ISSN: 2454-9940



E-Mail : editor.ijasem@gmail.com editor@ijasem.org





Enhancement of textual data for word embeddings in the classification of fraudulent news

KASANI DURGA SAI SRI	PERUMALLA NAGARAJU	ANDE BENJAMIN TYCHICUS
Student,Computer Science and	Student, Computer Science and	Student, Computer Science and
Engineering Department	Engineering Department	Engineering Department
Sasi Institute of Technology	sasi Institute of Technology	sasi Institute of Technology
Engineering	Engineering	Engineering
Tadepalligudem, INDIA	Tadepalligudem, INDIA	Tadepalligudem,INDIA
durgasri.kasani@sasi.ac.in	nagaraju.perumalla@sasi.ac.in	<u>benjamin.ande@sasi.ac.in</u>
ALLUR	I ABHINAY Dr.Kalli Srinivasa Nag	geswara Prasad

Student, Computer Science and Professor.Computer Science and Engineering Department Sasi Institute of Technology Engineering

Tadepalligusem, INDIA

abhinay.alluri@sasi.ac.in

Abstract— Data amount and quality affect categorization algorithm efficiency. Algorithm accuracy is great when the size of the data corpus is enormous. A low level of algorithm accuracy is indicative of a little data set. In order to better classify fake news, this article Text Data Augmentation. use In this study, the author evaluated methods for synonym replacement, back translation, and function word reduction. Word2Vec Skipgram models will transform text enriched using the aforementioned approach into a numerical vector. Bernoulli NB, RF, SVM, and LR are some of the classifiers that will be used to train Word2Vec. Random Forest on original text, improved Logistic Regression with 'Reduction of Function Words,' and SVM and Bernoulli naïve Bayes on Back Translation upgraded text all lead to high accuracy.

Engineering Department

Sasi Institute of Technology

Engineering

Tadepalligudem, INDIA

ksnprasad@sasi.ac.in

Keywords— Original Corpus, Synonym replacement (SR), Back translation (BT), Reduction of function words (FWD)

INTRODUCTION I.

Data augmentation eliminates the need to provide data while increasing the diversity of training examples [1]. By subjecting machine



learning models to a variety of scenarios, data augmentation aims to boost their performance and resilience. Augmenting data is useful in computer vision, NLP, and speech recognition. By expanding the variety of data the data, augmentation decreases overfitting. Putting an end to the process of the machine learning using example data used for training. Regardless of the apparent success, NLP studies including computer significant vision issues have not reaped advantage from DA systems thus far [2]. The most common methods for improving text data are paraphrasing, random insertion, swapping, and deletion; back translation; and synonym substitution. Outlined in papers [1], [2] are data augmentation campaigns. Many modern apps rely on text vectorization Manage natural language processing tasks involving classification. Word2Vec, Doc2Vec, and Glove are just a few of the popular word embedding methods that continue to thrive by capitalizing on words' semantic similarities.

Practically speaking, word vectors have a wide range of uses. In this part, the focus is on classification jobs. Improving classification performance metrics through data augmentation strategies is the focus of this study. To identify which data augmentation strategy offers the best assistance with classification difficulties, this post will take a look at a few popular options. From the various alternatives, the following techniques were chosen:

• Substituting a term with one of its synonyms is known as synonym replacement (SR). For this purpose, we use WordNet, an extensive database of words and their meanings.

• When it comes to improving written content, back translation (BT) is both easy and effective. It involves translating the source text from one language to another and vice versa. The real text is often slightly altered using this procedure yet important details are preserved.

• Random Deletion is similar to Function Word Reduction (FWD). Using a probability parameter, the approach extracts words from a given text. The likelihood parameter was restricted to words pertaining to function and content due to our context. • The original corpus is consulted for reference purposes only; no data augmentation methods are employed. We used the Original as a benchmark and tested various methods against it [1, 2].

Time and computing power are needed to apply these strategies to classification tasks. The relevance of these ideas is up for debate. We set out to find out in this post whether these tactics really do boost classification model performance metrics. We also want to find out which approach improves categorization model results the most.

The purpose of this essay is to compare and contrast different classification systems and their relative merits. By determining their relevance, researchers will be better able to choose approaches for future categorization issues.

In order to get the corpus ready, we follow the supplied procedures. Train the Word2Vec Skipgram model using the prepared corpus for classification tasks that involve word embeddings. For WELFakedataset, the task was to categorize false news. The classification's performance metrics will be evaluated next. So, we won't be evaluating Data. We will attack the classification issue by utilizing augmentation tactics and then evaluate their efficiency.

classification.

There are a number of options for evaluating different corpus preparation methods for huge language models. In their performance evaluation, Nazir et al. [3] used WordSim-353 [4] and SimLex-999 [5], two sets of records that contain the similarity between words. Word vectors are frequently used in educational examples to show how to compute word similarities.

II. LITERATURE REVIEW

1. A survey of data augmentation approaches for NLP

<u>A Survey of Data Augmentation Approaches for</u> <u>NLP</u>

Abstract: Interest in DA for natural language processing has grown in recent years, particularly in research pertaining to new problems, low-resource domains, and large-scale neural networks that require a substantial amount of training data. The discontinuity of the linguistic evidence may



explain why this subject has remained unexplored despite increasing studies in the field. This research compiles and organizes the literature on data augmentation in NLP. Next, we'll go over some important tactics for data augmentation in NLP. Neural Network Processing (NLP) methods for typical jobs and uses are outlined below. Finally, we cover current difficulties and potential directions for future study. Finally, we hope that this effort will either shed light on data augmentation for NLP literature or inspire other studies in this area.

2. Data augmentation techniques in natural language processing.

https://arxiv.org/pdf/2110.01852

Abstract: In cases when deep learning is unsuccessful, DA can be used to address data shortages. The extensive usage of it in computer vision and NLP has improved several issues. A primary goal of DA methods is to increase the model's generalizability to novel testing data by diversifying the training data. We classify DA methods as either paraphrasing, noising, or sampling in this study according to the different types of enhanced data. The aforementioned DA methods are the focus of our investigation. Aside from that, we go over the issues and uses of their natural language processing capabilities. Key resources are in Appendix A.

3. Toward the development of large-scale word embedding for low resourced language

<u>Toward the Development of Large-Scale Word</u> <u>Embedding for Low-Resourced Language | IEEE</u> <u>Journals & Magazine | IEEE Xplore</u>

Abstract: To syntactically and semantically alter unlabeled data, natural language processing employs word embedding. With the help of vector space representations of the recovered corpus features, It is feasible to do things like summarize, simplify, and forecast the following line, among other things. By considering the frequency and cooccurrence of words, the matrix Noisev contrastive estimation, factorization, skip-gram, and hierarchical-structure regularizer embed them. Researchers have concentrated on Urdu because of its large speaker population (231.3 million) and the fact that these methods have yielded fully developed word vectors for the majority of spoken languages. In this study, we look at word embedding in Urdu. Among the categories

represented in our dataset were: Industry, athletics, medicine, government, show business, scientific discoveries, and international news. This dataset yielded 288 million tokens. We used the skip-gram (word2vec) model for word vector creation. No more than 100, 200, 300, 400, 500, 128, 256, or 512 vector dimensions could be used for embedding. Assessments were conducted using Lexsim-999 and Wordsim-353 annotated datasets. The suggested study indicated that wordsim-353 and Lexsim-999 had Spearman correlation coefficients of 0.66 and 0.439, respectively. Compared to state-of-the-art, the results were superior.

4. A study on similarity and relatedness using distributional and WordNet based approaches

<u>A Study on Similarity and Relatedness Using</u> <u>Distributional and WordNet-based Approaches</u>

Abstract: Methods for similarity based on WordNet and distributional analysis are compared in this paper. Each relatedness and similarity method is discussed, along with its advantages and disadvantages, and a hybrid approach is proposed. Each of our methods outperforms the competition on the RG and WordSim353 datasets, and when combined, they provide the top results across the board. Our methods can be easily modified with little loss, and we pioneer crosslingual similarity.

5. SimLex-999: Evaluating seman tic models with (Genuine) similarity estimation

[1408.3456v1] SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation

Abstract: To evaluate distributional semantic models, we present SimLex-999, a benchmark that outperforms existing tools in multiple respects. In contrast to industry leaders like WordSim-353 and MEN, which place a broader emphasis on relatedness and association, it performs poorly categorizing things that are similar but different [Freud, psychology]. Compared to conceptual association models, SimLex-999 models are more versatile because they prioritize similarity. Along with concreteness and (free) association strength grades, SimLexprovides abstract and concrete noun, 999 adjective, and verb pairs. The ability to assess model performance on various ideas with granularity is made possible by this diversity, which in turn reveals opportunities to enhance



structures. When tested on SimLex-999, modern models consistently fall short of the interannotator agreement ceiling. In this way, SimLex-999 can gauge progress in distributional semantic models, which will power future representationlearning systems.

6. EDA: Easy data augmentation techniques for boost ing performance on text classification tasks

[1901.11196] EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

Abstract: Through the use of data augmentation, EDA simplifies text classification. Simple but successful methods used by EDA include swapping, deletion, random insertion, and synonym substitution. RNNs and CNNs both benefit from EDA's enhanced performance on five different text classification tasks. Training EDA with only half of the training set yields results that are on par with full training, according to five different datasets. Our reasonable numbers were the result of thorough ablation studies.

7. Improving short text classification through global augmentation methods

(PDF) Improving Short Text Classification Through Global Augmentation Methods

Abstract: Various techniques for improving text are investigated. To do this, we scour three databases of news articles and social media posts. Our goal is to enlighten researchers and practitioners on how to select augmentation for classification applications. We discover that augmentation based on Word2Vec can function even in the absence of a formal synonym model such as WordNet. By reducing overfitting in tested deep learning models, Mixup improves all text-based augmentations. For common and low-resource use cases, round-trip translation through a translation service is too costly and unavailable.

8. Contextual augmentation: Data augmentation by words with paradigmatic relations

<u>Contextual Augmentation: Data Augmentation by</u> <u>Words with Paradigmatic Relations - ACL</u> <u>Anthology</u>

Abstract: Our latest data augmentation technique is labeled sentence contextual augmentation. Even when paradigmatic relations stand in for actual words, we still consider

ISSN 2454-9940 www.ijasem.org Vol 19, Issue 2, 2025

sentences to be naturally occurring. Predicting arbitrary word substitutions is the job of a bidirectional language model. Word improvement is a good fit for the many context-predicted words. In addition, we extend sentences without impacting label compatibility by modifying a language model with a label-conditional design. We found that the proposed method improves the performance of classifiers built using convolutional or recurrent neural networks across six different text categorization tasks.

9. SynoExtractor: A novel pipeline for Arabic synonymextraction using Word2 Vecwordembeddings

<u>SynoExtractor: A Novel Pipeline for Arabic</u> <u>Synonym Extraction Using Word2Vec Word</u> <u>Embeddings - Al-Matham - 2021 - Complexity -</u> <u>Wiley Online Library</u>

Abstract: Automated synonym extraction is a feature of many NLP systems, such as those that retrieve information or answer questions. Recent work has used word embeddings to extract semantic linkages by capturing word relatedness and similarity. Synonym extraction is challenging since word embeddings do not have the ability to determine if two words are synonyms or have any other kind of semantic link. This research presents the SynoExtractor pipeline, a tool for preserving synonyms by filtering related word embeddings according to linguistic principles. For our experiments, we utilized Gigaword and KSUCCA embeddings, as well as CBOW and SG models. Using Alma'any Arabic synonym thesaurus, we compared automatically extracted synonyms. We had two native Arabic speakers review it by hand. When compared to cosine similarity, our results demonstrate that the SynoExtractor pipeline significantly enhances the accuracy of synonym extraction. Gigaword and the King Saud University Corpus of Classical Arabic were both enhanced using SynoExtractor; the former by 21% and the latter by 0.605. Comparing Sketch Engine to SynoExtractor, the latter performed 32% better when it came to MAP synonym extraction. The ability to extract synonyms demonstrates the method's generalizability to different languages.

10. On the use of text augmentation for stance and fake news detection

(PDF) On the use of text augmentation for stance and fake news detection



Abstract: DA(DATA Augmentation)can be used to generate latest training examples by updating old ones. Notable benefits of DA include alleviating class imbalance, avoiding data scarcity, and enhancing generalization. This study delves at the use of DA to identify bogus news and take a stance. The impact of DA methods on classification conventional algorithms is investigated in the first part of our study. The flawed "the more, the better" approach to text augmentation is exposed by our study, which demonstrates that no one-size-fits-all strategy exists. Part two of our research presents an ensemble learning method that uses augmentations. To enhance ensemble prediction performance, the proposed method augments base learners with text to increase their diversity and accuracy. Class imbalance with DA is the focus of our final experiment. As a result of social stratification, algorithms for detecting bias and fake news are skewed. Both moderate and severe imbalances can be remedied through text augmentation, according to this study.

III. PROPOSED SYSTEM

To make the suggested method better at identifying fake news, increase the amount of text data in the training dataset. To transform text into numerical vectors, the Word2Vec Skip-gram model employs SR, BT, and FWD reduction.

The impact of vector augmentation on classification accuracy will be assessed by a number of ML classifiers, including RF, SVM, Logistic Regression, and Bernoulli Naïve Bayes.

Performance will be enhanced with the addition of XGBoost and other modern ensemble techniques. In order to improve the system's ability to identify false news and overcome dataset restrictions, this strategy is employed. Measures of performance will include F1-score, recall, precision, and accuracy.

Extension: We enhanced the suggested system with deep learning models and powerful ensemble algorithms including LightGBM, CatBoost, and XGBoost. Training and categorization of features

are both improved by this extension. A news classification user interface was also developed by us using Flask.

Advantages:

- The proposed method makes advantage of text data augmentation to incorporate more diverse and variable samples into the training dataset, which in turn improves the model's classification abilities.
- To increase the accuracy and resilience of detecting faked news, the proposed method employs ensemble learning with advanced techniques like XGBoost. This combines the skills of several classifiers.
- The suggested method use the Word2Vec Skip-gram model to transform textual information into comprehensive numerical vectors that encompass semantic relationships and context. This enables machine learning classifiers to identify minute differences in false news.
- Look at performance measures like F1-score, accuracy, precision, and recall to see how well the system is doing and how far the model has come.
- Using complex algorithms such as XGBoost can enhance the accuracy and dependability of classifying bogus news.
- Flexibility to meet future demands is guaranteed by the ease of scaling to incorporate new algorithms or datasets.
- It is easy to upload news articles for classification using the Flask interface.
- Instantaneous feedback on the veracity of user-generated news is available.



II.SYSTEM ARCHITECTURE



FIG 1.system architecture

The proposed system for fake news classification integrates text data processing, augmentation, and machine learning techniques to enhance classification accuracy. Initially, the raw text data undergoes preprocessing steps such as tokenization, stop word removal, and normalization to prepare it for analysis. Following this, visualizations such as word clouds and frequency plots are generated to gain insights into the dataset. The cleaned data is then vectorized using techniques like TF-IDF or word embeddings to convert it into numerical format. To increase the diversity and generalization ability of the model, data augmentation methods—such as synonym replacement (SR), forward translation (FWD), and back translation (BT)—are applied to generate new samples. The dataset is split into training and testing sets, where multiple machine learning models including SVM, Bernoulli Naïve Bayes, Logistic Regression, Random Forest, and an extended version of XGBoost are trained. These models are evaluated using performance metrics like accuracy, precision, recall, and F1-score to determine their effectiveness. The architecture is designed to handle a variety of textual inputs and is optimized for improved detection of fake news content through robust augmentation and modeling strategies.



IV.EXPERIMENETAL RESULT AND DISUSSION



ISSN 2454-9940



www.ijasem.org

Vol 19, Issue 2, 2025

🗖 🗌 💭 Home F	Page - Select or cre	ate a n 🗙 🗮 AugmentFakeNe	ews - Jupyter No	ote 🗙 🖪	Text Data A	ugmentation	>	< +								_	ō
- 0 0 0	localhost:8888/no	tebooks/AugmentFakeNews.ip	ynb			2				A٩	\$	()	£≞	ſ h	\downarrow	~~	
UPDATE Read <u>the migra</u> xtensions.	<u>ration plan</u> to Note	book 7 to learn about the new fe	atures and the	e actions to) take if you a	are using ext	ensions - F	Please note the	at updating to	Notebook	7 might b	reak so	me of yo	our	٥	on't sho	w any
ຼີ ງເ	upyter Au	gmentFakeNews Last C	heckpoint: 2 h	hours ago	(autosaved)								ę	•	Logout		
File	Edit View	Insert Cell Kernel	Widgets	Help							Trusted	ø	Python	3 (ipyk	ernel)	•	
-	+ % 4	• ◆ ◆ ► Run ■ C	Code	~	-												
	Count	3000 - 2000 - 1000 - 0 - <u><u><u>w</u></u> Dataset Cl</u>	ass Label G	- lean Iraph													
1	In [140]: <i>#fu</i> def	<pre>nction to convert text in vectorize(sentence, mod words = sentence.split(words_vecs = [model.wv[if len(words_vecs) == 0 return np.zeros(100 words_vecs = np.array(w return words_vecs.mean()</pre>	nto wordvec el): word] for w :) ords_vecs) axis=0)	vord in w	vords if w	ord in mo	del.wv]										
I	In [141]: #pro	ocessing news text by ap	olying augm	nentation	n techniqu	е											
1	In [141]: #pro if c	ocessing news text by apposite the second se	olying augm ctor.npy'):	mentation	n techniqu	e muo)		-					~			16:2	4
Type here t	In [141]: #pr if to search	ocessing news text by apposite text by appointed tex	olying augm ctor.npy'): (voctor.npy	mentation	n techniqu Loicklo-T		Þ		•	30°C Pa	rtly sunn	y ^	<u>G</u>) <i>((</i> , <	い) ENG	16:2 06-10-	14 2024
Type here t	In [141]: #pr if o to search	FIG 3. to conve	ert word	ls into	o techniqu Picklo-T Povecto	e Puto \	b	<u>v</u>]	۲	30°C Pa	rtly sunn	y ^	(j) 🍋) <i>(i</i> . <	♭») ENG	16:2 06-10-	14 2024
₽ Type here t	In [141]: #pr if o to search	FIG 3. to conve	etying augm ctor.npy'): (voc	ls into	techniqu b nicklo-T C Vecto	e Involution		<u>*</u>	٠	30°C Pa	rtly sunn	y ^	ê e) // (り)ENG	16:2 06-10-	24 2024
✓ Type here t	In [141]: #pr if to search Page - Select or creat	FIG 3. to conve	elying augm ctor.npy'): (vector power (vector power) ert WORC s - Jupyter Note	ls into	n techniqu N picklo-T P P O Vecto Text Data Aug	e	×	₩] +	¢	30°C Pa	rtly sunn	y ^	<u></u> <u></u>		>) ENG —	16:2 06-10- 	24 2024
✓ ✓ Type here t ✓ ✓ Home P ✓ ✓ ✓	In [141]: #pr if to search Page - Select or creat scalhost:8888/note	FIG 3. to conve	s - Jupyter Not:	ls into	n techniqu	mentation	x	+	•	30°C Pa	rtly sunny	y ∧ Ι ζ≞	ē • •) <i>(</i> , ()») ENG —	16:2 06-10- ට [ි]	24 2024
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/note ation plan to Noteb	FIG 3. to convert te a no x Z AugmentFakeNews.ipy ook 7 to learn about the new feat	elying augm ctor.npy'): (useton.npy ert WOTC s - Jupyter Note nb ures and the a	ds into	techniqu picklor Vecto Text Data Aug	e aua) aua)	×	+ se note that u	C Polating to No	30°C Pa A ^N ℃ tebook 7 mi	rtly sunn ? () ght break	y ^ I €= some o	⊕ ¶ ⊡ f your) <i>(ii</i> , c)») ENG	16:2 06-10-	24 2024 > more
Type here t Type here t Home P To to to The migra nsions.	In [141]: #pr if to search Page - Select or creat ocalhost:8888/note ation plan to Noteb	pocessing news text by appos.path.exists('model/yee poc.path.exists('model/yee with the poly impose with the poly impose <t< td=""><td>eLying augm ettor.npy'): (voctor.npy) ert word s - Jupyter Note nb ures and the a</td><td>ds into</td><td>a techniqu h dickloar D Vecto Text Data Aug ike if you are</td><td>e Deuro) T DI using extense</td><td>× sions - Plea</td><td>+ +</td><td>e pdating to No</td><td>J0°C Pa A^N ☆</td><td>rtly sunny</td><td>y ^ I Ç≡ some o</td><td>ਉ ₩ • (ੇ) f your</td><td>) <i>(</i> (</td><td>)») ENG</td><td>16:2 06-10-</td><td>24 2024 more</td></t<>	eLying augm ettor.npy'): (voctor.npy) ert word s - Jupyter Note nb ures and the a	ds into	a techniqu h dickloar D Vecto Text Data Aug ike if you are	e Deuro) T DI using extense	× sions - Plea	+ +	e pdating to No	J0°C Pa A ^N ☆	rtly sunny	y ^ I Ç≡ some o	ਉ ₩ • (ੇ) f your) <i>(</i> ()») ENG	16:2 06-10-	24 2024 more
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/note ation plan to Noteb	books/AugmentFakeNews Last Ch	oLying augm ctor, npy'): (voctor, npy'): (voctor, npy ert WOTC s - Jupyter Not nb ures and the a eckpoint: 2 hou	dls into	n techniqu n tech	e Duro)	× sions - Plea	₩] + ise note that u	edating to No	30°C Pa A [®] ☆	rtly sunn) [] ght break	y へ I C≞ some o	ē ♥ ■ f your) <i>(ii</i> , (I») ENG	16:2 06-10-	2024
Type here t Type here t Home Pa C O Ioc DATE Read the migra insions.	In [141]: #pr if it o search Page - Select or creat ocalhost:8888/note ation plan to Noteb Upyter Aug Edit View	Decessing news text by appos.path.exists('model/ve. Sector an load("model FIG 3. to convert te a n: x X AugmentFakeNews.ipy books/AugmentFakeNews.ipy ook 7 to learn about the new feat gmentFakeNews Last Ch Insert Cell Kernel	elying augm ettor.npy'): (useton.npy ert WORC s - Jupyter Not: nb ures and the a eckpoint: 2 hor Widgets	entation	a techniqu A dickloar D VeCtC Text Data Aug utosaved)	e Bun) E 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	×	+ se note that u	pdating to No	30°C Pa A [®] ☆ tebook 7 mi	rtly sunn CD ght break	y ∧ I Ç≞ some o	⊡ 🐖 f your €) <i>(ii</i> , ¢	 Don't s out 	16:2 06-10- ඌ ,	24 2024 More
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/note ation plan to Noteb Upyter Aug Edit View + 3 2 5	poessing news text by appos.path.exists('model/yee pos.path.exists('model/yee poetan	eckpoint: 2 hou Widgets	entation Constant Allowing and a sections to take urs ago (au Help	a techniqu	e enal 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	×	₩] +	pdating to No	30°C Pa A ^N ☆ tebook 7 mi	rtly sunn c ght break usted	y ∧ I €≡ some o	E C C))) ENG — © Don't s out I) ●	16:2 06-10- ਹੈ	24 2024
Type here t C O Ioc OXTE Read the migra ensions. File ♥ ●	In [141]: #pr if to search Page - Select or creat ocalhost:8888/not ation plan to Noteb Upyter Aug Edit View + 🕱 🖉 🖪	Cesssing news text by apposed and the exists ('model/yee) and the exists ('model/yee) and ('model/yee) and ('model/yee) and ('model/yee) and ('model/yee) and ('model) and ('	eckpoint: 2 hor Widgets	Anticions to tal Anticions to tal Urs ago (ar Help	a techniqu Hoickloar Control Control Text Data Aug utosaved) Recall	e (INIC)	× sions - Plea	₩ + se note that u	pdating to No	30°C Pa A [®] ☆ tebook 7 mi	rtly sunny C ght break	y ∧ I Ç≞ some o Pytl	وَ مِعَادَ اللَّهِ اللَّهُ اللَّ	}))) ENG 	16:2 06-10-	24 2024
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/not ation plan to Noteb Upyter Aug Edit View + (2) (2) (5) 0	speciessing news text by apposed and the starts of the sta	eLying aug (voctor, npy'): (voctor, npy'): ert WOTC s - Jupyter Noto nb ures and the a eckpoint: 2 hou Widgets Midgets Code Accuracy s 88.161209	entation Constructions to tal urs ago (au Help Precision 88.175883	a techniqu a ciclor A ci	e bun) I Rentation using extens FSCORE 88.083788	× sions - Plea	₩] + Ise note that u	pdating to No	30°C Pa A ^N ☆ tebook 7 mi	rtly sunn C ght break usted	y ∧ I ∱≣ Some o	ē ਵ : € f your e hon 3 (ip) <i>ffi</i> ⊄ ⊥ Log))) ENG − Cont s out 1) ●	16:2 06-10-	24 2024 > more
Type here t	In [141]: #pr if it o search Page - Select or creat ocalhost:8888/note ration plan to Noteb Upyter Aug Edit View + 중 @ 1	Decessing news text by apposed and the start of the starts	eckpoint: 2 hor Widgets Code 8 Accuracy 5 8 8.161209 6 82.871537	entation Plant of the second	A techniqu A tech	e mentation using extens ESCORE 88.083788 82.766457	× sions - Plea	w] + se note that u	pdating to No	30°C Pa A ^N ☆ tebook 7 mi	rtly sunn CD ght break	y ∧ Some o	ලි දා : ම f your වේ. hon 3 (ip) <i>(ii</i> , < ↓ Log	→ ENG	16:2 06-10- 0 ,	24 2024
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/not calion plan to Noteb Upyter Aug Edit View + [%]] [] 0 1 2	Constant and the new feat Constant and the new	eckpoint: 2 hou Widgets Code Accuracy 8 8.8649874	entation Participations to tal urs ago (au Help Precision 88.175883 82.86543 86.697504	A techniqu	Figure 1 Image: Second seco	× sions - Plea	₩ + use note that u	pdating to No	30°C Pa A ^N ☆ tebook 7 mi	rtly sunn ght break	y ∧ Some o	ତ 또 : 대한 아이 3 (ip	ع رز در بل	→ ENG	16:2 06-10- 0'	24 2024
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/not ation plan to Noteb Upyter Aug Edit View + (2) (1) 0 1 2 3 4	Cesssing news text by apposed and the exists ("model/vectors on load/"model/vectors on load/"model vectors on loa	eckpoint: 2 hor Widgets Widget	entation () 110	A technique A dickloart Construction Construction Aug Construction	e buo) () mentation using extens using extens 88.083788 82.766457 86.644451 88.646546 88.646546	× sions - Plea	₩ + Ise note that u	pdating to No	30°C Pa A [™] ☆ tebook 7 mi	rtly sunn, cl ght break	y ∧ I ⊊≡ Some o	ତ 또 : (급 f your (ip hon 3 (ip	¢ رو در	 Don't s out 	16:2 06-10-	24 2024
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/not ation plan to Noteb Upyter Aug Edit View + 😹 🖓 🗈	Cesssing news text by apposed and the start of the starts	eckpoint: 2 hor widgets widget	entation ()))))))))))))))))))	a techniqu h dickloar Control of the second a techniqu D Vector Text Data Aug a text	e buo) mentation using extens second se	× sions - Plea	₩ + se note that u	pdating to No	30°C Pa A [®] ☆ tebook 7 mi	rtly sunn C ght break	y ∧ I Ç≞ some o	€ ¶ E e P our C C C C C C C C C C C C C C C C C C C	ک پ Log	 ⇒) ENG — Q Don't s out i) ● 	16:2 06-10-	24 2024
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/not ation plan to Noteb Upyter Aug Edit View + (2) (0) 1 2 3 4 5 6	Cesssing news text by apples, path, exists ('model/yee, years) Sectors on load('model/yee, years) FIG 3. to convert the and X X AugmentFakeNews ebooks/AugmentFakeNews Last Ch Insert Cell Kernel SVM Original Corput SVM FWD Corput SVM FWD Corput SVM FWD Corput SVM FWD Corput BernoulliNB Original Corput Corput Description Descript	blying augment ctor.npy'): (ucctor.npy'): ctor.npy'): ctor.npy'): ctor.npy'): ctor.npy'): ctor.npy': ctor.npy: ctor.npy: ctor.npy:	entation ())))) ())))) ()))) ()))) ())))) ()))))) ()	a techniqu h dickloar Text Data Aug ike if you are utosaved) Recall 88.020143 82.707792 86.645348 88.844662 72.550994 70.274678 66.423024	e e e e e e e e e e e e e e e e e e e	× sions - Plea	₩ + ise note that u	pdating to No	30°C Pa A [®] ☆ tebook 7 mi	rtly sunny c ght break usted	y ∧ I ≨= Pytł	ē € : € fyour e hon 3 (ip) <i>(k</i> , ζ	Don't state	16:2 06-10-	24 2024
Type here t	In [141]: #pr if it o search Page - Select or creat ocalhost:8888/notu ation plan to Noteb Upyter Aug Edit View + (2) (2) (0) 1 2 3 4 5 6 7	Cesssing news text by apples, path, exists ('model/yee) FIG 3. to conversion FIG 4. t	Public Public eff Word eff Word eff Word nb ures and the a ures and the a widgets widgets Code a Accuracy s 88.161209 s 88.649874 s 86.649874 s 86.649874 s 86.649874 s 86.649874 s 71.788413 s 66.498741 s 74.307305	entation () () () () () () () () () ()	techniqu techniqu <t< td=""><td>e buo) mentation using extens extenses</td><td>× sions - Plea</td><td>₩ + ise note that u</td><td>pdating to No</td><td>30°C Pa A^N ☆ tebook 7 mi</td><td>rtly sunn CD ght break</td><td>y ∧ some o Pytl</td><td>Ĝ ♥ I I I I I I I I I I I I I I I I I I I</td><td>) <i>(ii</i>, < ↓ Log</td><td>Don't s</td><td>16:2 06-10-</td><td>24 2024</td></t<>	e buo) mentation using extens extenses	× sions - Plea	₩ + ise note that u	pdating to No	30°C Pa A ^N ☆ tebook 7 mi	rtly sunn CD ght break	y ∧ some o Pytl	Ĝ ♥ I I I I I I I I I I I I I I I I I I I) <i>(ii</i> , < ↓ Log	Don't s	16:2 06-10-	24 2024
Type here t	In [141]: #pr if if it o search Page - Select or creat ocalhost:8888/not ration plan to Noteb Upyter Aug Edit View + * * * * 1 0 1 2 3 4 5 6 7 8	Cesssing news text by apposed and set of the set of th	Public Public ert WORC ert WORC s - Jupyter Note Note nb ures and the a ures and the a Widgets > Code > Accuracy > 88.161209 > 82.871537 > 86.649874 > 73.551637 > 66.498741 > 74.307305 > 80.100756	entation ())))))))))))))))))	a techniqu a techniqu a techniqu b techniqu b techniqu construction constructin constructin <td>e buo) mentation using extens value ESCORE 88.083788 82.766457 86.644451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 72.489688 69.301298 63.130111 74.299315 79.896673</td> <td>× sions - Plea</td> <td>₩ + Ise note that u</td> <td>pdating to No</td> <td>30°C Pa A[™] ☆ tebook 7 mi</td> <td>rtly sunn ght break</td> <td>y ∧ Some o</td> <td>ତ আ : @</td> <td>Log</td> <td> Image: second se</td> <td>16:2 06-10-</td> <td>24 2024</td>	e buo) mentation using extens value ESCORE 88.083788 82.766457 86.644451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 88.645451 72.489688 69.301298 63.130111 74.299315 79.896673	× sions - Plea	₩ + Ise note that u	pdating to No	30°C Pa A [™] ☆ tebook 7 mi	rtly sunn ght break	y ∧ Some o	ତ আ : @	Log	 Image: second se	16:2 06-10-	24 2024
Type here t	In [141]: #pr if to search Page - Select or creat ocalhost:8888/not ation plan to Noteb Upyter Aug Edit View + 2 2 5 0 1 2 3 4 5 6 7 8 9	Cesssing news text by apposed and the starts of the start	Public Public Image: Construction of the second of t	entation () 110	techniqu diclorer techniqu diclorer techniqu Text Data Aug Text Data Aug techniqu techniqu techniqu text Data Aug text DataAu	e buo) mentation using extens using extens 8.083788 82.766457 86.644451 88.646546 72.489688 69.301298 63.130111 74.299315 79.896673 80.022967 80.022967	× sions - Plea	₩ + ise note that u	pdating to No	A ^N ☆ tebook 7 mi	rtly sunn C ght break	y ∧ I Ç= Some o	ତ আ 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이	¢ رو در	 Don't s Out 	16:2 06-10-	24 2024
Type here t	In [141]: #pr if if if if if if if if if if if if if	Cesssing news text by apposed and the start of the starts	elying augm ctor, npy'): (vector, npy'): ert WOTC s - Jupyter Not- nb ures and the a eckpoint: 2 hor Widgets Code Accuracy 8.8.161209 8.8.84131 5.86.649874 5.88.6649874 5.88.6649874 5.88.6649874 5.80.100756 5.80.100756 5.80.100756 5.80.100756 5.80.910756 5	entation allow	a techniqu h dickloar Control of the second a techniqu D Vector Text Data Aug a techniqu a techniqu D Vector a techniqu a techni	e mentation) using extens using extens second sec	× sions - Plea	₩I + Ise note that u	pdating to No	30°C Pa A [®] ☆ tebook 7 mi	rtly sunny ght break	y ∧ I Ç≞ Some o	€ € f your €) اللہ ح Log	⇒) ENG Don't s out	16:2 06-10-	24 2024
Type here t	In [141]: #pr if if it o search Page - Select or creat ocalhost:8888/not ation plan to Noteb Upyter Aug Edit View + (2) (2) (1) 2 3 4 5 6 7 8 9 10 11 12	Cesssing news text by appose, path.exists('model/yee, years) FIG 3. to conversion of the second of	augment augment ctor, npy'): ctor, npy' ctor, npy' c	entation ())))) ())))) ())))) ()))))) ()	Recall 88.020143 88.020143 88.020143 88.020143 88.4662 72.55094 70.84346 79.784549 79.860368 88.911730 77.256838 87.243753	e mentation using extens vsing extens vsing extens vsing	× sions - Plea	₩ + ise note that u	pdating to No	30°C Pa A ^N ☆ tebook 7 mi	rtly sunny ght break	y ∧ Some o	Ĝ ♥) <i>(ii</i> , < ↓ Log	Don't s	16:2 06-10-	24 2024
Type here t	In [141]: #pr if if if if if if if if if if if if if	Cesssing news text by appose, path.exists ('model/yee) FIG 3. to conversion FIG 4. to	blying aug ctor, npy'): (ucctor, npy'): ctor, npy', npy' ctor, npy', npy', npy' ctor, npy', npy', npy' ctor, npy', npy', npy', npy' ctor, npy', npy', npy', npy' ctor, npy', npy', npy', npy', npy', npy' ctor, npy', npy	entation allowing allowing actions to tal actions to tal urs ago (au Help Precision 88.175883 82.863543 86.697504 88.639594 75.096476 77.155143 75.758299 74.996148 80.184755 80.904536 88.982965 77.277718 87.432730 79.848998	Recall 88.020143 88.020143 82.707792 86.645348 88.44662 72.550994 70.824549 79.860368 89.11730 77.256838 87.243753 79.213168	e buo) mentation using extens using extens extens 88.083788 82.766457 88.644451 88.645461 72.489688 69.301298 63.130111 74.299315 79.896673 80.022967 88.911178 76.826049 87.317915 79.343239	x sions - Plea	₩ + ise note that u	pdating to No	30°C Pa A ^N ☆ tebook 7 mi	rtly sunn ght break	y ∧ Some o	G 🖬	Log	 >)) ENG − Quit 0) 	16:2 06-10-	24 2024

Random Forest BT Corpus 75.818640 76.432525 76.317871 75.814804

01 607657 01 702105 01 520557 01 626220

w

Þ

15

 ${\cal P}$ Type here to search

H

16 Evt

5

ISSN 2454-9940



www.ijasem.org

Vol 19, Issue 2, 2025



ISSN 2454-9940

Gasem

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

www.ijasem.org

Vol 19, Issue 2, 2025

	🗖 🔶 Home Pag	ge - Select or create a no 🗙 🛛 🔀 AugmentFakeNews - Jupyter Noto 🗴 🕒 Text Data Augmentation	×	+						-	Ō	×
\leftarrow	C () 127	.0.0.1:5000/PredictAction				C)	۲j≡	Ē	\downarrow	~		0
+	C () 127	.0.0.1:5000/PredictAction Test News = RIO DE JANEIRO/SAO PAULO (Reuters) - Billionaire Marcelo Odebrecht, the highest-profile executive imprisoned in Brazil s massive graft scandal, was released from jail on Tuesday to continue his sentence for corruption under house arrest, according to a federal court. The former chief executive officer of Odebrecht SA [ODBES.UL], Latin America s largest construction firm, was arrested in 2015 during an investigation dubbed Car Wash that exposed billions of dollars in kickbacks to politicians and executives at state-run companies in exchange for inflated contracts. Odebrecht was set to travel to Sao Paulo to begin his house arrest under electronic surveillance on Tuesday, according to the federal court in Parana. A representative for the former executive suid he remained committed to collaborating with authorities under a leniency deal. Odebrecht was first sentenced to 19 years in prison in one of the many cases related to Car Wash. That was reduced to 10 years after be signed a leniency deal last December in exchange				¢	ć	ه ا	Ŧ	~		
		for paying a nearly \$2 billion fine, admitting guilt and providing evidence to authorities. He has already served two-and-a-half years in prison. Under the deal, he must serve another two-and-a-half years under house arrest. He will then be permitted to leave his home for work for another two-and-a-half years. He will then be required to do community service for the rest of his 10-year sentence. Separately Tuesday, Brazil s antitrust watchdog Cade said it was investigating two alleged cartels involved in bidding for Sao paulo infrastructure projects after receiving information provided by Odebrecht executives. News Predicted As ===> Fake								15:46		
-		search 🛛 👸 💽 📃 🗉 🖻 🧕 🔼 🖻 📆	b	💶 🤠 Brea	iking ne	ws ^	ê e) <i>(i</i> . ¢))) ENG	16:46 06-10-2	5 2024	1

FIG.7 Results obtained by Extention XGboost

IV. CONCLUSION

Results show that text data augmentation helps classify fake news better. With the use of Function Word Reduction, Synonym Replacement, and Back Translation, the classification accuracy and datasets were greatly enhanced.

The best algorithms for Back Translation-augmented text were SVM and Bernoulli Naïve Bayes, among all those that were tested. In contrast, Logistic Regression demonstrated the impact of augmentation strategies on classifier performance by outperforming other methods while using Reduction of Function Words. When testing on the original corpus without any augmentation, Random Forest yielded the best results. The most accurate, though, was XGBoost, an ensemble algorithm that boosts predictive strength by combining decision trees.

This approach demonstrates that textual data can improve classification accuracy in limited datasets. The accuracy of the system's fake news identification is enhanced by advanced technologies such as XGBoost.

FUTURE SCOPE

Utilizing Contextualized Word Embeddings and Word Embedding Averaging can enhance this project.

Transformers and BERT are two examples of advanced deep learning models that could improve classification accuracy.

Potentially more effective methods include hybrid models that use machine learning techniques and ensemble procedures.

To further aid the algorithm in detecting fake news in many languages, multi-lingual data augmentation can be considered.



REFERENCES

[1] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," 2021, arXiv:2105.03075.

[2] L.F.A.O.Pellicer, T.M.Ferreira, and A.H.R.Costa, "Dataaugmentation techniques in natural language processing," Appl. Soft Comput., vol. 132, Jan. 2023, Art. no. 109803, doi: 10.1016/j.asoc.2022.109803.
[3] S. Nazir, M. Asif, S. A. Sahi, S. Ahmad, Y. Y. Ghadi, and M. H. Aziz, "Toward the development of large-scale word embedding for low resourced language," IEEE Access, vol. 10, pp. 54091–54097, 2022, doi: 10.1109/ACCESS.2022.3173259.

[4] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet based approaches," in Proc. Human Lang. Technologies, Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2009, pp. 19–27.

[5] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating semantic models with (Genuine) similarity estimation," Comput. Linguistics, vol. 41, no. 4, pp. 665–695, Dec. 2015, doi: 10.1162/coli_a_00237.

[6] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 6381–6387, doi: 10.18653/v1/d19-1670.

[7] G. A. Miller, "WordNet," Commun. ACM, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[8] V. Marivate and T. Sefara, "Improving short text classification through global augmentation methods," in Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction, 2020, pp. 385–399, doi: 10.1007/978-3030-57321-8_21.

[9] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2018, pp. 452–457, doi: 10.18653/v1/n18-2072.

[10] R. N. Al-Matham and H. S. Al-Khalifa, "SynoExtractor: A novel pipeline for Arabic synonymextraction using Word2 Vecwordembeddings," Complexity, vol. 2021, pp. 1–13, Feb. 2021, doi: 10.1155/2021/6627434.

[11] I. Salah, K. Jouini, and O. Korbaa, "On the use of text augmentation for stance and fake news detection," J. Inf. Telecommun., vol. 7, no. 3, pp. 359–375, Jul. 2023, doi: 10.1080/24751839.2023.2198820.

[12] I. Salah, K. Jouini, and O. Korbaa, "Augmentation-based ensemble learning for stance and fake news detection," in Proc. Int. Conf. Comput. Collective Intell., 2022, pp. 29–41, doi: 10.1007/978-3-031-16210-7_3.

[13] M. Bucos and G. Țucudean, "Text data augmentation techniques for fake news detection in the Romanian language," Appl. Sci., vol. 13, no. 13, p. 7389, Jun. 2023, doi: 10.3390/app13137389.

[14] A.J.Keya, M.A.H.Wadud, M.F.Mridha, M.Alatiyyah, and M.A.Hamid, "AugFake-BERT:Handling imbalance through augmentation of fake news using BERT to enhance the performance of fake news classification," Appl. Sci., vol. 12, no. 17, p. 8398, Aug. 2022, doi: 10.3390/app12178398.

[15] G. Haralabopoulos, M. T. Torres, I. Anagnostopoulos, and D. McAuley, "Text data augmentations: Permutation, antonyms and negation," Expert Syst. Appl., vol. 177, Sep. 2021, Art. no. 114769, doi: 10.1016/j.eswa.2021.114769.

[16] A. Dahou, A. A. Ewees, F. A. Hashim, M. A. A. Al-Qaness, D. A. Orabi, E. M. Soliman, E. M. Tag-Eldin, A. O. Aseeri, and M. A. Elaziz, "Optimizing fake news detection for Arabic context: A multitask learning approach with transformers and an enhanced nutcracker optimization algorithm," Knowl.-Based Syst., vol. 280, Nov. 2023, Art. no. 111023, doi: 10.1016/j.knosys.2023.111023.



[17] J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, "Multimodal fake news detection through data augmentation-based contrastive learning," Appl. Soft Comput., vol. 136, Mar. 2023, Art. no. 110125, doi: 10.1016/j.asoc.2023.110125.

[18] M. I. Marwat, J. A. Khan, M. D. Alshehri, "Sentiment analysis of product reviews to identify deceptive rating information in social media: ASentiDeceptive approach," KSII Trans. Internet Inf. Syst., vol. 16, no. 3, pp. 830–860, Dec. 2022, doi: 10.3837/tiis.2022.03.005.

[19] J. A. Khan, A. Yasin, R. Fatima, D. Vasan, A. A. Khan, and A. W. Khan, "Valuating requirements arguments in the online user's forum for requirements decision-making: The CrowdRE-VArg framework," Softw., Pract. Exper., vol. 52, no. 12, pp. 2537–2573, Dec. 2022, doi: 10.1002/spe.3137.

[20] M. Risdal. (2016). Getting Real About Fake News. Kaggle. Accessed: Dec. 28, 2023. [Online]. Available: https://www.kaggle.com/code/ anthonyc1/gathering-real-news-for-oct-dec-2016/output 31550 [21] S.Bird,E.Klein,andE.Loper,NaturalLanguageProcessingWithPython: Analyzing Text With the Natural Language Toolkit, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2009.

[22] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word embedding over linguistic features for fake news detection," IEEE Trans. Computat. Social Syst., vol. 8, no. 4, pp. 881–893, Aug. 2021. [Online].

Available: https://www.kaggle.com/datasets/saurabhshahane/fake-newsclassification/data

[23] J. Tiedemann and S. Thottingal, "OPUS-MT—Building open translation services for the world," in Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl., A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, M. L. Forcada, Eds., 2020, pp. 479–480.

[24] R. Řehuřek and P. Sojka, "Software framework for topic modelling with large corpora," in Proc. LREC Workshop New Challenges for NLP Frameworks, ELRA, 2010, pp. 45–50.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, arXiv:1301.3781.

[26] M. Zhai, J. Tan, and J. Choi, "Intrinsic and extrinsic evaluations of word embeddings," in Proc. AAAI Conf. Artif. Intell., Nov. 2016, vol. 30, no. 1, pp. 4282–4283, doi: 10.1609/aaai.v30i1.9959.

[27] Y. Shi, Y. Zheng, K. Guo, L. Zhu, and Y. Qu, "Intrinsic or extrinsic evaluation: An overview of word embedding evaluation," in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Nov. 2018, pp. 1255–1262,

doi: 10.1109/ICDMW.2018.00179.