org
[8] https://github.com/rbewoor/FakeNewsDetection.git
[9] https://scikit-learn.org/stable/

	Count of features	Count of effective features
Morphological Features	50	24
Psychological Features	93	24
Readability Features	18	7
Tweet Features	22	9
Total	**183**	**64**

Table 1: List of total vs important features.

It can be seen that the number of effective features are 64 and these will be effective in differentiating fake and reliable URLs.

A description of the important features is listed in ***Table 2*** below.

	Abbreviation	Description	Abbreviation	Description	Abbreviation	Description
Morphological	M_ADD	Email	M_HYPH	Punctuation mark hyphen	M_SYM	Symbol
	M_AFX	Affix	M_JJ	Adjective	M_UH	Interjection
	M_CD	Cardinal number	M_NN	Noun singular or mass	M_VB	Verb base form
	M_colon	Punctuation mark colon or ellipsis	M_NNP	Noun proper singular	M_VBD	Verb past tense
	M_comma	Punctuation mark comma	M_NNPS	Noun proper plural	M_VBG	Verb gerund or present participle
	M_dblApostrophe	Closing quotation mark	M_NNS	Noun plural	M_VBN	Verb past participle
	M_dblSpecialApoxstrophe	Opening quotation mark	M_PRP	Pronoun personal	M_VBP	Verb non-3rd person singular present
	M_dot	Punctuation mark sentence closer	M_RB	Adverb	M_VBZ	Verb 3rd person singular present
Psychological	L_AllPunc	All punctuations	L_Exclam	Exclamation marks	L_OtherP	Other punctuation
	L_Analytic	Analytical thinking	L_focuspresent	Present focus	L_ppron	Personal pronouns
	L_anger	Anger	L_informal	Informal language	L_pronoun	Total pronouns
	L_auxverb	Auxiliary verbs	L_insight	Insight	L_QMark	Question marks
	L_certain	Certainity	L_leisure	Leisure	L_Quote	Quotation marks
	L_cogproc	Cognitive processes	L_negemo	Negative Emotions	L_time	Time
	L_Comma	Commas	L_netspeak	Netspeak	L_Tone	Emotional tone
	L_Dash	Dashes	L_number	Numbers	L_WC	Word count
Twitter	T_avgTimeBetweenTweets	Average Time between two consecutive Tweets	T_countMonday	Number of Tweets on Monday	T_timeAbsBin1count0to6	Number of Tweets tweeted within the 0-6 hour bucket (from time of Aggregate of Likes for all the Tweets
	T_avgTimeOfNextTweet	Average Time between each Tweet against the First Tweet	T_countWeekDay	Number of Tweets from Monday-Friday	T_totalLikes	
	T_avgTweetsPerUniqUser	Average number of Tweets per Unique User	T_span	Time between Last and First tweet	T_totalRetweets	Aggregate of Retweets for all the Tweets
Readability	R_capitalized_words	Capitalized words	R_gunning_fog	Gunning fog	R_urls_counts	Urls
	R_coleman_liax	Coleman-Liau index	R_linear_write	Linear write	R_words_per_sentence	Words per sentence
	R_Flesch_reading_ease	Flesch reading ease				

Table 2: Linguistic and Twitter features used to represent news articles.

The comparison of average values of these effective features for fake and reliable URLs is provided in ***Figure 3***.

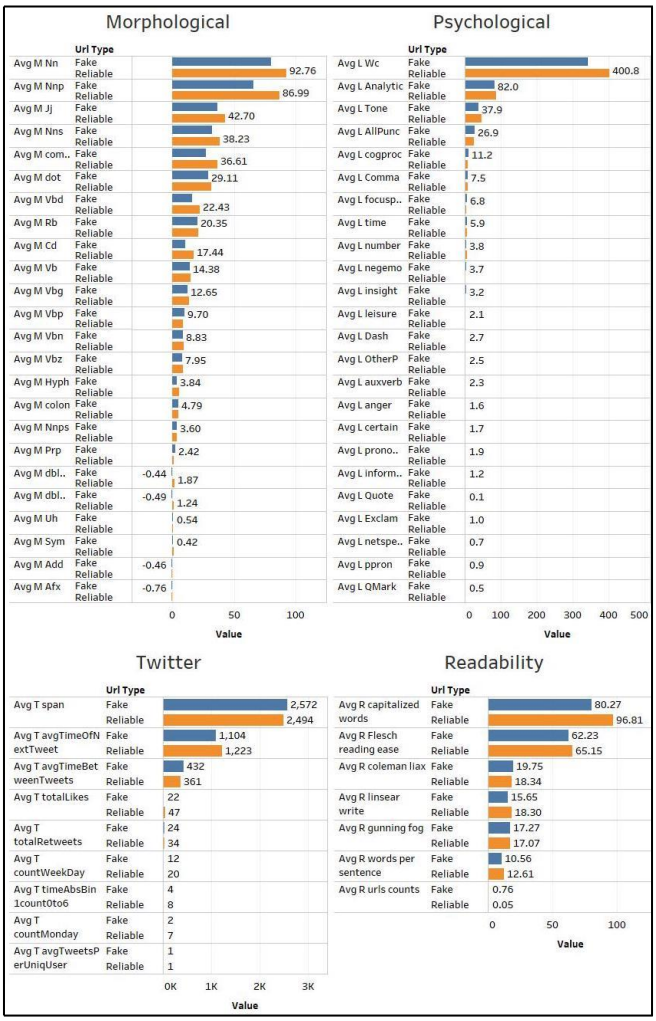

Figure 3: Comparison of average values of effective features.

One interesting trend observed is that the values for Morphological features are higher for reliable URLs as compared to fake URLs.

Discussion on Tweet features:

Each URL had one or more tweets and various features were extracted at an aggregate level for every unique URL. The following information from the tweet was used: timestamp of tweet, counts of likes, retweets and replies and the username of the Twitter User. Note: we only captured Tweets containing the URL at the first hop level.

These features can be broadly classified into four subcategories:

a. Based on the user:

- Average number of tweets per unique user. This is an important feature but the average value after rounding is 1 for both URL Types. So even a small difference still plays a role.

b. Based on the timestamp of tweet:

- Span is the time difference between the earliest and the latest tweet found. While the average value is around 2,500 hours for both categories, it is slightly higher for Fake.
- Average Time between Tweets. It is the average time between Tweets for that particular URL. For Fake URLs, the average value is 432 hours which is almost 20% higher than that of Reliable type.
- Average Time of next Tweet is the average of the time difference between the earliest tweet and any subsequent tweets. The average value for Reliable is higher by around 10%.
- Binning based on Span of the tweets. We used five bins of 0-6 hours, 6-12, 12-18, 18-24 and 24+ hours. The first bin value is an important feature.
- Counts of tweets by the day of the week. We calculated for each of the 7 days as well as for Weekdays (Monday to Friday) and Weekends (Saturday-Sunday). The numbers for Reliable are substantially higher from Sunday to Thursday. But is lower only on Friday and Saturday. The average value on Monday is 7 for Reliable vs 2 for Fake. The Monday count is an important feature. ***(Refer Figure 4.)***

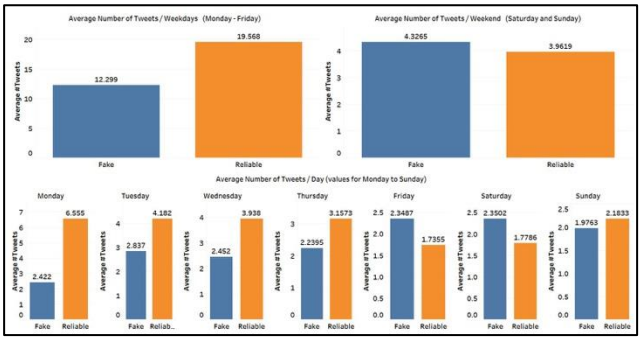

Figure 4: Daily and Weekly Twitter trends.

c. **Count of total tweets for the URL. (*Refer Figure 4*.)**

d. **Based on the actions taken by other Twitter users after seeing the tweets captured:**

- Aggregates of the number of Likes, Retweets and Replies. From ***Figure 4*** we see the average number of tweets is 50% higher for Reliable, average number of Likes is more than double for Reliable, average number of Retweets is around 40% higher for Reliable and the average number of Replies is around the same for both types of URLs.

As seen from ***Table 1*** out of the 22 total features, the following 9 were found to be important using Random Forest classifier: T_totalRetweets, T_avgTweetsPerUniqUser, T_timeAbsBin1count0to6, T_avgTimeBetweenTweets, T_avgTimeOfNextTweet, T_span, T_countMonday, T_totalLikes, and T_countWeekDay.

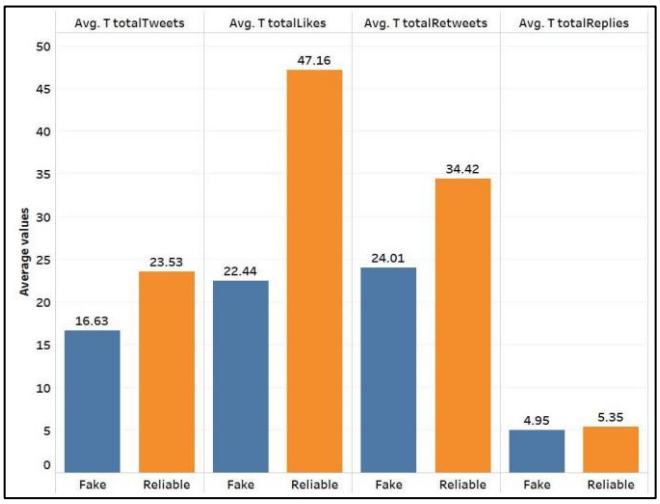

Figure 5: Average count of tweet details per URL

Effectiveness of different classification models:
The following classification models were created.

a. Decision Trees with Entropy[10] (DT-E)
b. Decision Trees with Gini Index[11] (DT-G)
c. K-Nearest Neighbors[12] (KNN)
d. Random Forest[13] (RF)
e. Support Vector Machine[14] (SVM)
f. Logistic Regression[15] (LR)
g. XGBoost[16]

These were performed on the individual as well as on a combination of certain extracted features in the following order:

a. All features combined (ALL)
b. Psychological (L)
c. Morphological (M)
d. Readability (R)
e. Twitter (T)
f. Morphological-Psychological-Readability (MLR)

The accuracy for different models trained on different input data are as follows with 80:20 for train:test split.

In addition, using the feature selection functionality of scikit-learn[17] with the estimator as Random Forest, for all six types of input data important features were identified. ***Table 4*** below shows classifier accuracies on using subsets of data.

	DT (E)	DT (G)	KNN	RF	SVM	LR	XGBoost
ALL	0.84	0.85	0.71	0.90	0.88	0.88	0.93
L	0.82	0.82	0.70	0.88	0.84	0.83	0.90
M	0.78	0.78	0.77	0.87	0.80	0.79	0.86
R	0.70	0.71	0.65	0.77	0.72	0.73	0.76
T	0.68	0.69	0.68	0.72	0.65	0.63	0.76
MLR	0.84	0.83	0.71	0.91	0.89	0.90	0.92

Table 3: Classification results (accuracies) without feature selection.

	DT (E)	DT (G)	KNN	RF	SVM	LR	XGBoost
ALL	0.85	0.84	0.70	0.90	0.85	0.87	0.93
L	0.82	0.82	0.72	0.87	0.83	0.81	0.89
M	0.76	0.77	0.75	0.83	0.77	0.76	0.84
R	0.69	0.66	0.67	0.75	0.71	0.71	0.72
T	0.66	0.68	0.69	0.72	0.62	0.61	0.74
MLR	0.84	0.83	0.70	0.90	0.86	0.85	0.91

Table 4: Classification results (accuracies) with feature selection.

Comparing accuracies between ***Table 3*** vs. ***Table 4,*** we see no substantial difference.

As can be seen above in ***Table 3***, the model analysis per feature is as follows:

a. Decision Trees with Entropy performed best on a combination of all features with an accuracy of 84% and least on Twitter with 68%
b. Decision Trees with Gini Index performed best on a combination of all features with an accuracy of 85% and least on Twitter with 69%
c. K-Nearest Neighbors performed best on Morphological features with an accuracy of 77% and least on Readability with 65%
d. Random Forest performed best on a combination of Morphological, Psychological and Readability features with an accuracy of 91% and least on Twitter with 72%
e. Support Vector Machine performed best on a combination of Morphological, Psychological and Readability features with an accuracy of 89% and least on Twitter with 65%
f. Logistic Regression performed best on the combination of Morphological, Psychological and Readability features with an accuracy of 90% and least on Twitter with 63%
g. XGBoost performed best on a combination of all features with an accuracy of 93% and least on Readability with 76%.

[10] https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8
[11] http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/
[12] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[13] https://en.wikipedia.org/wiki/Random_forest
[14] https://en.wikipedia.org/wiki/Support-vector_machine
[15] https://en.wikipedia.org/wiki/Logistic_regression
[16] https://en.wikipedia.org/wiki/XGBoost
[17] https://scikit-learn.org/stable/

5. DISCUSSION

We proposed a new approach for detection of fake URLs that uses topic and author agnostic features including Twitter features. It is found that Twitter features from just the first hop are not effective in improving the accuracy of the classifier. Classifier that incorporates linguistic features and Twitter features is as accurate as the classifier that uses only linguistic features. We would need to further study the effect of Twitter features from more hops.

Sonia Castelo and team achieved a maximum accuracy of 0.86 in their approach that incorporated topic agnostic features including web mark-up features using linear SVM classifier. We achieved an accuracy of 0.93 with our topic agnostic features using XGBoost. It seems like web mark-up features are not as effective as other features in distinguishing the fake and reliable URLs.

6. CONCLUSION & FUTURE SCOPE

Based on our methodology and the results, to identify URLs with fake news we conclude the following:

a. In the context of analysis of the content of URLs, one can ignore identification of topics and any author related characteristics. Therefore, a topic and author agnostic approach holds water.
b. Identifying fakeness based purely on Psychological group features consistently gives very high accuracy as compared to other groups taken individually viz. Morphological, Readability and Twitter.
c. Accuracy of models using only Twitter group features used by us is the least. Maybe if more Twitter features are used, the accuracy may have improved.
d. XGBoost consistently gives the best accuracy.
e. We will now discuss about improvements that could be made in our approach to further enhance the accuracy of fake news detection.

One could incorporate web mark-up features like information related to author, number of advertisements, presence of images/videos/audios, etc. for each of the URLs. These features could be extracted by web scraping using Beautifulsoup and Newspaper library.

One could also include Twitter features derived from second hop and onwards, as currently in our approach, we are only using features extracted from the first hop. Further, one could collect features related to the Twitter User (statistics about Followers, statistics about Following, Twitter Verified account, location of User), age of the Twitter account, presence of media (like images, videos), text analysis on tweets and replies, etc.

Our approach is limited to fake URL detection for English language only. Corpus can be prepared for other languages and the model can be trained to predict fake news detection for other languages as well.

REFERENCES

1. Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., & Freire, J. (2019). A Topic-Agnostic Approach for Identifying Fake News Pages. San Francisco, CA, USA: International World Wide Web Conference Committee.
2. Shearer, E., & Grieco, E. (2019, October 2). Americans Are Wary of the Role Social Media Sites Play in Delivering the News. Retrieved from Pew Research Center: https://www.journalism.org/2019/10/02/americans-are-wary-of-the-role-social-media-sites-play-in-delivering-the-news/
3. Shariatmadari, D. (2019, September 2). Could language be the key to detecting fake news? Retrieved from The Guardian: https://www.theguardian.com/commentisfree/2019/sep/02/language-fake-news-linguistic-research
4. Asr, F. T., & Taboada, M. (2019). Big Data and quality data for fake news and misinformation detection. SAGE Journals , 1.
5. Horne, B. D., & Adali, S. (2017). This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. The 2nd International Workshop on News and Public Opinion at ICWSM.
6. Zhou, X., & Zafarani, R. (2018). Fake News: A Survey of Research, Detection Methods, and Opportunities.
7. Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group polarization on facebook. Scientific Reports, 6, 2016.
8. James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. 10.15781/T29G6Z.