Identification of Synthetic Sentence in Bengali News using Hybrid Approach

Soma Das
soma_phd_2018july@iiitkalyani.ac.in
Indian Institute of Information Technology Kalyani,
West Bengal, India

Sanjay Chatterji
sanjayc@iiitkalyani.ac.in
Indian Institute of Information Technology Kalyani,
West Bengal, India

Abstract

Often sentences of correct news are either made biased towards a particular person or a group of persons or parties or maybe distorted to add some sentiment or importance in it. Engaged readers often are not able to extract the inherent meaning of such synthetic sentences. In Bengali, the news contents of the synthetic sentences are presented in such a rich way that it usually becomes difficult to identify the synthetic part of it. We have used machine learning algorithms to classify Bengali news sentences into synthetic and legitimate and then used some rule-based postprocessing on each of these models. Finally, we have developed a voting based combination of these models to build a hybrid model for Bengali synthetic sentence identification. This is a new task and therefore we could not compare it with any existing work in the field. Identification of such types of sentences may be used to improve the performance of identifying fake news and satire news. Thus, identifying molecular level biasness in news articles.

Keywords: Synthetic Sentence, Engaged Reader, Machine Learning technique, Rule Base Approach

1 Introduction

The Bengali language is rich in terms of the usage of its words. It is also a relatively free word order language. By changing the order of the same set of words, the author can add some emphasis to some part of the sentence. It is usually observed that the Bengali sentences are frequently distorted like this way. The number of ways English sentences can be distorted is much less than the number of ways a Bengali sentence can be. But all the distorted sentences not necessarily have added biasness or emphasis.

Some readers take the inherent meaning of the sentences without getting into involved in the biased part of it. They can take out an overview of the text. But often, an engaged reader gets engaged with the writer's views. Sometimes it is not so harmful or it is preferred to be an engaged reader. For example, to get the full flavour of a literary work, the reader has to be engaged. But often it is not desirable. For example, in a piece of political news, it is not recommended to engage a reader without his concern. So, it is essential to notify the reader about synthetic sentences.

Often sentences of correct news are either made biased towards a particular person or a group of persons or parties or maybe distorted to add some sentiment or importance in it. We refer such types of sentences as synthetic sentences. Engaged readers often are not able to extract the inherent meaning of synthetic sentences. In Bengali, the news contents of the synthetic sentences are presented in such a rich way that it usually becomes difficult to identify the synthetic part of it.

In this paper, we wish to identify the synthetic sentences in Bengali news automatically. We use the linguistic features in multiple Binary Machine Learning Classifiers to decide whether it is synthetic or legitimate. Then analyzing a confusion matrix, we apply a set of rules. Finally, we combine these models using a voting-based approach. We test this hybrid technique in a Bengali news corpus covering the news in Politics, Sports, Entertainment, and Social domains. The final hybrid technique is able to provide 86% accuracy.

The rest of the paper is organized as follows. Section 2 illustrates a background study related to synthetic news detection. Section 3 and Section 4 discuss how we have prepared the experimental dataset and model building part. Section 5 shows the results of different steps. Finally, Section 6 presents concluding remarks of the task.

D M Sharma, P Bhattacharyya and R Sangal. Proc. of the 16th Intl. Conference on Natural Language Processing, pages 193–200
Hyderabad, India, December 2019.

2 Related Work

There are some works in the detection of fake news. They detect fake news in a news corpus Bovet and Makse (2019); Batchelor (2017); Shu et al. (2017); Conroy et al. (2015) that is, misleading news stories which come from non-reputable sources. These papers mainly focus on fake news from four perspectives: the false knowledge, its writing style, its propagation patterns, and the credibility of its creators and spreaders.

Rubin et al. (2016) describes three types of fake news in contrast to reporting. These are - serious fabrications (uncovered in mainstream or participant media); large-scale hoaxes; humorous fakes (news satire, parody).

Zellers et al. (2019) discussed the threats posed by automatically generated propaganda articles that closely mimics the style of real news. They have designed a language model-based system called Grover for the controllable generation of text from the title of the news. Humans may find this generated text to be more trustworthy compared to the actual news article. Such type of fake news called neural fake news is discriminated best using the generator system itself.

Bradshaw and Howard (2017) compared the teams who spread manipulated information, also called disinformation through social media and news across 28 countries including India. These types of fake news are created manually to influence the voters and domestic audiences purposely. Melford and Fagan (2019) designs a Global Disinformation Index (GDI) to combat the disinformation.

Another essential type of fake news is created by proliferating stylistic bias in the text. Pérez-Rosas et al. (2017) have used linguistic features in Support Vector Machine (SVM) to detect these fake news in some English newspapers. Rubin et al. (2016) discriminated between synthetic and legitimate news using 5 features namely Absurdity, Humor, Grammar, Negative Affect, and Punctuation.

3 Dataset Preparation

We evaluate our proposed framework on two datasets, Kaggle Bengali news, and Online Bengali news. For the time being, we are not using the name of the newspapers to avoid the controversy. In total, we have 25K news covering seven different domains, namely Kolkata, State, National, Sports, Entertainment, World, and Travel. Each news contains on an average of 15 sentences. The distribution of the Bengali news dataset in the seven domains is shown in Fig. 1

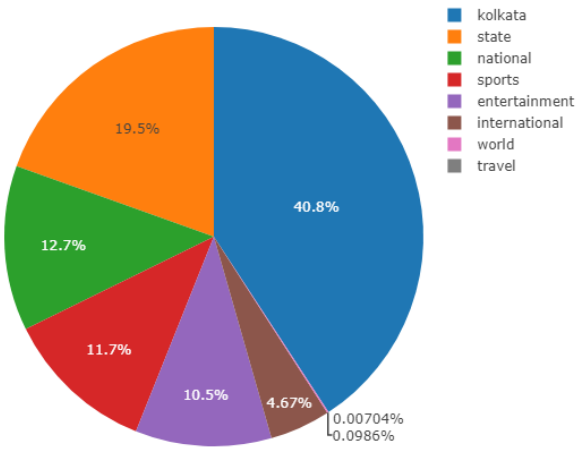

Figure 1: Bengali News Dataset Distribution in Kolkata, State, National, Sports, Entertainment, World, and Travel Domains

3.1 News Content Features

The structure of the news dataset is listed below:

1. **Source:** Author or publisher of the news article.
2. **Domain:** The domain of the news is defined in this field. In this dataset, we have seven different domains viz, Kolkata, State, National, Sports, Entertainment, World, and Travel.
3. **Headline:** Short title text that aims to catch the attention of readers and describes the main topic of the article
4. **Body Text:** Main text that elaborates the details of the news story, there is usually a significant claim that shapes the angle of the publisher.

Depending on these raw content attributes, different kinds of feature representations can be built to extract discriminative characteristics of synthesis news. The news content we are looking at mostly be linguistic-based features, discussed in the following section.

3.2 Preprocessing of News Sentences

The overall framework of the machine learning-based classifier is divided into three parts: Cleaning of raw text, feature extraction and synthetic

news classification. The collected news sentences are annotated manually as a legitimate or synthetic sentence. It is difficult to deal with raw news due to noise. The noisy news includes:

- Keyphrases: - নিজস্ব প্রতিবেদন, ওয়েব ডেস্ক:, এই বিষয়ে অন্যান্য খবর, ব্যুরো, ডিজিটাল ডেস্ক, সূত্রের খবর
  Different publication media use the mentioned key phrases which are actually not part of the news. We remove these phrases from the news sentences.

- English Sentences: News contains some English sentences along with Bengali sentences. The following English sentence is highlighted inside a Bengali sports news, e.g.,*"Delhiites get a bite of #ViratKohli quite literally at #MadameTussauds PC Statesman pic.twitter.com/FNLARdIQi6 - Bharat Sharma (sharmabharat45) June 7, 2018", "ISIS," "JNU"*. We remove such non Bengali sentences from our dataset.

- Stop Words: Stop words are described as the most common words that occur in any corpus of a particular language. At the preprocessing step, we remove stop words such as - 'এ', 'এবং', 'আর' from the sentences. Here we created a stop word list of 360 words and then these words are removed from the corpus.

- Word Stemming: Word stemming is applied to map the words with different endings to a single one such as চ্যালেঞ্জের,চ্যালেঞ্জ becomes চ্যালেঞ্জ. Bengali is a very inflectional language for which stemming is required for further processing.

- Other: News contains emoticons, symbols, and pictographs. We remove them by using Unicode.

By using the above-mentioned list of phrases, we preprocess the raw news and generate a clean text for further processing.

3.3 Annotation Guideline

We have annotated Bengali news sentences into two categories: synthetic and legitimate. In this section, we discuss the method we have followed in tagging with some examples.

- *Example-1:* এই সম্প্রদায়ের এক সদস্যের মতে, ২০১৫ সালে আমেরিকাতে এবং ২০১৭ সালে ইংল্যান্ডে সভা করেছিলেন প্রধানমন্ত্রী।
  [According to a member of this community, the Prime Minister had a meeting in the United States in 2015 and England in 2017.]
  In this sentence, it is claimed that the statement is taken from somebody, but the name is not mentioned explicitly. This is why, we consider such types of sentences as synthetic. If the name of the claimer is added, then it is converted to legitimate.

- *Example-2:* সুপার ওভারেও খেলার শেষ না হওয়ায় বাউন্ডারির সংখ্যার ভিত্তিতে ইন্ডিয়াকে চ্যাম্পিয়ন ঘোষণা করে দেওয়া হয়।
  [India was declared champion on the basis of the number of boundaries as the game did not end in the Super Over.]
  The cause-effect sentence of Example-2 is considered legitimate as it is based on a true fact cause, and the relation is an established relation: if a game does not end in Super Over then go for a number of boundaries.

- *Example-3:* সুপার ওভারে উত্তেজনা, শিষ্য নিশামের ছক্কা দেখে শেষ নিঃশ্বাস গুরুর।
  [Tension in the Super Over, the master releases last exhale after seeing the six of the disciple Neesham.]
  In the cause-effect sentence of Example-3, the cause is a true fact, but the relationship is based on an assumption or probability. There is no rule in the environment to state that one will die after seeing one's six. Therefore this sentence is tagged as synthetic.

- *Example-4:* সম্ভবত মন খারাপের কারনে, বোল্ট তার সেরা পারফরম্যান্স দিতে পারলনা।
  [Probably due to distress, Bolt could not give his best performance.]
  This is also a cause-effect sentence. In this sentence, the cause is not a true fact as a probability is associated with it. Though, the relation "if somebody is in distress, then he will not be able to give his best performance" is an established relation this sentence has synthetic property.

- *Example-5:* The phrases containing synthetic adjectives have a synthetic property like: মন ভাঙা বোল্ট বললেন

[broken heart Bolt told]
We observed that verbs carry the best clue about the synthetic property. Therefore we have created a clue verb list and used it as a feature. But there are some adjectives, adverbs, nouns, which can also be considered a clue. We considered them during annotation.

- *Example-6:* কিন্তু, আমরা তাঁদের আশাপূরণ করতে পারিনি।
[But we could not meet their hope.]
This sentence talks about an abstract mental state (hope). It is not defined how to measure whether it is met or not. Therefore, this is a synthetic sentence.

- *Example-7:* তবু মণীশ পান্ডের শতরানের সৌজন্যে বড় রান তুলে ফেলে ভারতীয়।
[Yet by courtesy of Centurion Manish Pandey India built a big score.]
Here, the word 'yet' makes this sentence synthetic, as it means it would not be possible without him. But it is not correct as others are not tested. Dropping this word leads to a legitimate sentence.

4 Proposed System

After preprocessing of the raw news, the news is tokenized and segmented into sentences level. In this paper, we create a hybrid approach by combining Binary Machine Learning Classifiers using the Voting approach and then postprocessing with a Rule-Based approach. The proposed system is as follows:

1. For machine learning, features are generated from the sentences, and after that, we apply supervised machine learning algorithms, namely Support Vector Machine, Naive Bayes, K Nearest Neighbors, Random Forest, Decision Tree, and Logistic Regression.

2. According to the results of supervised algorithms, we are creating some rules based on the mismatched outputs.

3. Lastly, we are combining the supervised algorithms using a voting approach. We are giving a higher preference for synthetic tagging. If among the six classifiers, 3 classifiers tell the sentence is synthetic and 3 classifiers tell it is legitimate, then we annotate it as synthetic. However, if more than 3 classifiers tell that the sentence is legitimate, then it became legitimate.

In this paper, we consider synthetic news classification for independent sentences as each sentence carries some synthetic or legitimate property. Our approach is to use a committee of classifiers, each trained on a set of text features. The entire list of features is presented in this section.

4.1 Feature Selection in Synthetic Classification

Feature Selection of any classification problem takes a crucial part. Each sentence is represented by a feature vector which contains numerical features that represent the occurrence frequency or weight of a feature or binary features (occurrence or non-occurrence of a feature) or a ternary feature. The features are listed below:

1. Punctuation: Various punctuation is used to indicate different types of sentiments. The use of punctuation can help the synthetic news detection algorithm to differentiate between funny, entertaining, deceptive, and truthful texts. This feature has two values (binary) - Exclamation (!) and Question Mark (?). This feature is used because, according to our observation the exclamatory sentences and question sentences are more prone to be synthetic. This feature helped us to improve the precision of synthetic sentence identification.

2. Named Entities: News tells a story related to a particular incidence. Legitimate news is having the property of some person telling something or making comments. Named Entities are used many times inside news. We have observed that the sentences having a Named Entity mostly become a legitimate sentence. This is because even if an incidence is false or synthetic, but it is told by a particular person we consider it to be true. Consider the example: "a person x told that he is probably sick". Here the 'probably' word makes the statement synthetic. But the sentence is legitimate as the person x has actually told it. Thus, use this binary feature (there exist a Named Entity or not) may be helpful to predict synthetic news sentence.

   To find the named entity in the news sentence, we manually annotate 42k words from 6k

news sentences to train CRF++ model. Here, we are using POS tag as a feature of CRF++ to find the presence of named entity in the sentence. To get the POS tag we have used some of the features proposed by Dandapat et al. 2007 which are available with us. Along with the POS tag the Bengali gazetteer list of 1200 names is also used in training the CRF++ model. Then the model is tested with our 530 news examples and we get 94.3% accuracy. We have used this Bengali gazetteer list as a binary feature (presence or non-presence of named entity).

3. Domain-specific Clue Verb: We have seen that maximum synthetic sentences have some verbs which carry some indication for the sentence to have synthetic property. Similarly, some verbs are indications of the legitimate property of the sentence. Theses verbs are defined as Clue Verbs. We are attempting to make a list of words for Legitimate sentences and another for synthetic sentences manually. By analyzing a large corpus, we make a list as shown in 1, which is not exhaustive.

   We have considered this clue verbs as a ternary feature as follows. If there is a Legitimate Clue Verb then it is 1 (we do not need to consider Synthetic Clue Verb); if there is no Legitimate Clue Verb, but there is a Synthetic Clue Verb then it is 2, and if there is neither Legitimate Clue Verb nor Synthetic Clue Verb then it is 3. The value 1 indicates that the sentence is a strong candidate for being legitimate; the value 2 indicates that the sentence is a strong candidate for being synthetic and the value 3 indicates that there is no clue about its property. According to our observation, this feature is the most effecting feature in the classification task.

4. TF-IDF: Term Frequency - Inverse Document Frequency (TF-IDF) of a term is used to denote its importance. TF(w,d) denotes the raw count of the word (w) in a news document (d) and IDF(w,D) is a measure of how much information the word (w) provides, i.e., if it is usual or rare across all news documents (D). Finally, TF-IDF is defined as follows.

$$tf\_idf(w, d, D) = tf(w, d) \times idf(w, D) \quad (1)$$

   Thus, TF-IDF is used to determine the importance of words in news domain. We have combined the TF-IDF of words of the input sentence to calculate the importance of sentence in news domain. Considering that the synthetic sentences carry more importance we have used it as a feature in our task.

4.2 Machine Learning-Based Classification

Synthetic news sentence classification may be done at the document level, sentence level, and phrase level. We are considering the news article is based on an actual fact. Some of its sentences are synthesized by the author to attract the engaged reader. Our objective is to identify those synthetic sentences. So, in this paper, sentence-level classification is considered where an independent sentence is classified as synthetic sentences and legitimate sentences.

A supervised binary classifier algorithm may be used to identify the synthetic sentences. Several supervised machine learning techniques have been examined in the paper to classify the sentences into classes. Those are Support Vector Machine, Naive Bayes, K Nearest Neighbors, Random Forest, Decision Tree, and Logistic Regression (LR). We have considered the Punctuation feature, Named Entity feature, clue verb, and TF-IDF feature, as discussed in Section 4.1.

The Confusion Matrix is one of the most intuitive metrics used for finding the correctness and accuracy of the model. The Confusion Matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on the Confusion Matrix and the numbers inside it. The confusion matrix is a table with two dimensions (“Actual” and “Predicted”), and sets of ”classes” in both dimensions.

The following terms are associated with a confusion matrix.

- True Positive (TP): when predicted synthetic sentences pieces are actually annotated as a synthetic sentence;
- True Negative (TN): when predicted legitimate sentences pieces are actually annotated as true sentences;

Label	Clue Verb
Legitimate	বলছেন [Says-H], জানিয়েছেন [Said-H],জানান [Said-H], বললেন [Said-H],ঘোষেণামতো [Declaration], কথা বলেন[Speak-H]কথা বললেন [Speak-H], জানান[Tell me], জানালেন[Told], ঘোষণা করলেন[Announced]
Synthetic	খতিয়ে দেখা[Check it out], খতিয়ে দেখেন[Check it out],খতযি়ে দেখা হয়[Is checked], প্রশ্ন করেছেন [Asked-H], প্রশ্ন করে [Asked-H], প্রশ্ন করা হয়েছে[Asked-H], প্রশ্ন করা[Asked-H], প্রশ্ন হয়েছে [The question has been], প্রশ্ন উঠেছে [The question arises], কথা ছিল [There was talk], কথা দেওয়া[Promise-H], কথা দিয়েছিল[Promised-H], কথা দিয়েছিলেন[Promised-H],কথা দিয়ে রাখেননি[Didn't talk], প্রশ্ন উঠতে শুরু করেছে[The question has started to arise], আস্থা নেই[Not confident], মনে করা হচ্ছে[It seems], চ্যালেঞ্জ নিয়েছেন [Have taken up the challenge], চ্যালেঞ্জ ছোড়েন[Throw the challenge], চ্যালেঞ্জ নেওয়া [Take up the challenge], চ্যালেঞ্জ ছোড়া[Throw the challenge], চ্যালেঞ্জ[challenge-H],মনে করা হচ্ছে[It seems], প্রমাণিত হবে[Will prove], নীতি নিয়েছে[Policy taken], সক্রিয় ভাবে পথে নামতে দেখা গিয়েছে[Actively shown on the way down], দাবি <Null verb>[Claim-H], খবর চাউর হয়ে যায়[The news goes sour], আতঙ্কের ছাপ <Null Verb>[The impression of terror], আতঙ্কের ছাপ নেই[No sign of panic], তদারকি করেন[Take care], ফের সতর্ক করে দেন[Warns again], দাবি তুলেছিলেন[Claimed-H], তবু স্বস্তি ছিল[Yet there was relief], দেখছেন স্থানীয়রা[The locals are watching], আস্থা নেই[Not confident], আশ্বাস দিয়েছেন[Assured-H], আশঙ্কা[Fear], আশঙ্কা করেছেন[Have feared], আশঙ্কা করা[Do not be afraid], আশঙ্কা করল[Apprehensive], রয়েছে বলে[Say there is], দেওয়ার অভিযোগ[The charge to give], চাঙ্গা হয়ে ওঠে[Became stronger], অভিযোগ তুললেন[Complain-H], অভিযোগ করল[Complained-H], অভিযোগ করা হয়েছে[The complaint was made], অভিযোগ[Complain-H], কড়া বার্তা[Strong message], রুখে দাঁড়িয়েছিলেন [Standing in the stands], রুখে দাড়ান-H[Stand up], রুখে দাড়ালেন-H[Stand up], তোপের মুখে-H[Under the cannon], মাঠে নামাচ্ছেন-H[Getting down on the field], স্বীকৃতি দিয়েছেন[Recognized], স্বীকৃতি দেওয়া[recognition-H], স্বীকৃতি মিললে[Acceptance-H], বিচারের মুখোমুখি[Facing trial], অভিযোগ উঠল[The complaint arose], অভিযোগ উঠা[Complaints-H arise], বলে অভিযোগ[Complain-H], একাংশের দাবি ছাপ <Null Verb>[Part claim impression], একাংশের মত[Like a part], উস্কে দিল[Instigated-H], কীসের ইঙ্গিত[What a hint], জানা গিয়েছে[Got it]

Table 1: List of Clue Verb [H: Honorific]

- False Negative (FN): when predicted legitimate sentences pieces are actually annotated as synthetic sentences;
- False Positive (FP): when predicted synthetic sentences, pieces are actually annotated as legitimate sentences.

We have created a separate set of 106 sentences to create the Confusion Matrix. The Confusion Matrices of all the Binary Classifier techniques for these sentences are shown in Fig. 2.

4.3 Rule Based Postprocessing

After analyzing the errors in the confusion matrix of the Binary Classifier techniques, as shown in Fig. 2, we have formulated an initial set of rules. These rules are used in the postprocessing step to correct some of the errors. The rules we formulated are discussed below.

1. If the **topic** of the news sentence is clubbed with old news of different **topic**, then it is considered as synthetic. Consider the following example.
   সারদা মামলায় রমেশ গাঁধীকে নিজেদের হেফাজতে নেওয়ার পরে এ বার রোজ ভ্যালির দুই কর্তাকে গ্রেফতার করল তারা|
   [After taking Ramesh Gandhi in their custody in Sarada case now they arrested two heads of Rose Valley]
   In this sentence, the leading news on arresting two heads of Rose Valley is clubbed with old news of different topics. Therefore it is considered synthetic.

2. If in the news sentence the reason for an incidence is written abstractly, then it is considered as synthetic. Consider the following example.
   শাসকদলের গোষ্ঠীকোন্দলের জন্য সোমবার চাষিদের নথিপত্র জমা দেওয়ার শিবির বেড়াবেড়ি থেকে সরে গেল সিঙ্গুর বিডিও অফিসে|
   [Due to infighting in the governing party Monday the camp for submitting documents of farmers is moved from Beraberi to Singur BDO office.]
   We have prepared an initial list of abstract reasons. In this sentence, the reason ”infighting” is not concrete. Therefore it is considered synthetic.

3. If there is an incomplete list of names, matters, topics, or any other thing in the sentence, then it is considered as synthetic. Consider the following example.

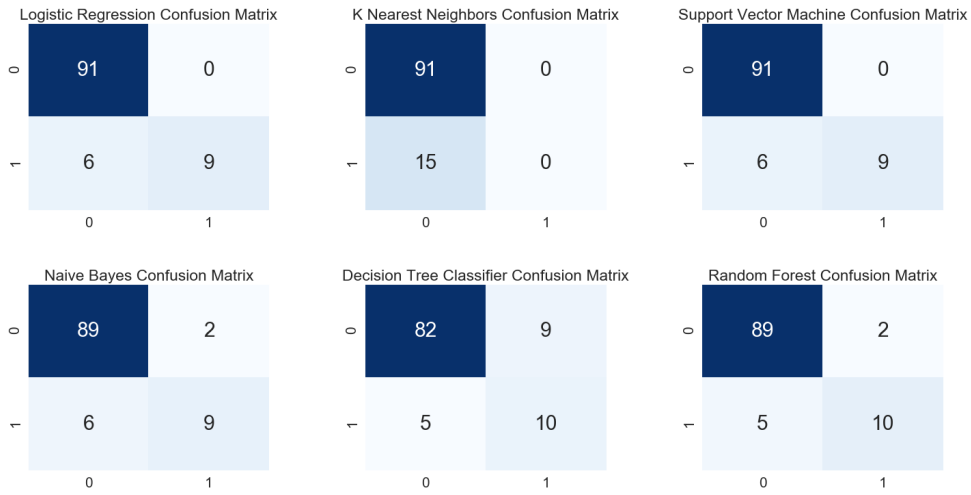

Figure 2: Confusion Matrix after Applying 6 Binary Classifiers on 106 Sentences[X axis denotes predicted value and Y axis denotes actual value and both cases 0 indicates legitimate sentence and 1 indicates Synthetic Sentence]

এজন্য বর্ধমান, কল্যাণী এবং অন্যান্য কিছু বিশ্ববিদ্যালয়কে আবেদন জানাতে বললেন শিক্ষামন্ত্রী|
[For this, the education minister asked the Bardhaman, Kalyani and some other universities to apply.]
In this sentence, the phrase ”some other universities” indicates that an incomplete list is used. Therefore it is considered synthetic.

4. If a sentence contains a legitimate clue verb, then it is classified as a legitimate sentence. But with the clue verb if there exists an adverb, then the sentence becomes synthetic. Consider the following sentence.
শিক্ষার সর্বস্তরে শিক্ষকশিক্ষিকাদের হাজিরায় বিশেষ নজর দেওয়ার কথা তিনি বারবার বলেছেন
[He has repeatedly said to pay special attention to the attendance of teachers in all levels of education.]
In these sentences, an adverb is used with the clue verb. Therefore it is considered as synthetic.

5. If in the sentence there exist any phrase in Double Quotation indicating a comment, then it is a legitimate sentence irrespective of the property of the comment. Consider the following example.
তিনি জানান, "এই বিষয়ে সকলের জন্য একটি সাধারণ নিয়মাবলী থাকলে ভাল হয়"
[He said, ”it is better to have a general rule for everyone in this regard”.]
This sentence tells the comment made by some entity. Therefore it is considered legitimate.

5 Experimental Result

Various evaluation metrics have been used to evaluate the performance of machine learning models. We want to test the performances of our models in terms of Accuracy, Precision, Recall, and F1-Score. These metrics are commonly used to evaluate the machine learning models and enable us to evaluate the performance of a classifier from different perspectives. The results of the k-fold cross-validations for each of our hybrid models are shown in section 5.1.

Then we have applied the voting approach to combine the models and then the rules. The final accuracy of this hybrid system is discussed in section 5.2.

5.1 Classification Performance of Individual Hybrid Machine Learning Models

Firstly, we have tested the six binary classifiers namely Logistic Regression, K Nearest Neighbors, Support Vector Machine, Naive Bayes, Decision Tree, and Random Forest. We have used 530 sentences annotated as legitimate or synthetic. These sentences are folded into Training and Test data in the proportion of 80:20.

We have used k-fold cross-validation ($k = 5$) and calculated the above metrics in each of the k-folds for each model. Then, we have manually applied the rules on all the six prediction models. We prepared a comparison chart of the final performances of these six individual hybrid systems which is given in Fig. 3.

In Fig. 3, the dark gray colour bars indicate the k-fold cross-validation results of the binary clas-

Approach	Accuracy	Precision	Recall	F1-Score
Logistic Regression based Hybrid System	0.82	0.67	0.82	0.74
Combined Model based Hybrid System	0.86	0.86	0.87	0.85

Table 2: Performances of different approaches for Synthetic Sentence classification

Figure 3: Classifier Result

sification. The light gray colour bars indicate the result we got after applying the rules on the best output of the corresponding technique. The result shows that the rules improved the Support Vector Machine and LR based techniques most. The Logistic Regression based technique combined with the rules gave the highest accuracy up to this stage. We are getting highest around 82% accuracy by applying rules on the Logistic Regression based model.

5.2 Classification Performance of Combined Hybrid Machine Learning Model

Finally, we have used a voting based combination of these six machine learning classifiers. If a sentence is tagged as synthetic by 3 or more classifiers then we consider it to be synthetic. Otherwise, it is considered to be legitimate. Then, we have applied the rules on the combined classifier. The final result of this hybrid system is shown in Table 2.

6 Conclusion and Future Scope

In this paper, we defined and compared synthetic and legitimate sentences and highlighted many interesting differences between these two categories. We then utilized these differences as features to detect synthetic sentences. We have proposed a hybrid approach that can detect synthetic news. To the best of our knowledge, our work is the first attempt to detect synthetic news at the Bengali news sentence level. In the future, we want to extend it to use semantic features in the Machine Learning model and calculate the degree of synthetic property in the synthetic sentences. Then we want to compare the renowned Bengali newspapers in terms of the usage of different types of synthetic sentences.

References

Oliver Batchelor. 2017. Getting out the truth: the role of libraries in the fight against fake news. *Reference services review*, 45(2):143–148.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7.

Samantha Bradshaw and Philip Howard. 2017. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *Technical report, Oxford Internet Institute*.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Sandipan Dandapat, Sudeshna Sarkar, and Anupam Basu. 2007. Automatic part-of-speech tagging for bengali: An approach for morphologically rich languages in a poor resource scenario. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 221–224. Association for Computational Linguistics.

Clare Melford and Craig Fagan. 2019. Cutting the funding of disinformation: The ad-tech solution. *Technical report, The Global Disinformation Index*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

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