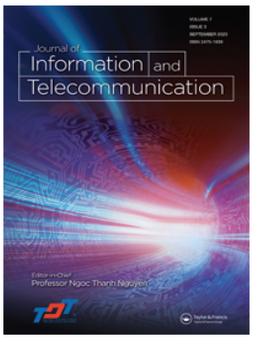

Taylor & Francis
Taylor & Francis Group

Journal of Information and Telecommunication

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjit20

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

Ilhem Salah, Khaled Jouini & Ouajdi Korbaa

To cite this article: Ilhem Salah, Khaled Jouini & Ouajdi Korbaa (2023) On the use of text augmentation for stance and fake news detection, Journal of Information and Telecommunication, 7:3, 359-375, DOI: 10.1080/24751839.2023.2198820

To link to this article: https://doi.org/10.1080/24751839.2023.2198820

Published online: 19 Apr 2023.

JOURNAL OF INFORMATION AND TELECOMMUNICATION
2023, VOL. 7, NO. 3, 359–375
https://doi.org/10.1080/24751839.2023.2198820

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

Ilhem Salah, Khaled Jouini and Ouajdi Korbaa

MARS Research Lab LR17ES05, ISITCom, University of Sousse, H. Sousse, Tunisia

ABSTRACT

Data Augmentation (DA) aims at synthesizing new training instances by applying transformations to available ones. DA has several well-known benefits such as: (i) increasing generalization ability; (ii) preventing data scarcity; and (iii) helping resolve class imbalance issues. In this work, we investigate the use of DA for stance and fake news detection. In the first part of our work, we explore the effect of various DA techniques on the performance of common classification algorithms. Our study reveals that the motto '*the more, the better*' is the wrong approach regarding text augmentation and that there is no *one-size-fits-all* text augmentation technique. The second part of our work leverages the results of our study to propose a novel augmentation-based, ensemble learning approach. The proposed approach leverages text augmentation to enhance base learners' diversity and accuracy, ergo the predictive performance of the ensemble. The third part of our work experimentally investigates the use of DA to cope with the class imbalance problem. Class imbalance is very common in stance and fake news detection and often results in biased models. In this work we show how and to what extent text augmentation can help resolving moderate and severe imbalance.

ARTICLE HISTORY
Received 30 December 2022
Accepted 26 March 2023

KEYWORDS
Stance and fake news detection; text augmentation; ensemble learning; class imbalance

1. Introduction

In the era of the Internet and social media, where a myriad of information of various types is instantly available and where any point of view can find an audience, access to information is no longer an issue, and the key challenges are veracity, credibility, and authenticity. The reason for this is that any user can readily gather, consume, and break news, without verification, fact-checking, or third-party filtering.

By directly influencing public opinions, major political events, and societal debates, fake news has become the scourge of the digital era, and combating it has become a dire need. The identification of fake news is however very challenging, not only from a machine learning and Natural Language Processing (NLP) perspective, but also

CONTACT Ilhem Salah ilhemsalah53@gmail.com MARS Research Lab LR17ES05, ISITCom, University of Sousse, H. Sousse 4011, Tunisia

sometimes for the most experienced journalists (Pomerleau & Rao, 2017). That is why the scientific community approaches the task from a variety of angles and often breaks down the process into independent sub-tasks. A first practical step towards automatic fact-checking and fake news detection is to estimate the opinion or the point of view (i.e. *stance*) of different news sources regarding the same topic or claim (Pomerleau & Rao, 2017). This (sub-) task, addressed in recent research as *stance detection*, was popularized by the Fake News Challenge – Stage 1 (or FNC-1) (Pomerleau & Rao, 2017), which compares article bodies to article headlines and determines if a body agrees, disagrees, discusses or is unrelated to the claim of a headline. As aptly stated by Momchil et al. (2022), automated stance detection can help in identifying fake news in two key ways. First, it enables human fact-checkers to quickly and efficiently, identify controversial claims, gather relevant opinions about a claim, and evaluate the arguments for and against it (i.e. *evidence retrieval*). Second, it can be integrated as a component of an automated fact-checking pipeline, which would give a preliminary label to a claim, based on the stances taken by various sources, weighted by their credibility (Guo et al., 2021).

In a previous work (Salah et al., 2022), we proposed a novel *Augmentation-based Ensemble learning* approach for stance and fake news detection. Data augmentation aims at synthesizing new training instances that have the same ground-truth labels as the instances that they originate from (Xie et al., 2019). Data augmentation has several well-known benefits: (i) preventing overfitting by *improving the diversity* of training data; (ii) *preventing data scarcity* by providing a relatively easy and inexpensive way to collect and label data; (iii) increasing the *generalization ability* of the obtained models; and (iv) helping resolve *class imbalance* issues. Data augmentation is extensively used in Computer Vision (CV) where it is considered as one of the anchors of good predictive performance. Despite promising advances, data augmentation remains however less explored in NLP where it is still considered as the 'cherry on the cake' which provides a steady but limited performance boost (Shorten et al., 2021).

Ensemble learning combines the knowledge acquired by base learners to make a consensus decision which is supposed to be superior to the one attained by each base learner alone (Suting & Ning, 2020). Research on ensemble learning proves that the greater are the skills and the diversity of base learners, the better are the accuracy and the generalization ability of the ensemble (Suting & Ning, 2020). In our work we leverage text augmentation to enhance both, the diversity and the skills of base learners, ergo the predictive performance of the ensemble.

Class imbalance refers to situations where the distribution of examples across the classes is not equal, i.e. the number of examples available for one or more classes (minority classes) is far less than other classes (majority classes). Class imbalance appears in many domains, including fraud detection, disease screening and fake news detection. When a dataset is imbalanced, most classifiers have *a strong bias toward majority classes* (Fernndez et al., 2018). This paper extends our previous work (Salah et al., 2022) by presenting more comprehensive experimental results, additional related work, and deeper insights into our augmentation-based ensemble approach. Furthermore, this paper evaluates the effectiveness of text augmentation in addressing severe and moderate class imbalance, a common issue in stance and fake news detection, not explored in (Salah et al., 2022).

The main contributions of our work are therefore: (i) an extensive experimental study on the effect of different text data augmentation techniques on the performance of common classification algorithms; (ii) a novel augmentation-based ensemble learning approach; and (iii) an experimental study on the use of text augmentation to mitigate the effects of class imbalance.

The remainder of this paper is organized as follows. Section 2 outlines the main steps we followed to vectorize text and reduce dimensionality. Section 3 exposes the key motifs of data augmentation and the text augmentation techniques adopted in our work. Section 4 details the architecture of our novel augmentation-based ensemble learning. Section 5 briefly reviews existing work on stance and fake news detection. Section 6 presents an experimental study on two real-world fake news datasets and discusses the main results and findings. Finally, Section 7 concludes the paper.

2. Text as vectors

2.1. Pre-processing and feature extraction

Machine Learning (ML) algorithms operate on numerical features, expecting input in the form of a matrix where rows represent instances and columns features. Raw texts have therefore to be transformed into feature vectors before feeding into ML algorithms (Jouini et al., 2021). In our work, we first eliminated stop words and reduced words to their roots (i.e. base words) by stemming them using Snowball Stemmer from the NLTK library (NLTK.org., n.d.). We next vectorized the corpus with a TF-IDF (*Term Frequency – Inverse Document Frequency*) weighting scheme and generated a term-document matrix.

TF–IDF is computed on a per-term basis, such that the relevance of a term to a text is measured by the scaled frequency of the appearance of the term in the text, normalized by the inverse of the scaled frequency of the term in the entire corpus. Despite its simplicity and its wide-spread use, the TF-IDF scheme has two severe limitations: (i) TF-IDF does not capture the co-occurrence of terms in the corpus and makes no use of semantic similarities between words. Accordingly, TF-IDF fails to capture some basic linguistic notions such as synonymy and homonymy; and (ii) The term-document matrix is high dimensional and is often noisy, redundant, and excessively sparse. The matrix is thus subject to the curse of dimensionality: as the number of features is large, poor generalization is to be expected.

2.2. Dimensionality reduction

Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is an unsupervised statistical topic modelling technique, overcoming some of the limitations of TF-IDF. As other topic modelling techniques, such as LDA (Latent Dirichlet Allocation (Blei et al., 2003)), LSA is based on the assumptions that: (i) each text consists of a mixture of *topics*; and (ii) each topic consists of a set of (weighted) terms that regularly co-occur together. Put differently, the basic assumption behind LSA is that words that are close in meaning, appear in similar contexts and form a 'hidden topic'. The idea behind LSA is then to represent words that form a topic not as separate dimensions, but by a single dimension. LSA

represents thus texts by 'semantic' or 'topic' vectors, based on the words that these texts contain and the set of weighted words that form each of the topics.

To uncover the latent topics that shapes the meaning of texts, LSA performs a Singular Value Decomposition (SVD) on the document-term matrix (i.e. decomposes it into a separate text-topic matrix and a topic-term matrix). Formally, SVD decomposes the term-document matrix $A_{t\times n}$, with t the number terms and d the number of documents, into the product of three different matrices: orthogonal column matrix, orthogonal row matrix and one singular matrix.

$$A_{t\times n} = U_{t\times n} S_{n\times n} D^{T}_{n\times d} \tag{1}$$

where $n = \min(t, d)$ is the rank of A. By restricting the matrices T, S and D to their first $k<n$ rows, we obtain the matrices $T_{t\times k}$, $S_{k\times k}$ and $D_{d\times k}$, and hence obtain k-dimensional text vectors. From a practical perspective, the key ask is to determine k, which would be reasonable for the problem (i.e. without major loss). In our work we used the transformer TruncatedSVD from sklearn (Pedregosa et al., 2011). As in Li et al. (2019) we set the value of k to 100D. The experimental study conducted in Li et al. (2019) showed that using LSA (with k set to 100D) instead of TF-IDF allows a substantial performance improvement for the tasks of stance and fake news detection.

3. Text data augmentation

The success of data augmentation in Computer Vision has been fuelled by the ease of designing semantically invariant transformations (i.e. label-preserving transformations), such as rotation, flipping, etc…While recent years witnessed significant advancements in the design of transformation techniques, text augmentation remains less explored and adopted in NLP than in CV. This is mainly due to the intrinsic properties of textual data (e.g. polysemy), which make defining label-preserving transformations much harder (Shorten et al., 2021). In the sequel we mainly focus on off-the-shelf text augmentation techniques and less on techniques that are still in the research phase, waiting for large-scale testing and adoption. For a more exhaustive survey on text augmentation techniques, we refer the reader to Karnyoto et al. (2022), Li et al. (2021), and Tesfagergish et al. (2021).

3.1. Masked language models

The main idea behind *Masked Language Models* (MLMs), such as BERT (Devlin et al., 2018), is to mask words in sentences and let the model predict the masked words. BERT, which is a pretrained multi-layer bidirectional transformer encoder, has the ability to predict masked words based on the bidirectional context (i.e. based on its left and right surrounding words). In contrast with other context-free models such as GLOVE and Word2Vec, BERT alleviates the problem of ambiguity since it considers the whole context of a word.

BERT is considered as a breakthrough in the use of ML for NLP and is widely used in a variety of tasks such as classification, Question/Answering, and Named Entity Recognition (Shi et al., 2020). Inspired by the recent work of Li et al. (2021) and Shi et al. (2020), we use BERT as an augmentation technique. The idea is to generate new sentences by randomly masking words and replacing them by those predicted by BERT.

3.2. Back-translation (*a.k.a.* round-trip translation)

Back-Translation is the process of translating a text into another language, then translating the new text back into the original language. Back-translation is one of the most popular means of paraphrasing and text augmentation (Marivate & Sefara, 2019). Google Cloud Translation API, used in our work to translate sentences to French and back, is considered as the most common tool for back-translation (Li et al., 2021).

3.3. Synonym (*a.k.a.* thesaurus-based augmentation)

The *Synonym* technique, also called lexical substitution with dictionary, was until recently the most widely (and for a long time the only) augmentation technique used for textual data classification. As suggested by its name, the Synonym technique replaces randomly selected words with their respective synonyms. The types of words that are candidates for lexical substitution are: adverbs, adjectives, nouns and verbs.

The synonyms are typically taken from a lexical database (i.e. dictionary of synonyms). WordNet (Shoemaker, 2019), used in our work for synonym replacement, is considered as the most popular open-source lexical database for the English language (Li et al., 2021).

3.4. TF-IDF based insertion and substitution

The intuition behind these two noising-based techniques is that uninformative words (i.e. words having low TF-IDF scores) should have no or little impact on classification. Therefore, the insertion of words having low TF-IDF scores (at random positions) should preserve the label associated with a text, even if the semantics are not preserved. An alternate strategy is to replace randomly selected words with words having the same low TF-IDF scores (*TF-IDF based substitution*).

Section 6 presents an extensive study on the effect of the aforementioned augmentation techniques on the predictive performance of ten common classification algorithms, namely, Decision Tree (DT), Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Bagged Random Forests (Bagged RF), Light Gradient Boosting Machine (LightGBM), Gradient Boosting (GradBoost), Logistic Regression (LR), and Naive Bayes (NB).

4. Augmentation-based ensemble learning

4.1. Diversity and skilfulness in ensemble learning

Ensemble Learning finds its origins in the 'Wisdom of Crowds' theory (Surowiecki, 2005). The 'Wisdom of Crowds' theory states that the collective opinion of a group of individuals can be better than the opinion of a single expert, provided that the aggregated opinions are diverse (i.e. diversity of opinion) and that each individual in the group has a minimum level of competence (e.g. better than a random guess). Similarly, Ensemble Learning combines the knowledge acquired by a group of base learners to make a consensus decision which is supposed to be superior to the one reached by each of them separately (Suting & Ning, 2020). Research on Ensemble Learning proves that the greater are the skills and the diversity of base models, the better is the generalization ability of the ensemble model (Suting & Ning, 2020). Alternatively stated, to generate a good ensemble model, it is

necessary to build base models that are, not only skilful, but also skilful in a different way from one another.

Bagging and stacking are among the main classes of parallel ensemble techniques. *Bagging* (i.e. Bootstrap aggregating) involves training multiple instances of the same classification algorithm, then combining the predictions of the obtained models through hard or soft voting. To promote diversity, base learners are trained on different subsets of the original training set. Each subset is typically obtained by drawing random samples with replacement from the original training set (i.e. bootstrap samples). *Stacking* (*a.k.a*. stacked generalization) involves training a learning algorithm (i.e. meta-classifier) to combine the predictions of several heterogeneous learning algorithms, trained on the same training data. The most common approach to train the meta-model is via k-fold cross-validation. With the k-fold cross-validation, the whole training dataset is randomly split (without replacement) into independent equal-sized k-folds. $k-1$ folds are then used to train each of the base models and the kth fold (holdout fold) is used to collect the predictions of base models on unseen data. The predictions made by base models on the holdout fold, along with the expected class labels, provide the input and the output pairs used to train the meta-model. This procedure is repeated k times. Each time a different fold acts as the holdout fold while the remaining folds are combined and used for training the base models.

4.2. Novel augmentation-based approach

As mentioned earlier, in conventional stacking base learners are trained on the same dataset and diversity is achieved by using heterogeneous classification algorithms. As depicted in Figure 1, the classical approach for combining augmentation and stacking, is to: (i) apply one or several augmentation techniques to the original dataset, (ii) fuse the original dataset and data obtained through augmentation; and (iii) train base learners on the fused dataset. In our work we adopt a different approach and train heterogeneous algorithms on different data to further promote diversity. More specifically, through an extensive experimental study (Section 6), we first identify the most accurate *(augmentation technique, classification algorithm)* pairs. Our meta-model is then trained on the predictions

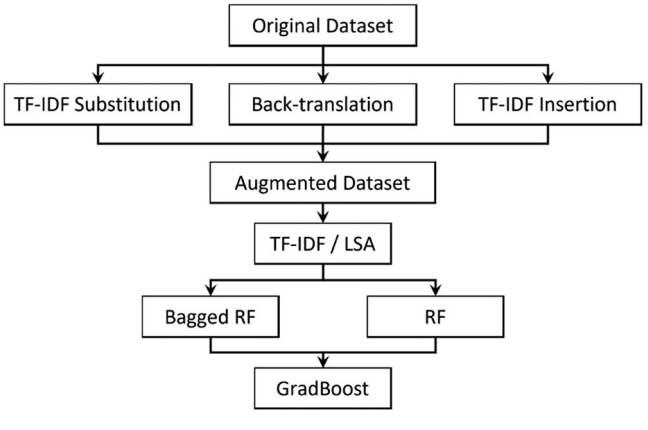

Figure 1. Conventional approach for combining augmentation and stacking.

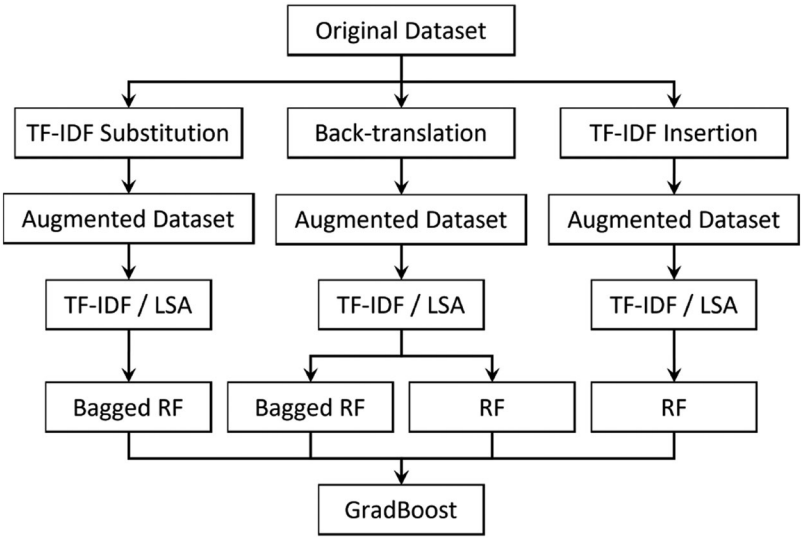

Figure 2. Novel augmentation-based ensemble learning approach.

made by the most accurate pairs, using a stratified k-fold cross-validation. Figure 2 depicts the overall architecture of the proposed augmentation-based ensemble learning.

Our augmentation-based ensemble learning approach, can be seen as a mixture between stacking and bagging. In contrast with Bagging and like Stacking, we use an ensemble of heterogeneous learning algorithms. In contrast with stacking and like Bagging, base learners are trained on different datasets. However, unlike Bagging the considered datasets are not obtained through bootstrap sampling. Instead, they are obtained by combining the original training data with the data obtained by applying one of the aforementioned text augmentation techniques. Finally, like in conventional Stacking, the meta-model is trained using a stratified K-fold cross-validation.

5. Related work

Salient stance and fake news detection approaches adopt a wide range of different features (e.g. context-based, content-based), classifiers, and learning tactics (e.g. stacking, bagging, etc.). Due to the lack of space, we mainly focus hereafter on ensemble approaches and on approaches that rely on content-based features. We suggest readers to refer to surveys and retrospectives on recent challenges (Hanselowski et al., 2018; Khan et al., 2021) for a more comprehensive overview of the current state of research. In the sequel, we distinguish between approaches dedicated to stance classification (*multinomial classification*) and those intended to fake news classification (*binary classification*).

5.1. Stance classification

The authors of the fake news challenge (FNC-1) (Slovikovskaya, 2019), released a simple baseline model for the stance detection task. The proposed model achieves an

F1-score of 79.53% and uses a gradient boosting (GradBoost) classifier on global co-occurrence, polarity and refutation features. The three best performing systems in the FNC-1 competition were 'SOLAT in the SWEN' (Pan, 2018), 'Team Athene' (Hanselowski et al., 2018) and 'UCL Machine Reading' (UCLMR) (Riedel et al., 2017). 'SOLAT in the SWEN' won the competition using an ensemble approach based on a 50/50 weighted average between gradient-boosted decision trees and a Convolutional Neural Network (CNN). The proposed system is based on several features: Word2Vec pretrained embeddings, TF-IDF, Single Value Decomposition and WordCount. The convolutional network uses pre-trained Word2Vec embeddings passed through several convolutional layers followed by three fully-connected layers and a final softmax layer for classification. Hanselowski et al. (2018), the second place winner, used an ensemble composed of 5 Multi-Layer Perceptrons (MLPs), where labels are predicted through hard voting. The system of UCLMR (Riedel et al., 2017), placed third, used an MLP classifier with one hidden layer of 100 units and a softmax layer for classification.

Recently, other published work used FNC-1 in their experiments.In particular, several recent approaches (Dulhanty et al., 2019; Sepúlveda-Torres et al., 2021; Slovikovskaya, 2019) construct stance detection language models by performing transfer learning on pre-trained variants of BERT (mainly BERT, RoBERTa and XLNet).

Among these approaches, Sepúlveda-Torres et al. (2021) stands out for its integration of *text summarization*. Text summarization involves reducing a long text into a shorter version, while preserving its most important information. From an overarching perspective, text summarization and text augmentation can both be seen as techniques that alter the form of a text (making it more concise or more diverse) to better capture its core meaning.

The approach proposed by Sepúlveda-Torres et al. (2021) involves two stages: *Relatedness Stage* and *Stance Stage*. The Relatedness Stage is in charge of determining whether or not a headline and a news summary are related. This stage uses TextRank extractive algorithm for body text summarization and a fine-tuned RoBERTa pre-trained model that classifies a headline-summary pair as related or unrelated. Once the related pairs are identified, the Stance stage determines their type with respect to the remaining stances (*agree*, *disagree* or *discuss*). Similarly to the Relatedness Stage, the Stance stage uses a fine-tuned RoBERTa pre-trained model to yield predictions. While not presented as such by the authors, we believe that by discarding unrelated pairs (the majority class) at an early stage of the process, the approach of Sepúlveda-Torres et al. (2021) partially resolves the class imbalance problem. The work of Sepúlveda-Torres et al. (2021) on text summarization and ours on text augmentation, both demonstrate that modifying the form of texts while preserving their respective meanings can result in improved predictive performance. A downside of the work of Sepúlveda-Torres et al. (2021) is that it is not suitable for short text messages, which are prevalent on social media.

5.2. Fake news classification

Besides stance detection, several ensemble learning models have been proposed to tackle the binary (*True News*/*False News*) content-based classification task. Notably, Jiang et al. (2021) proposed a stacking-based ensemble that uses Random Forest (RF) as meta-learner and Support Vector Machine (SVM), Logistic Regression (LR), Decision

Tree (DT), k-nearest neighbours (KNN), Random Forest (RF), Convolution Neural Network (CNN), Long short-term memory (LSTM) and Gated Recurrent Network (GRU) as base learners. The approach of Jiang et al. (2021) uses three different text vectorization methods: Word2Vec embedding, TF-IDF and TF. The proposed approach was evaluated on ISOT fake news (Ahmed et al., 2017) and KDnuggets (McIntire, 2017) datasets. Similarly, Patil (2022) proposed a majority voting ensemble model involving nine base learners, namely: SVM, DT, LR, RF, X-Gradient Boosting (XGBoost), Extra Trees (ET), AdaBoost, Stochastic Gradient Descent (SGD) and Naive Bayes (NB). The proposed approach was evaluated on Kaggle Fake News dataset (Kaggle.com, n.d.). In the same vein as Patil (2022) and Mahabub (2020) uses a majority voting classifier with an ensemble composed of three base learners, namely, MLP, LR and XGBoost. Mahabub (2020) experimented their approach on the dataset LIAR proposed by Wang (2017).

The work of Li et al. (2019), which is the closest to the spirit of our work, uses LSA for dimensionality reduction and a stacking-based ensemble having five base learners: GradBoost, Random Forest (RF), XGBoost, Bagging and Light Gradient Boosting Machine (Lightgbm). Besides, Li et al. (2019) compared LDA and LSA and found that LSA yields better accuracy. The authors in Li et al. (2019) experimented their approach on FNC-1 and FNN datasets. Li et al. (2019) has not addressed the issue of class imbalance. An experimental comparison between (Li et al., 2019) and our work is given in Section 6.

It is worth noticing that in all the aforementioned studies, ensemble approaches yielded better results than those attained by their contributing base learners. On the other hand, despite the substantial potential improvement that text augmentation can carry out, to the best of our knowledge, there exists no previous work on stance and fake news detection that compares text augmentation techniques, uses text augmentation in conjunction with ensemble learning or mitigates the effects of class imbalance through text augmentation.

6. Experimental study

6.1. Tools & datasets

Our system was implemented using NLTK (NLTK.org., n.d.) for text preprocessing, nlpaug (Ma, 2019) for text augmentation, SciKit-Learn (version 0.24.2) (Pedregosa et al., 2011) for classification and Beautiful Soup for web scraping. A stratified 10-fold cross-validation was used for model fusion. The Li & al. approach was implemented as described in Li et al. (2019). The experimental study was conducted without any special tuning. A large number of experiments have been performed to show the accuracy and the effectiveness of our augmentation-based ensemble learning. Due to the lack of space, only few results are presented herein.

As there are no agreed-upon benchmark datasets for stance and fake news detection (Li et al., 2019), we used two publicly available and complementary datasets: FNC-1 (Slovikovskaya, 2019) and FNN (i.e. FakeNewsNet) (Shu, 2019). FNC was released to explore the task of stance detection in the context of fake news detection. As reported in Table 1, the FNC-1 dataset consists of approximately 50k headline-article pairs in the training set and 25k pairs in the test set. Stance detection is a multinomial classification problem, where the relative stance of each headline-article pair has to be classified as

Table 1. Corpus statistics and class distribution in the FNC-1 dataset.

Dataset	# Instances of training data	# Agree	# Disagree	# Discuss	# Unrelated
FNC-1	49,972	7.36%	1.68%	17.82%	73.31%

Headline: Hundreds of Palestinians flee floods in Gaza as Israel opens dams	
Agree (AGR)	GAZA CITY (Ma'an) – Hundreds of Palestinians were evacuated from their homes Sunday morning after Israeli authorities opened a number of dams near the border, flooding the Gaza Valley in the wake of a recent severe winter storm. The Gaza Ministry of Interior said in a statement that civil defense services and teams from the Ministry of Public Works had evacuated more than 80 families from both sides of the Gaza Valley (Wadi Gaza) after their homes flooded as water levels reached more than three meters [..]
Discuss (DSC)	Palestinian officials say hundreds of Gazans were forced to evacuate after Israel opened the gates of several dams on the border with the Gaza Strip, and flooded at least 80 households. Israel has denied the claim as "entirely false". [..]
Disagree (DSG)	Israel has rejected allegations by government officials in the Gaza strip that authorities were responsible for released storm waters flooding parts of the besieged area. "The claim is entirely false, and southern Israel does not have any dams," said a statement from the Coordinator of Government Activities in the Territories (COGAT). "Due to the recent rain, streams were flooded throughout the region with no connection to actions taken by the State of Israel." At least 80 Palestinian families have been evacuated after water levels in the Gaza Valley (Wadi Gaza) rose to almost three meters. [..]
Unrelated (UNR)	Apple is continuing to experience 'Hairgate' problems but they may just be a publicity stunt [..]

Figure 3. Headline and text snippets with respective stances from the FNC dataset (Slovikovskaya, 2019).

either: *Agree* if the article agrees with the headline claim, *Disagree* if the article disagrees with the claim, *Discuss* if the article is related to the claim, but takes no position on the subject, and *Unrelated* if the content of the article is unrelated to the claim. An illustration of the above classification task is given in Figure 3.

It is worth mentioning that the discovery of a disagreeing headline-article pair does not necessarily correspond to the discovery of a fake article, but is an automated first step which could make human fact-checkers aware of a discrepancy (Momchil et al., 2022). Human fact-checkers or specialized algorithms can then ultimately decide which articles are fake, based on the credibility of agreeing and disagreeing sources.

FNN data was collected from two fact-checking websites (i.e. GossipCop and PolitiFact) containing news contents, along with context information. In comparison with FNN, FNC-1 provides fewer data features (4 vs. 13 features), but more data (≈ 75k vs. ≈ 16k) (Table 2).

6.2. Results and discussion

We ran our experiments with four objectives in mind: (i) identify the best performing (*Augmentation technique*, *Classifier*) pairs; (ii) quantify the actual performance improvement allowed by each text augmentation technique; (iii) evaluate the effectiveness of our augmentation-based ensemble approach; and (iv) evaluate the effectiveness of text augmentation in addressing class imbalance.

Table 2. Class distribution in the FNN dataset (Shu, 2019).

Dataset	# Instances	% True news	% Fake news
FNN	16,118	76.23%	23.77%

6.2.1. Best performing pairs

The two Tables 3 and 5 (resp. Figures 4 and 6), report the F1-scores and Accuracy obtained on FNC (resp. FNN). The results presented in these tables allow to draw important conclusions regarding text augmentation:

(1) *Text augmentation does not always improve predictive performance*. This can be especially observed for SVM, LightGBM, GradBoost (Tables 3 and 5) and AdaBoost

Table 3. F1-score on FNC without and with data augmentation.

Classification algorithm	Without data augmentation	With data augmentation					
		Synonym	BERT	Tf-IDF subtitution	TF-IDF insertion	Back-translation	Combination
DT	81.64%	75.51%	84.01%	80.59%	81.71%	**86.80%**	78.05%
SVM	**61.78%**	61.78%	61.78%	61.78%	61.78%	61.78%	61.78%
AdaBoost	61.71%	**62.08%**	61.86%	61.91%	61.82%	61.78%	61.83%
RF	85.47%	80.22%	88.27%	84.77%	85.55%	**89.93%**	82.95%
XGBoost	75.04%	71.67%	72.65%	71.91%	71.92%	73.91%	71.67%
Bagged RF	85.99%	80.54%	88.25%	85.04%	85.80%	**90.24%**	83.23%
LightGBM	**86.48%**	78.74%	84.07%	82.16%	82.61%	86.20%	80.14%
GradBoost	**72.98%**	70.89%	71.68%	70.97%	70.88%	72.75%	70.87%
LR	62.40%	62.65%	62.39%	62.45%	**62.67%**	62.50%	62.42%
NB	63.54%	62.56%	62.91%	62.71%	63.03%	**63.65%**	63.31%

Table 5. Accuracy on FNC without and with text augmentation.

Classification algorithm	Without data augmentation	With data augmentation					
		Synonym	BERT	TF-IDF substitution	TF-IDF insertion	Back-translation	Combination
DT	81.48%	75.27%	84%	80.34%	81.60%	**86.72**%	77.92%
SVM	**73.13**%	73.13%	73.13%	73.13%	73.13%	73.13%	73.13%
AdaBoost	72.45%	72.59%	72.73%	72.91%	**73.07**%	72.77%	72.84%
RF	86.82%	82.77%	89.09%	86.12%	86.69%	**90.60%**	84.77%
XGBoost	**80.13**%	77.89%	78.47%	78.03%	77.97%	79.26%	77.84%
Bagged RF	87.06%	82.66%	88.99%	86.14%	86.61%	**90.85%**	84.84%
LightGBM	**87.67**%	81.94%	85.77%	84.37%	84.66%	87.36%	82.88%
GradBoost	**78.66%**	77.28%	77.73%	77.38%	77.22%	78.36%	77.26%
LR	73.30%	73.44%	73.33%	73.38%	**73.47%**	73.33%	73.35%
NB	68.81%	68.09%	68.43%	67.99%	68.71%	68.79%	**69.64%**

Table 4. F1-score on FNN without and with data augmentation.

Classification algorithm	Without data augmentation	With data augmentation					
		Synonym	BERT	TF-IDF substitution	TF-IDF insertion	Back-translation	Combination
DT	81.80%	86.18%	86.31%	**87.70%**	87.66%	85.97%	86.30%
SVM	86.78%	86.79%	**86.80%**	86.65%	86.65%	86.78%	86.77%
AdaBoost	**87.92%**	87.83%	87.84%	87.88%	87.89%	87.85%	87.84%
RF	88.31%	90.19%	90.38%	**91.15%**	**91.15%**	90.03%	90.68%
XGBoost	88.15%	88.68%	88.67%	88.72%	**88.79%**	88.74%	88.71%
Bagged RF	88.09%	90.46%	90.40%	91.29%	**91.34%**	90.33%	90.68%
LightGBM	87.84%	89.41%	89.35%	**89.84%**	89.81%	89.29%	89.49%
GradBoost	88.07%	88.23%	88.15%	**88.23%**	88.21%	88.20%	88.18%
LR	86.85%	**87%**	86.98%	86.96%	86.97%	86.92%	**87%**
NB	86.46%	86.57%	86.51%	86.56%	**86.59%**	86.56%	**86.59%**

(Tables 4 and 6), where the F1-scores and Accuracy on the original dataset are higher than to those obtained on the augmented datasets;

(2) *There is no one-size-fits-all augmentation technique that performs well in all situations*. As shown in Tables 3 and 4 (resp. Figures 5 and 6), an augmentation technique may perform well when combined with a classification algorithm and poorly when combined with another. This is the case for example for the 'Synonym' technique which yields the highest F1-score when combined with AdaBoost and the lowest score when used with Naive Bayes (Table 3).

It is worth noting that even if BERT doesn't achieve the highest F1-scores, it provides a steady performance improvement for almost all classifiers;

(3) *The motto 'the more, the better' is the wrong approach regarding text augmentation and targeted approaches allow often better results*. This can be observed in Tables 3 and 4 (resp. Figures 5 and 6), where in almost all cases, combining all augmentation techniques does not yield the best F1-scores and Accuracy.

As shown in Tables 3 and 5, the pairs (*Back-translation, Bagged RF*) and (*Back-translation, RF*) yield the best performance on FNC and increase substantially the predictive performances ($\approx +4.16\%$ in comparison with the highest F1-Score that can be achieved without text augmentation). Similarly, as shown in Tables 4 and 6, the pairs (*Substitution TF-IDF, RF*) and (*Insertion TF-IDF, Bagged RF*) yield the best performance on the dataset FNN ($\approx +5.87\%$).

6.2.2. Augmentation-based ensemble learning

As previously stated, base learners' diversity and competency are the two key success factors of any ensemble learning approach. Our ensemble approach leverages text augmentation to enhance both. Figure 2 depicts our classification model which is a mixture of stacking and bagging. In our model, we use Bagged RF and Random Forest (RF) as base classifiers and GradBoost as meta-classifier. As depicted in Figure 2, each of the base classifiers is trained on a dataset composed of the original dataset and the data obtained by applying one of the augmentation techniques. The choice of the (classifier, augmentation technique) pairs was driven by the experimental study conducted in Subsection 6.2.1. We

Table 6. Accuracy on FNN without and with data augmentation.

Classification algorithm	Without data augmentation	With data augmentation					
		Synonym	BERT	TF-IDF substitution	TF-IDF insertion	Back-translation	Combination
DT	72.66%	78.98%	79.04%	81.10%	**81.34%**	78.60%	79.17%
SVM	76.83%	76.86%	**76.88%**	76.53%	76.53%	76.84%	76.81%
AdaBoost	**79.95%**	79.77%	79.77%	79.86%	79.87%	79.87%	79.79%
RF	81.02%	84.14%	84.42%	**85.75%**	85.72%	83.87%	84.94%
XGBoost	80.38%	81.27%	81.27%	81.36%	**81.50%**	81.39%	81.33%
Bagged RF	80.75%	84.48%	84.59%	86.06%	**86.07%**	84.36%	85.10%
LightGBM	80.30%	82.80%	82.70%	**83.57%**	83.49%	82.64%	82.97%
GradBoost	80.23%	80.46%	80.30%	80.45%	80.41%	**80.49%**	80.36%
LR	77.26%	**77.63%**	77.59%	77.52%	77.56%	77.52%	77.63%
NB	77.06%	**77.24%**	77.14%	77.17%	77.23%	77.22%	77.23%

Table 7. F1-score and accuracy achieved by conventional stacking, Li et al. (2019) and the proposed approach.

Model	F1-score		Accuracy	
	FNC	FNN	FNC	FNN
(TF-IDF Insertion. Stacking)	85.58%	**90.92%**	86.62%	**85.67%**
(TF-IDF Substitution. Stacking)	84.57%	90.43%	85.78%	84.72%
(Back-Translation. Stacking)	**90.31%**	89.80%	**89.70%**	83.47%
(BERT. Stacking)	87.93%	90.26%	88.62%	84.23%
(Synonym. Stacking)	80.71%	90.28%	82.71%	84.17%
(Combination. Stacking)	83.11%	90.73%	84.98%	85.28%
Li et al. (2019)	83.72%	88.45%	83.67%	79.31%
Proposed approach	**90.15%**	**91.07%**	**90.67%**	**85.78%**

compare our model to a more classical stacking approach, where all base classifiers are trained on the same dataset, consisting of the original dataset and the data obtained by applying one of the augmentation techniques (Figure 1). We also compare our model to the approach of Li et al. (2019), which is one of the state-of-the-art approaches that uses LSA, stacking-based ensemble learning and K-fold cross-validation. Table 7 synthesizes the predictive performances achieved by each approach.

As reported in Table 7, the use of text augmentation allows better performances than those achieved by Li et al. (2019) in almost all situations. On the other hand, except for the Synonym technique over the FNC dataset, our model outperforms the classical approach in all situations. Overall, our stacking approach achieves an increase in F1-score and Accuracy of 7,72% and 7,13% (resp. 7,54% and 2,88%) over FNC (resp. FNN) when compared to Li et al. (2019).

6.2.3. Class balancing

The class imbalance problem arises when data is distributed unevenly among classes; i.e. when one or more of the predicted outputs happen much less frequently than others. As stated in Fernndez et al. (2018), when working with an imbalanced classification problem:

- *The minority classes are typically of the most interest*, meaning that a model's skill in correctly predicting a minority class is more important than in correctly predicting a majority class.
- *Minority classes are harder to predict*. The main reason is that with few available training examples, it is often challenging to identify regularities and learn characteristics. That is why most classification algorithms tend to be biased towards the majority class(es), causing bad classification of the minority class(es).

The above two observations hold for the tasks of stance and fake news detection. As stated by the authors of the FNC challenge (Pomerleau & Rao, 2017),

> *'The related/unrelated classification task is expected to be much easier and is less relevant for detecting fake news…. The Stance Detection task (classify as agrees, disagrees or discuss) is both more difficult and more relevant to fake news detection, …'.*

When dealing with class imbalance, evaluation metrics that give equal importance to each observation (and not to each class), such as the *Accuracy*, can be misleading as they

Table 8. Per class F1-score on balanced data (FNN).

Model	Original dataset		Balanced FNN	
	False	True	False	True
DT	44.90%	82.43%	**85.18%**	**84.47%**
SVM	25.22%	87.93%	**66.26%**	74.16%
AdaBoost	40.42%	88.49%	**69.54%**	73.67%
RF	49.59%	88.35%	**88.21%**	88.02%
XGBoost	42.46%	88.74%	**69.90%**	74.58%
Bagged RF	48.04%	86.20%	**86.79%**	85.42%
LightGBM	43.01%	88.07%	**72.15%**	76.09%
GradBoost	42.15%	88.69%	**70.40%**	74.81%
LR	14.82%	87.07%	**66.89%**	72.50%
NB	24.37%	86.77%	**53.92%**	72.48%
Proposed approach	46.11%	88.84%	**87.14%**	87.65%

fail to reflect the poor performance over the minority classes (*Accuracy Paradox*). In the sequel, in order to better perceive the effects of class imbalance we provide per class F1-scores.

The FNN dataset is moderately imbalanced and contains 12287 *True News* and only 3831 *Fake News*. The FNC dataset has a more severe imbalance as it contains 36 545 *Unrelated*, 8 909 *Discuss*, and only 3 678 *Agree*, and 840 *Disagree* examples. The bias towards the majority classes can be observed in Tables 8 and 9 where all algorithms perform poorly over the minority classes *Fake News* (Table 8) and *Agree/Diagree* (Table 9). Some algorithms, such as SVM, Logistic Regression and Naïve Baies even obtain zero F1-scores over the minority classes *Agree/Disagree*.

To balance the FNN dataset, 6765 additional samples was generated for the minority class *Fake News*, using the five augmentation techniques of section 3. Similarly, we generated 18 390 additional examples for the *Agree* class and 4 200 additional examples for the *Disagree* class. We should notice here that the number of examples of the *Diagree* class remains much less than the number of instances of the majority class *Unrelated*. As shown in Tables 8 and 9, text augmentation allows a substantial improvement in the F1-scores obtained over the minority classes: an average improvement of 94.47% for the *Fake News* class (Table 8), 189.14% for the *Agree* class and 586.39% for the *Disagree* class (Table 9).

Table 9. Per class F1-score on balanced data (FNC).

Model	Original dataset				Balanced FNC			
	Agree	Disagree	Discuss	Unrelated	Agree	Disagree	Discuss	Unrelated
DT	40.70%	28.49%	68.80%	89.95%	**69.47%**	**84.52%**	59.06%	81.13%
SVM	0%	0%	33.51%	81.36%	**36%**	**38.89%**	21.48%	63.27%
AdaBoost	0%	0%	0.77%	84.30%	**31.26%**	**42.30%**	6.08%	59.09%
RF	45.99%	27%	74.74%	92.53%	**78.11%**	**89.53%**	68.08%	87.11%
XGBoost	1.85%	0%	41.08%	86.98%	**43.34%**	**45.28%**	34.10%	66.14%
Bagged RF	45.68%	28.87%	72.61%	91.94%	**74.56%**	**87.25%**	66.09%	84.77%
LightGBM	14.65%	8.56%	57.81%	88.99%	**54.40%**	**74.54%**	47.30%	75.13%
GradBoost	6.17%	3.45%	46.96%	87.56%	**44.87%**	**54.85%**	39.81%	68.12%
LR	0%	0%	2.42%	84.50%	**1.01%**	**28.20%**	2.86%	59.04%
NB	0%	0%	24.70%	80.92%	**15.27%**	**28.87%**	18.83%	57.75%
Proposed approach	47.78%	26.34%	75.35%	92.64%	**77.23%**	**89.26%**	65.85%	86.82%

7. Conclusion

Combating fake news on social media is a pressing need and a daunting task. Most of existing approaches on fake news detection, focus on using various features to identify those allowing the best predictive performance. Such approaches tend to undermine the generalization ability of the obtained models.

In this work, we investigated the use of text augmentation in the context of stance and fake news detection. In the first part of our work, we studied the effect of text augmentation on the performance of various classification algorithms. Our experimental study quantified the actual contribution of data augmentation and identified the best performing (*classifier, augmentation technique*) pairs. Besides, our study revealed that the motto 'the more, the better' is the wrong approach regarding text augmentation and that there is no one-size-fits-all augmentation technique. In the second part of our work, we proposed a novel augmentation-based ensemble learning approach. The proposed approach is a mixture of bagging and stacking and leverages text augmentation to enhance the diversity and the performance of base classifiers. We evaluated our approach using two real-world datasets. Experimental results show that the proposed approach is more accurate than state-of-art methods. In the third part of our work we investigated the use of text augmentation to cope with class imbalance, a very common problem in stance and fake news detection. As shown by our experimental study, even in presence of severe imbalance, text augmentation can highly alleviate its effects and substantially improve the predictive performance over the minority classes.

As a part of our future work, we intend to explore the use of a multimodal data augmentation that involves linguistic and extra linguistic features.We also intend to explore the detection of fake news from streams under concept drifts and to connect the dots between stance detection and fake news detection through the use of media profiling and multi-source credibility scores.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on Contributors

Ilhem Salah is currently pursuing a Ph.D. in Computer Science at Sousse University (Tunisia) after earning her Master's degree in Distributed Computing from the same institution. Her research interests include Machine Learning, Distributed Ledgers, and NLP.

Khaled Jouini received the Ph.D. degree in Computer Science from Paris-Dauphine University (France) in 2008. He was a research staff member at Telecom ParisTech (France). Since 2011, he has been with Sousse University (Tunisia), where he is currently an Associate Professor. His research interests include Data Engineering, Machine Learning, NLP, and Large-scale data management and mining.

Ouajdi Korbaa is a full-time professor at the University of Sousse (Tunisia). He received his Engineering Diploma from the Ecole Centrale de Lille (France) in 1995 and his Master's degree in Production Engineering and Computer Science from the University of Lille (France) in the same year. He obtained his Ph.D. in Production Management, Automatic Control, and Computer Science from the University of Science and Technologies of Lille (France) in 1998 and his "Habilitation to Supervise Researches" degree in Computer Science from the same University in 2003. Pr. Korbaa

has published around 150 research papers on Optimistation, Simultation and Modeling, Applied and Computational Mathematics, Manufacturing Engineering and Computer Engineering.

ORCID

Ilhem Salah http://orcid.org//0000-0002-3375-3637
Khaled Jouini http://orcid.org//0000-0001-5049-4238
Ouajdi Korbaa http://orcid.org/0000-0003-4462-1805

References

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, secure, and dependable systems in distributed and cloud environments: First international conference* (pp. 127–138). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(2003), 993–1022.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805

Dulhanty, C., Deglint, J. L., Ben Daya, I., & Wong, A. (2019). Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. CoRR, abs/1911.11951

Fernndez, A., Garca, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (1st ed.). Springer Publishing Company, Incorporated.

Guo, Z., Schlichtkrull, M. S., & Vlachos, A. (2021). A survey on automated fact-checking. CoRR, https://arxiv.org/abs/2108.11896

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Association for Computational Linguistics.

Jiang, T., Li, J. P., Haq, A. U., Saboor, A., & Ali, A. (2021). A novel stacking approach for accurate detection of fake news. *IEEE Access*, *9*(2021), 22626–22639.

Jouini, K., Maaloul, M. H., & Korbaa, O. (2021). Real-time CNN-based assistive device for visually impaired people. In *2021 14th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI)* (pp. 1–6). IEEE.

Kaggle.com (n.d.). *Fake and real news dataset*. Retrieved February 19, 2023, from https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/discussion

Karnyoto, A. S., Sun, C., Liu, B., & Wang, X. (2022). Augmentation and heterogeneous graph neural network for AAAI2021-COVID-19 fake news detection. *International Journal of Machine Learning and Cybernetics*, *13*(7), 2033–2043. https://doi.org/10.1007/s13042-021-01503-5

Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, *4*(June), 2666–8270. https://doi.org/10.1016/j.mlwa.2021.100032

Li, B., Hou, Y., & Che, W. (2021). Data augmentation approaches in natural language processing: A survey. CoRR, abs/2110.01852

Li, S., Ma, K., Niu, X., Wang, Y., Ji, K., Yu, Z., & Chen, Z. (2019). Stacking-based ensemble learning on low dimensional features for fake news detection. In *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (pp. 2730–2735). IEEE.

Ma, E. (2019). *NLP augmentation*. Retrieved May 15, 2021, from https://github.com/makcedward/nlpaug

Mahabub, A. (2020). A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers. *SN Applied Sciences*, *2*(4), 525. https://doi.org/10.1007/s42452-020-2326-y

Marivate, V., & Sefara, T. (2019). Improving short text classification through global augmentation methods. CoRR, abs/1907.03752

McIntire, G. (2017). *Machine learning finds 'fake news' with 88% accuracy*. Kdnuggets, ODSC. Retrieved February 19, 2023, from https://www.kdnuggets.com/2017/04/machine-learning-fake-news-accuracy.html

Momchil, H., Arnav, A., Preslav, N., & Isabelle, A. (2022). A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 1259–1277). Association for Computational Linguistics.

NLTK.org. (n.d.). *Natural language toolkit*. Retrieved May 15, 2021, from https://github.com/nltk/nltk

Pan, Y. (2018). *Fake news challenge – team solat in the swen*. Retrieved February 22, 2023, from https://github.com/Cisco-Talos/fnc-1/

Patil, D. R. (2022). Fake news detection using majority voting technique. arXiv, https://arxiv.org/abs/2203.09936

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(2011), 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Pomerleau, D., & Rao, D. (2017). *The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news*. Retrieved April 1, 2022, from https://www.fakenewschallenge.org/

Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the fake news challenge stance detection task. CoRR, http://arxiv.org/abs/1707.03264

Salah, I., Jouini, K., & Korbaa, O. (2022). Augmentation-based ensemble learning for stance and fake news detection. In *Advances in Computational Collective Intelligence – 14th International Conference, ICCCI 2022, Proceedings of Communications in Computer and Information Science* (Vol. 1653, pp. 29–41). Springer.

Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., & Palomar, S. M. (2021, November). HeadlineStanceChecker: Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, *71*, 100660 https://doi.org/10.1016/j.websem.2021.100660

Shi, L., Liu, D., Liu, G., & Meng, K. (2020). AUG-BERT: An efficient data augmentation algorithm for text classification. In *Communications, signal processing, and systems* (pp. 2191–2198). Springer.

Shoemaker, E. (2019). *Using data science to detect fake news*. James Madison University JMU Scholarly Commons, https://orcid.org/0000-0002-7955-5441

Shorten, C., Khoshgoftaar, T., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data*, *8*(1), 1–34. https://doi.org/10.1186/s40537-021-00492-0

Shu, K. (2019). FakeNewsNet. Retrieved December 15, 2021, from https://doi.org/10.7910/DVN/UEMMHS, Harvard Dataverse, V2

Slovikovskaya, V. (2019). Transfer Learning from Transformers to Fake News Challenge Stance Detection {(FNC-1)} Task. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association.

Surowiecki, J. (2005). *The wisdom of crowds* (1st ed.). Anchor Books.

Suting, Y., & Ning, Z. (2020). Construction of structural diversity of ensemble learning based on classification coding. In *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* (Vol. 9, pp. 1205–1208). *IEEE*.

Tesfagergish, S. G., Damaševičius, R., & Kapočiūtė-Dzikienė, J. (2021). Deep fake recognition in tweets using text augmentation, word embeddings and deep learning. In *ICCSA 2021: 21st International Conference Computational Science and Its Applications* (pp. 523–538). Springer Nature.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new Benchmark dataset for fake news detection. CoRR, abs/1705.00648

Xie, Q., Dai, Z., Hovy, E. H., Luong, M. T., & Le, Q. V. (2019). Unsupervised data augmentation. CoRR, abs/1904.12848