

Enhancing prediction of user stance for social networks rumors

Kholoud Khaled, Abeer ElKorany, Cherry A. Ezzat

Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt

Article Info

Article history:

Received Jan 22, 2023

Revised Apr 21, 2023

Accepted Apr 24, 2023

Keywords:

Augmentation

Machine learning

Social network analysis

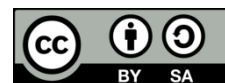
Stance prediction

Text analysis

ABSTRACT

The spread of social media has led to a massive change in the way information is dispersed. It provides organizations and individuals wider opportunities of collaboration. But it also causes an emergence of malicious users and attention seekers to spread rumors and fake news. Understanding user stances in rumor posts is very important to identify the veracity of the underlying content as news becomes viral in a few seconds which can lead to mass panic and confusion. In this paper, different machine learning techniques were utilized to enhance the user stance prediction through a conversation thread towards a given rumor on Twitter platform. We utilized both conversation thread features as well as features related to users who participated in this conversation, in order to predict the users' stances, in terms of supporting, denying, querying, or commenting (SDQC), towards the source tweet. Furthermore, different datasets for the stance-prediction task were explored to handle the data imbalance problem and data augmentation for minority classes was applied to enhance the results. The proposed framework outperforms the state-of-the-art results with macro F1-score of 0.7233.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Kholoud Khaled

Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University

Cairo, Egypt

Email: k.khaled@fci-cu.edu.eg

1. INTRODUCTION

Social media has become an important part of people's lifestyle. The role of traditional information channels such as newspapers and television on how we collect and consume news has become less prominent than in the past. The growth of social media platforms has definitely played a crucial role in this transformation. In fact, social networks like Twitter and Facebook have registered an exponential spike in popularity. Twitter has been used in various research as a data source for stance detection [1]–[3]. Many people use social media platforms not only to communicate with friends and family, but also to gather information and news from all around the world. Thus, social media play a fundamental role in the news fruition [4].

Social media have become a critical publishing tool for journalists [5]–[7] and the main consumption method for people looking for the latest news. Journalists may use social media to discover new stories and report on public opinions about breaking news stories, whereas people may follow the evolution of breaking news and events through posts published on their own network or through official channels. Indeed, social networks have proved to be extremely useful especially during crisis situations, because of their inherent ability to spread breaking news much faster than traditional media. This positive impact of the social networks comes at a cost: the absence of control and fact-checking over posts makes social networks a fertile ground for the spread of fake news and rumors. Rumor is defined as a circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety [8]. People often publish posts or share other people's posts verifying neither the source nor the information validity and reliability.

User stance classification involves identifying the attitude of users towards the truthfulness of the rumor they are discussing. Derczynski *et al.* [3] proposed a 4-way classification task to encompass all the different kinds of reactions to rumors. The schema of classifications includes supporting, denying, querying, and commenting. User Stance classification is an important step towards rumors detection and prevention by utilizing different knowledge sources.

Thus, in this paper, a framework for prediction of user stance for social networks rumors (PSSNR) is presented. The basic idea of PSSNR is to enhance the prediction by exploring two main issues. The first one is the type of features that are fed to the machine learning model. In PSSNR, new user features that represent the user's influence as well as communication aspects are considered. The second challenge was the imbalance of benchmark dataset available. Therefore, data augmentation was used to balance SemEval-2019 task 7 dataset [9] to a certain degree. Finally, traditional as well as ensemble machine learning algorithms were applied for stance prediction. The structure of this paper is as follows: section 2 discusses different related work for rumor stance prediction for both single as well as conversation thread. The proposed framework is presented in section 3. The proposed PSSNR framework is decomposed of two phases each are described in subsections 3.1 and 3.2 respectively. A set of conducted experiments to proof the validity of the PSSNR framework is discussed in section 4. Finally, the conclusions are presented in section 5.

2. RELATED WORK

Stance detection on social media has been a growing research interest measuring different aspects of online human behavior, including the public stance toward various social and political aspects on social media. There are two different types of stance detection tasks: user stance classification for a single tweet, stance classification for a conversation thread. Each type will be discussed in this section.

2.1. Single tweet stance classification

The task is formulated as follows: given a tweet and a target entity (person, organization, ...), stance detection aims to determine from the tweet whether the author (tweeter) is in favor of, against, or neutral towards the given target [10]. Stance classification was defined as a binary classification task (believe versus refute or question the rumor) [11]. Qazvinian *et al.* [11] used tweets observed in the past to train a classifier, which is then applied to new tweets discussing the same rumor. The proposed dataset is a claim-based manually annotated dataset containing 10,417 tweets related to 5 different controversial topics, 6,774 marked as rumor tweets, 2,971 of which show belief, and 3,803 tweets show that the user is doubtful, denies, or questions it. They used content, twitter-based, and network-based features. Their approach is based on Bayes classifiers. They achieved F1-score of 0.932 using content-based features only such as unigrams and bigrams. [12] used bag-of-words autoencoder trained along with Hillary-labeled data for extracting features and classification was performed using logistic regression (LR), they achieved F1-score 0.327.

Zen *et al.* [13] that applied stance classification for rumors emerging during crises, built supervised machine learning models (logistic regression, naïve Bayes and random forest) and obtained F1-Score of 0.917. They only used a two-way classification approach where tweets were classified into affirm and deny. The proposed dataset contains over 4,300 manually annotated tweets. The best results were achieved using a random forest (RF) classifier. They enriched the feature sets of earlier studies by linguistic features using linguistic inquiry and word count (LIWC) dictionary. A variety of feature categories, including n-grams, basic textual features, sentiment, part of speech, and lexical LIWC features were used in this model. Stance classification towards a target on Twitter has been addressed in SemEval-2016 task 6 [2] where about 19 teams competed. In task A, the stance of tweets had to be determined towards five different targets with three stances either favor, against or none. The dataset of the competition was not related to rumors or breaking news. The dataset contains 2,914 labeled training data instances for five targets and 1,249 testing data instances. The winning team achieved a macro F1-score of 0.67. Zarrella and Marsh [14] was composed of two recurrent neural network (RNN) classifiers: the first was trained to predict task-relevant hashtags on a very large unlabeled Twitter corpus. This RNN was used to initialize the second RNN classifier, embeddings of words and phrases trained with the word2vec skip-gram method, then used those features to learn sentence representations via a hashtag prediction auxiliary task. These sentence vectors were then finetuned for stance detection, which was trained with the provided task A data. The most frequent stance classes uniformly outperformed the minority classes by all metrics.

2.2. Thread stance classification

The task is formulated as a tree-structured conversation consisting of a parent tweet (rumor) followed by a set of tweets replying to it either directly or indirectly. Each tweet presents its own type of stance supporting, denying, querying, or commenting (SDQC) with respect to the rumor. The tree structure of tweets improved the performance of an independent classifier [15].

The first work which initiated this idea was during SemEval-2017 task 8 [3], they presented an annotation scheme and a large dataset covering multiple topics. Training dataset comprises 297 rumorous threads collected for 8 targets in total, which include 297 source and 4,222 reply tweets, amounting to 4,519 tweets in total. The testing dataset includes, in total, 1,080 tweets, 28 of which are source tweets and 1,052 are replies. For task A, the dataset is annotated by four different classes: support, deny, query and comment. The distribution of the stance classes is clearly skewed towards the comments class. Systems generally viewed task A as a four-way classification task. The winning system [16] achieved F1-score for support class of 0.403 while the F1-score for deny class was 0 with overall macro F1-score 0.434 and accuracy 0.784. They propose long short-term memory (LSTM)-based sequential model that models the conversational structure of the tweets with content-based features such as tweet lexicon, punctuation, attachment, tweet role. The low macro F1-score of deny affected the overall macro F1-score. Bahuleyan and Vechtomova [17] ranked second with accuracy 0.780 and F1-score 0.45, they used topic independent features from two categories, namely cue features and message specific features to fit a gradient boosting classifier. Chumachenko *et al.* [18] applied three models (random forest (RF), k-nearest neighbors (KNN) and gradient boosting) for predicting the dynamics of epidemic process in specific areas using statistical machine learning methods. In [19], a trust-based access control approaches have been proposed to show how behavioral parameters of different cloud users and service providers taken into consideration to find trusted resources for cloud users and calculate the trust value of cloud service providers based on the quality-of-service parameters and user feedback.

In SemEval-2019 task 7 [9], the dataset is expanded with new twitter posts. SemEval-2017 task 8 dataset became training data in 2019 and was augmented with new twitter test data. The winning system [20] in task A (stance classification) achieved macro F1-score of 0.6178. They proposed an inference chain-based system fine-tuned on generative pre-trained transformer (OpenAI GPT), which fully utilizes conversation structure-based knowledge. They divide an inference chain into four parts: source tweet, other tweets, parent tweet and target tweet. They used Language-based features for word-level as well as other tweet-level features. They used word-level features such as whether the tweet content has punctuation, hashtags, URLs, negative words, positive words. While for the tweet-level features, verified, followers, friends, retweet count were used. They alleviate class imbalance in stance classification by expanding training data in the under-represented classes with pre-screened external data from similar datasets.

Prakash and Madabushi [21] integrate count-based feature into pretrained models such as bidirectional encoder representations from transformers (BERT) and RoBERTa using ensemble on the dataset released at SemEval-2019 task 7. They ensemble the pooled output of RoBERTa with the output of multilayered-perceptron MLP (consisting of four units, one for each class). This combination is then connected to a linear layer followed by a SoftMax function to make predictions. They achieve Macro F1-Score of 0.64. Khandelwal [22] proposed a multi-task learning framework for jointly-predicting rumor stance and veracity on the dataset released at SemEval-2019 task 7. The proposed method represents the model averaging based on three different architectures with language-based features trained with varying parameters encoder and learning rate. They achieved a Macro F1-score of 0.6720. We believe that there is still room for improvement using other combinations of features, in addition to dataset balancing by applying one of the current state-of-the-art augmentation methods with similar datasets.

3. PROPOSED PSSNR FRAMEWORK

As shown in Figure 1, PSSNR framework aims to enhance the prediction of user stance against the source tweet in a conversation thread. PSSNR framework consists of two main phases. Phase 1 describes the dataset preparation process which includes 2 main steps: i) dataset augmentation to solve the dataset imbalance problem, and ii) applying text analysis for the content of tweets to be considered as features for the machine learning (ML) models. Phase 2 describes applying different ML algorithms. Phase 2 is also decomposed of different steps such as: Feature engineering for extracting conversation thread and user-based features and categorizing them, and stance prediction using different ML algorithms. In the following subsections, each of those phases are described in detail.

3.1. Phase 1: dataset preparation process

This phase is composed of 2 steps: i) dataset augmentation, and ii) text analysis. In dataset augmentation phase, we explore different methods used to solve imbalanced dataset problems. In text analysis phase we applied two types of text analysis, sentiment analysis using different tools and semantic analysis as we found that content features are the most affecting features in stance prediction.

3.1.1. Dataset augmentation

A major issue with the current available benchmark datasets is the class imbalance in which the distribution of data is skewed towards comment class, accounting for about 50% of the tweets, which affects

any model generated using the imbalanced dataset. Therefore, we explore two different methods for addressing this problem. The first method, which is used in [20], solved this by augmenting the under-represented classes (support, deny, question) from other datasets so that all classes are balanced. The second method is the use of two-step classification [23], where the first classifier classifies the tweets to comment and non-comment, while the second classifier distinguishes non-comments as support, deny or comment. After exploring, we find that the first method is more reliable and gives better results. We explore and merge multiple similar datasets on Twitter in different topics to expand the training data for minority classes to alleviate class imbalance in the SemEval-2019 task 7 dataset.

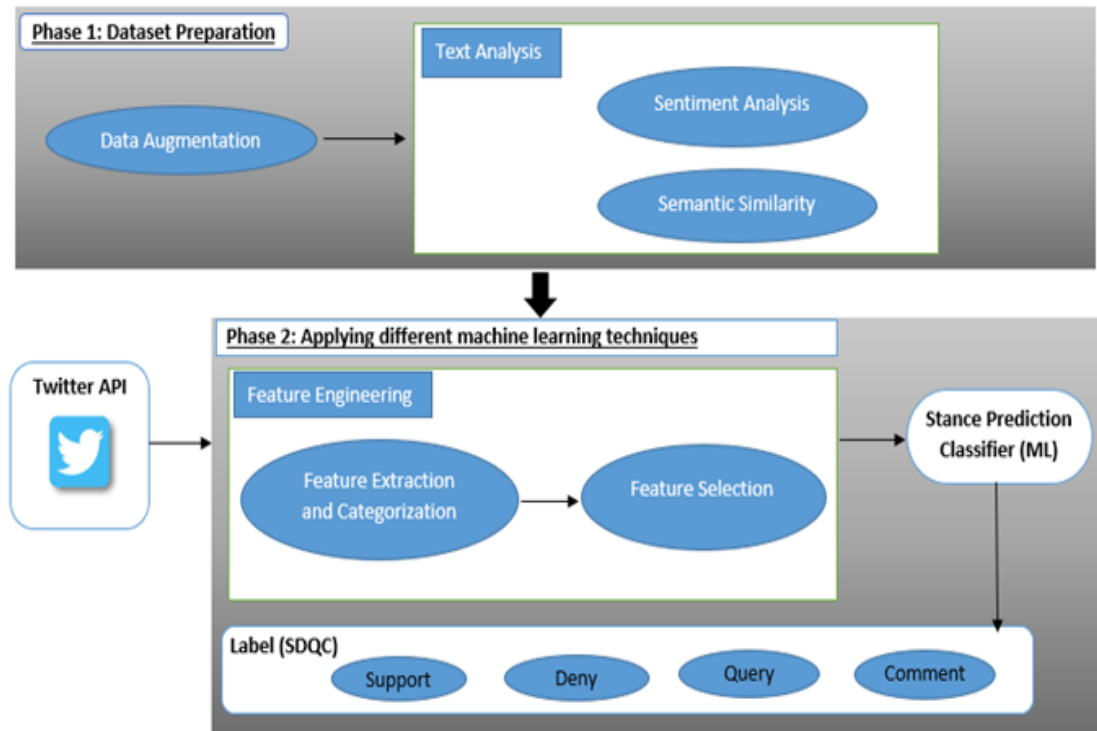


Figure 1. PSSNR framework phases

a. Exploring benchmark datasets

Starting from SemEval-2017, task 8: rumor eval was introduced, where participants analyzed rumors in the form of claims made in user-generated content (source tweet), and users responding to it within conversations. SemEval-2019 task 7 Dataset [9]: a large dataset covering multiple topics. Training dataset comprises 297 rumor threads collected for 8 targets in total, which includes 297 source and 4,222 reply tweets, amounting to 4,519 tweets in total. SemEval-2017 task 8 dataset became training data in 2019, and was augmented with new Twitter test data, amounting to 5,568 tweets in total. According to Table 1, the imbalanced classes are deny, query, as well as support, therefore, other datasets which will be described in augmented datasets section are required to be added to balance the overall dataset.

Table 1. Distribution of tweets between classes in SemEval-2019 task 7

	Support	Deny	Query	Comment	Total
Training set	1004	415	464	3685	5568
Testing set	141	92	62	771	1066

b. Augmented datasets

Data augmentation (DA) is an effective strategy for handling scarce data situations to balance these datasets. Synthetic data augmentation method such as SMOTE, random oversampling, and under sampling, have been traditionally used in many application domains. Therefore, several benchmark datasets have been investigated to be used in data argumentation of SemEval-2019 task 7. According to Table 1, it has been found

that if we down-sample both the comment and support classes to get balanced sample size, extensive cases will be neglected. Accordingly, we only down-sampled the comment class, and we applied random oversampling in which data is added to the minority classes-support and deny-to balance the data. To do so, similar benchmark datasets used previously in stance detection are used to balance SemEval-2019 task 7 training dataset for down-sampled classes. The following datasets [22], [23] are used to balance the distribution of tweets after merging Table 2, given that we could not find more examples in query class.

- SRQ-2020 dataset (4 Events) [24]:

A human-labeled stance dataset for Twitter conversations (both replies and quotes) with over 5,200 stance labels. The collection methodology of tweets favors the selection of denial-type responses as this class is expected to be more useful in the identification of rumors. We added 1,165 tweets in the three classes: support, deny and query.

- (WT-WT)-2020 dataset [25] in two different domains (healthcare and entertainment):

A large dataset of English tweets targeted at stance detection for the rumor verification task. The dataset contains 51,284 tweets. We added only 95 tweets in support class. After augmentation the dataset became more balanced with a total of 3,837 training examples and 463 testing examples.

Table 2. Distribution of tweets between classes in augmented dataset

	Support	Deny	Query	Comment	Total
Training set	1111	1248	468	1010	3837
Testing set	122	177	41	123	463

3.1.2. Text analysis

Pamungkas *et al.* [26] thread tweets content is the most affecting feature in enhancing the prediction of user stance. Therefore, it is crucial to apply different text analysis for the content of tweets. In this step, two main text analysis methods are applied in order to add text-based features that well represent the thread tweets. The first one is the sentiment analysis of each of the thread tweets which is expressed by the sentiment of the tweet. The second one is calculating the semantic similarity between each of the thread tweets and the source tweet. Details of each of those steps are described in this section.

a. Sentiment analysis

Mohammad *et al.* [10] shows that knowing the sentiment expressed by a tweet is useful in stance detection. Therefore, the following tools for sentiment analysis were explored.

- A lexicon-based sentiment analyzer: text blob is used which takes a single sentence as input and returns polarity and subjectivity. Polarity score lies between [-1,1] where -1 identifies the most negative words and 1 identifies the most positive words.
- A pre-trained sentiment analysis model: we used BERTweet-sentiment-analysis trained with SemEval 2017 task 8 corpus contains around 40K tweets. The base model is BERTweet, a RoBERTa model trained on English tweets.
- A rule-based sentiment analysis tool: we used Valence aware dictionary and sentiment reasoner (VADER) Sentiment tool in sentiment analysis of reply to tweets. VADER is specifically attuned to sentiments expressed in social media. It can very well understand the sentiment of a text containing emoticons, slangs, conjunctions, capital words, punctuations and much more.

Among the different tools, the results showed that VADER sentiment analyzer is the best performing one. VADER sentiment analyzer gives us four scores: positive, negative, neutral and compound. The compound score is the sum of positive, negative and neutral scores which is then normalized between -1 (most negative) and +1 (most positive). The compound score is used with a threshold value for the analysis of the text. We follow the standard scoring metric followed by most of the analyzers for the threshold value [27]. Positive sentiment if compound score ≥ 0.05 , neutral sentiment if compound score > -0.05 and compound score < -0.05 and negative sentiment if compound score ≤ -0.05 .

b. Semantic analysis

Semantic textual similarity score represents the relationship between texts or documents using a defined metric. The similarity score indicates whether two texts have similar or different meanings. We tested two different approaches in calculating the semantic score using the cosine similarity metric:

- In the first approach, steps include the following: i) tokenize both reply and source tweet, ii) remove stop words, iii) form a set containing keywords in both tweets and convert them to vectors, and iv) use cosine similarity formula between two vectors.
- In the second approach, we used a pre-trained model 'sts-b-RoBERTa-large' which uses RoBERTa-large as a base model. In this approach, we used the model to encode the tweets and then calculate the cosine similarity of the resulting embeddings.

We applied both approaches as two different features used in stance detection. We found that the second approach is the best performing approach which has a useful effect on stance detection. It is significant to mention that feature normalization to scale values between the range of 0 and 1 was applied such that machine learning models could interpret features varying in degrees of range and unit on the same scale.

3.2. Phase 2: Applying different machine learning techniques

This phase is composed of 2 steps: i) feature engineering and ii) stance prediction model. In the first step, different features were extracted and selected to improve the performance of the machine learning models. While in the stance prediction step, we described the details of applying different machine learning algorithms.

3.2.1. Feature engineering

In this phase, different features that could be fed to the machine learning model are extracted. There are two main categories of features extracted: user-based and thread features using the Twitter application programming interface (API) and other libraries. In this section, feature extraction and feature selection steps are described in detail.

a. Feature extraction and categorization

After data has been cleaned and classes have been balanced, features related to the stance should be extracted. Many features could be used to build a model to predict the user stance, but most of them are not provided in the current datasets available. Therefore, Twitter API is used to extract other features that would be used to enhance the prediction using the tweet ID which is provided in the datasets. In order to improve the user stance prediction, two categories of features are investigated, one is related to the conversation thread itself and the other related to users who shared and responded to the rumor thread. Conversation thread features are also classified based on their type either content or structural while user features are all structural features as described in Tables 3 and 4 respectively.

Table 3. Conversation thread features

Type	Feature	Explanation
Content	Encoded tweet	TF-IDF Vectorizer for transforming text to numerical features.
	Punctuation Flag	Boolean (True/False) value that represent whether text contains these characters '!(),-;:<=>?^{} ~'
	Media Flag	Boolean (True/False) value that represent whether text contains image or video.
	Emoji Flag	Boolean (True/False) value that represents whether text contain emoji.
	Sentiment (positive, negative or neutral)	Using Vader Sentiment tool.
	Similarity score between reply and source.	Using two methods: -Cosine Similarity. -Transformers.
Structural	Content Length	Count of characters
	Retweet count	Number of retweets.
	Favorite count	Number of likes.
	Reply count	Number of users replied to the tweet.

Table 4. User-based features

Type	Feature	Explanation
User popularity	Followers count	Number of users who follows this user
	Listed count	Number of lists that include this user
	Verified	Indicates if this user is a verified twitter account
User behavior	Following count	Number of users this user is following
	Statuses count	Number of tweets (including retweets) posted by this user
	Favorite count	Total Number of likes in the user timeline
Similarity with source tweet user	Common friends	Number of common friends between source and reply tweet users.
	Following flag	Boolean (True/False) value that determines if user follows source tweet user.

b. Feature selection

We used feature selection to determine the most important features that have a huge influence on stance prediction. Irrelevant features can negatively impact the model's performance. Therefore, we calculate the mutual information value for each of the independent variables with respect to dependent variable and select the ones which have the most information gain. Feature importance is presented in Figure 2. Experiments were done starting with the top 5 features and then gradually adding more features one by one until reaching the top 10 features. We found that the top 9 features are the best combination that achieves the highest F1-score of 0.7233.

3.2.2. Stance prediction model

We apply different machine learning algorithms for stance prediction where the model is built with features extracted in Phase 2. Logistic regression, support vector machine (SVM) and random forest are common algorithms for stance prediction. A stacking ensemble technique is also used for a majority voting of algorithms to decide on the final stance. The stacking ensemble technique using logistic regression, support vector machine, random forest and decision tree (DT) as base models with default models hyperparameters except for RF with 60 n estimators and logistic regression as meta model outperforms state of the art. Base models are the models that fit on the training data creating a new feature matrix for the meta-classifier layer, while meta model is the model that learns how to best combine the predictions of the base models. The training set is divided into k-folds, where k-equal 5 and each base model is trained upon k-1 part and prediction is done on the kth part, the process is iterated until every fold is predicted.

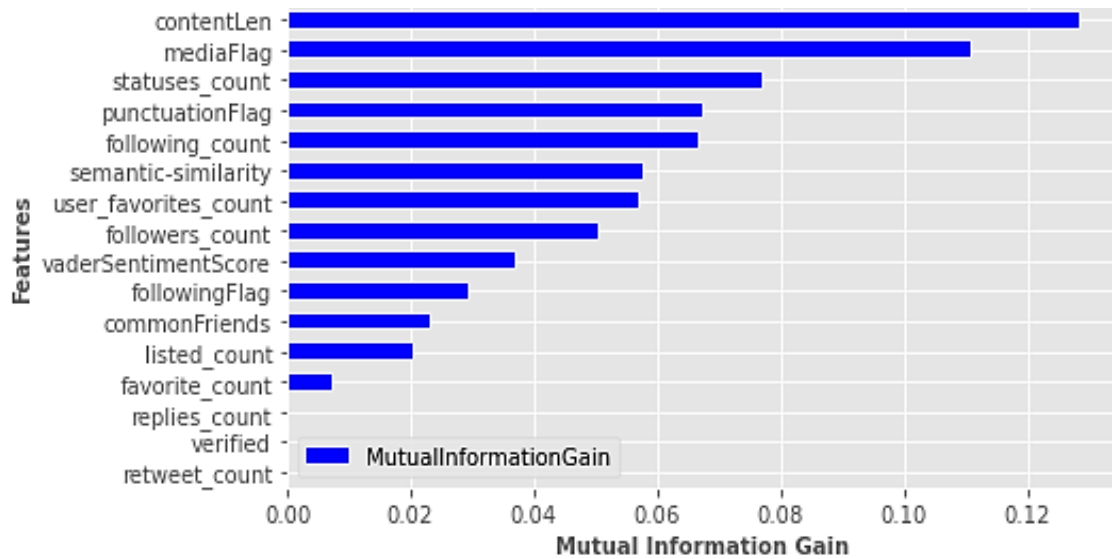


Figure 2. Feature selection results using information gain method

4. RESULTS AND DISCUSSION

The results of experiments done on applying data augmentation to balance the training set of RumorEval 2019 dataset [9] is presented in Table 5 using the same evaluation metric as in [4]. The feature applied in this experiment include conversation thread features, user features, all features and top 9 features using information gain method in feature selection (feature selected). Figure 3 shows the confusion matrix of the outperforming method on augmented dataset. Figure 4 presents a comparison between the value of F1 measure when using different feature sets.

Table 5. Results using different algorithms and features in augmented datasets

Algorithm	Feature	Support	Deny	Query	Comment	Macro F1
LR	Conv.Thread	0.56	0.67	0.67	0.81	0.679
SVM	Conv.Thread	0.51	0.64	0.64	0.84	0.658
RF	Conv.Thread	0.58	0.66	0.65	0.83	0.678
Ensemble	Conv.Thread	0.58	0.67	0.69	0.85	0.698
LR	User	0.49	0.58	0.57	0.79	0.606
SVM	User	0.54	0.62	0.51	0.77	0.610
RF	User	0.56	0.64	0.59	0.80	0.598
Ensemble	User	0.52	0.71	0.59	0.85	0.669
LR	All	0.57	0.65	0.70	0.83	0.686
SVM	All	0.54	0.65	0.62	0.83	0.661
RF	All	0.58	0.70	0.54	0.81	0.656
Ensemble	All	0.54	0.68	0.67	0.83	0.679
LR	Selected Features	0.56	0.66	0.69	0.81	0.682
SVM	Selected Features	0.53	0.63	0.63	0.84	0.656
RF	Selected Features	0.57	0.65	0.63	0.82	0.669
Ensemble	Selected Features	0.60	0.70	0.73	0.87	0.723

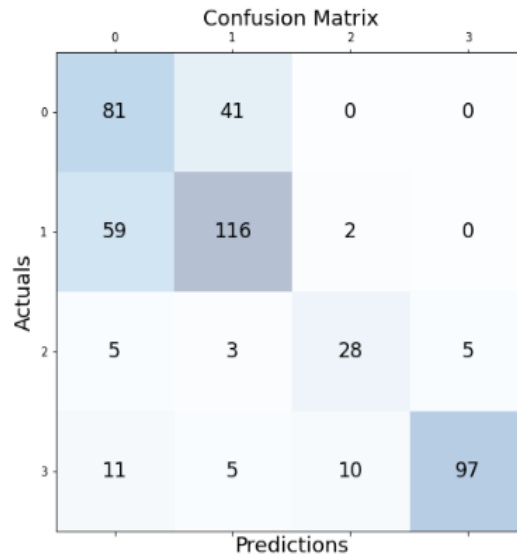


Figure 3. Confusion matrix of ensemble approach on augmented dataset

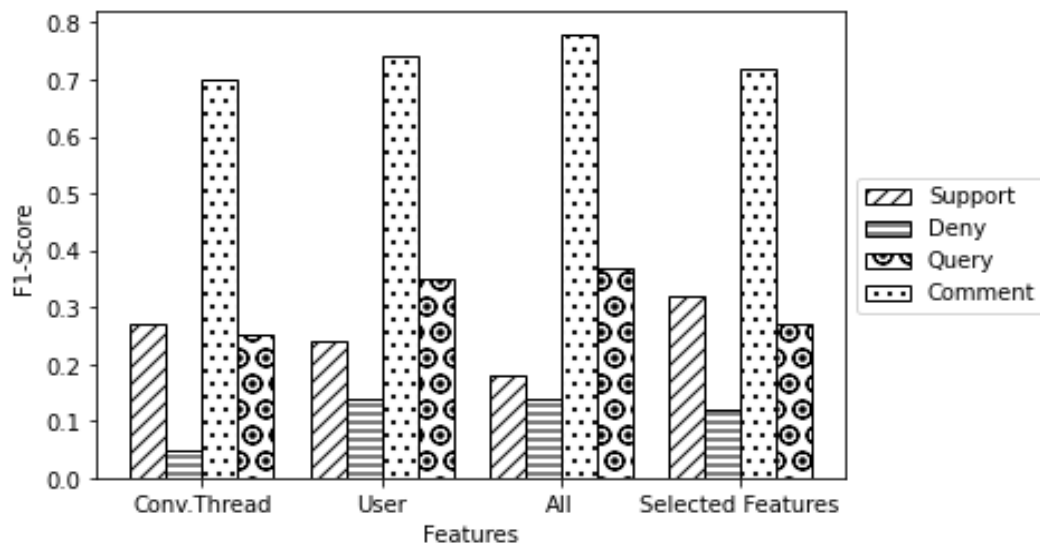


Figure 4. F1-Score after applying ensemble approach on different set of features before augmentation

According to Figure 4, the variation in the value of F1 using different sets of features is remarkable due to the imbalance of dataset. Therefore, different machine algorithms are applied on different feature sets on SemEval-2019 task 7 dataset with augmentation according to Figure 5 and Table 6. We observed the outstanding effect of data augmentation on deny and support classes by a sufficient margin. The Macro F1-score increased from 0.371 before augmentation using user-based features with 0.24 in support class and 0.14 in deny class to 0.723 using a specific combination of features as a result of feature selection method. Information gain for each independent value with respect to the stance is applied as the feature selection method to select the top features. Stacking ensemble technique outperforms other machine learning algorithms where multiple base models such as SVM, logistic regression and random forest fit on the training data and logistic regression is used as a meta model that combines the predictions of the base models. Based on the previous experiments results of applying different ML algorithms on augmented dataset shown in Table 5, stacking ensemble is the outperforming algorithm with macro F1-score 0.723. A comparison between the outperforming algorithm and other recent state-of-the-art work that also handled class-imbalance of SemEval 2017, task 8 is presented in Table 7. According to Table 7, the proposed framework that applied stacking ensemble using top selected features is beating the current state of the art using the same datasets.

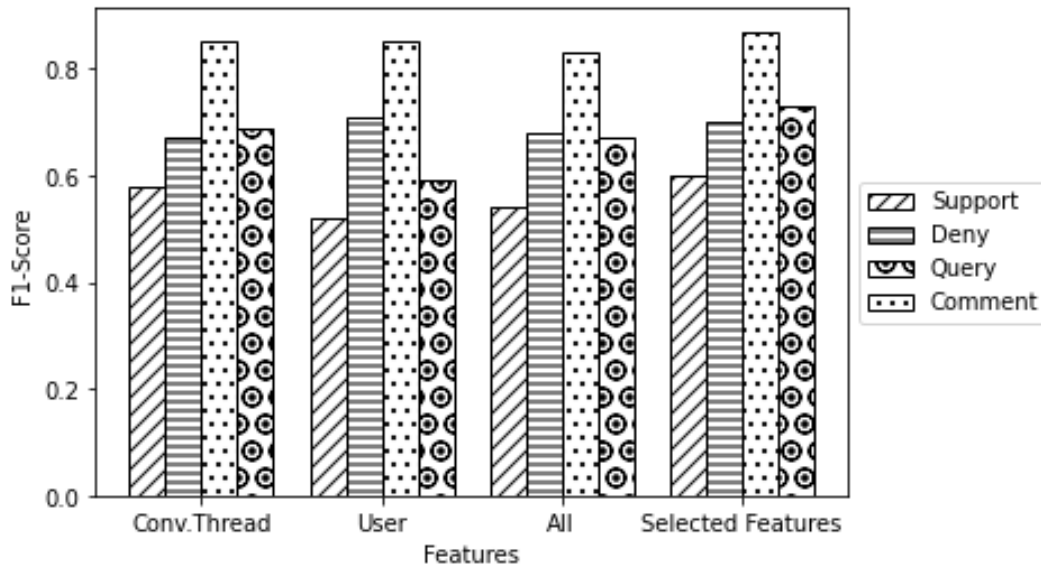


Figure 5. F1-Score after applying ensemble approach on different sets of features after augmentation

Table 6. Results using different algorithms and features in SemEval-2019 task 7 dataset

Algorithm	Feature	Support	Deny	Query	Comment	Macro F1
LR	Conv.Thread	0.22	0.09	0.23	0.66	0.301
SVM	Conv.Thread	0.16	0.05	0.22	0.70	0.285
RF	Conv.Thread	0.18	0.07	0.23	0.60	0.267
Ensemble	Conv.Thread	0.27	0.05	0.25	0.70	0.314
LR	User	0.26	0	0.31	0.79	0.342
SVM	User	0.27	0	0.29	0.80	0.338
RF	User	0.21	0.04	0.36	0.66	0.32
Ensemble	User	0.24	0.14	0.35	0.74	0.371
LR	All	0.24	0.09	0.28	0.69	0.325
SVM	All	0.20	0.11	0.28	0.63	0.304
RF	All	0.17	0.15	0.35	0.72	0.349
Ensemble	All	0.18	0.14	0.37	0.78	0.368
LR	Selected Features	0.22	0.09	0.23	0.66	0.302
SVM	Selected Features	0.20	0.05	0.22	0.70	0.292
RF	Selected Features	0.22	0.08	0.31	0.72	0.333
Ensemble	Selected Features	0.32	0.12	0.27	0.72	0.359

Table 7. Results of outperforming method compared to the state of the art showing macro F1-scores for each class

Model	Support	Deny	Query	Comment	Macro F1
RoBERTa large+MLP [21]	0.48	0.55	0.60	0.93	0.64
Fine tune long former [22]	0.51	0.92	0.59	0.63	0.672
PSSNR	0.60	0.70	0.73	0.87	0.723




5. CONCLUSION

In this paper, we introduced a PSSNR framework that is composed of two main phases i) dataset preparation and ii) applying different ML algorithms with an augmented dataset. The dataset preparation is composed of 2 steps, data augmentation to solve the imbalanced dataset problem and text analysis as we observed that content features are the most effective features in stance prediction. We have done experiments before and after data augmentation using different ML algorithms and different combinations of features including conversation thread and user-based features. We conclude that data augmentation in deny and support classes and a specific combination of content and user features using feature selection improve the model's performance. Our proposed framework outperforms the state-of-the-art results from macro F1-score 0.672 to macro F1-score 0.7233.




REFERENCES

- [1] W. Ferreira and A. Vlachos, "Emergent: A novel data-set for stance classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1163–1168, doi: 10.18653/v1/N16-1138.
- [2] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 31–41, doi: 10.18653/v1/S16-1003.
- [3] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours," *Prepr. arXiv.1704.05972*, Apr. 2017.
- [4] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, pp. 38–55, Sep. 2019, doi: 10.1016/j.ins.2019.05.035.
- [5] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 2012, pp. 2451–2460, doi: 10.1145/2207676.2208409.
- [6] P. Tolmie *et al.*, "Supporting the use of user generated content in journalistic practice," in *Proceedings of the 2017 CHI Conference on Computing Systems*, May 2017, pp. 3632–3644, doi: 10.1145/3025453.3025892.
- [7] M. Broersma and T. Graham, "Social media as beat: Tweets as a news source during the 2010 British and Dutch elections," *Journalism Practice*, vol. 6, no. 3, pp. 403–419, Jun. 2012, doi: 10.1080/17512786.2012.663626.
- [8] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie, "Towards detecting rumours in social media," *Prepr. arXiv.1504.04712*, Apr. 2015.
- [9] G. Gorrell *et al.*, "SemEval-2019 Task 7: RumourEval, determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Sep. 2019, pp. 845–854, doi: 10.18653/v1/S19-2147.
- [10] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *Prepr. arXiv.1605.01655*, May 2016.
- [11] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1589–1599.
- [12] I. Augenstein, A. Vlachos, and K. Bontcheva, "USFD at SemEval-2016 task 6: Any-target stance detection on Twitter with autoencoders," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 389–393, doi: 10.18653/v1/S16-1063.
- [13] L. Zeng, K. Starbird, and E. Spiro, "#Unconfirmed: Classifying rumor stance in crisis-related social media messages," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 1, pp. 747–750, Aug. 2021, doi: 10.1609/icwsm.v10i1.14788.
- [14] G. Zarrella and A. Marsh, "MITRE at SemEval-2016 Task 6: Transfer learning for stance detection," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 458–463, doi: 10.18653/v1/S16-1074.
- [15] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, and M. Lukasik, "Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Sep. 2016, pp. 2438–2448.
- [16] E. Kochkina, M. Liakata, and I. Augenstein, "Turing at SemEval-2017 Task 8: Sequential approach to rumour stance classification with branch-LSTM," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 475–480, doi: 10.18653/v1/S17-2083.
- [17] H. Bahuleyan and O. Vechtomova, "UWaterloo at SemEval-2017 Task 8: Detecting stance towards rumours with topic independent features," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 461–464, doi: 10.18653/v1/S17-2080.
- [18] D. Chumachenko, I. Menailov, K. Bazilevych, T. Chumachenko, and S. Yakovlev, "Investigation of statistical machine learning models for COVID-19 epidemic process simulation: Random forest, K-nearest neighbors, gradient boosting," *Computation*, vol. 10, no. 6, May 2022, doi: 10.3390/computation10060086.
- [19] A. Kesarwani and P. M. Khilar, "Development of trust based access control models using fuzzy logic in cloud computing," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1958–1967, May 2022, doi: 10.1016/j.jksuci.2019.11.001.
- [20] R. Yang, W. Xie, C. Liu, and D. Yu, "BLCU_NLP at SemEval-2019 Task 7: An inference chain-based GPT model for rumour evaluation," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 1090–1096, doi: 10.18653/v1/S19-2191.
- [21] A. Prakash and H. T. Madabushi, "Incorporating count-based features into pre-trained models for improved stance detection," *Prepr. arXiv.2010.09078*, Oct. 2020.
- [22] A. Khandelwal, "Fine-tune longformer for jointly predicting rumor stance and veracity," in *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, Jan. 2021, pp. 10–19, doi: 10.1145/3430984.3431007.
- [23] F. Wang, M. Lan, and Y. Wu, "ECNU at SemEval-2017 Task 8: Rumour evaluation using effective features and supervised ensemble models," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 491–496, doi: 10.18653/v1/S17-2086.
- [24] R. Villa-Cox, S. Kumar, M. Babcock, and K. M. Carley, "Stance in replies and quotes (SRQ): A new dataset for learning stance in Twitter conversations," *Prepr. arXiv.2006.00691*, May 2020.
- [25] C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier, "Will-they-won't-they: A very large dataset for stance detection on Twitter," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1715–1724, doi: 10.18653/v1/2020.acl-main.157.
- [26] E. W. Pamungkas, V. Basile, and V. Patti, "Stance classification for rumour analysis in Twitter: Exploiting affective information and conversation structure," *Prepr. arXiv.1901.01911*, Jan. 2019.
- [27] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/icwsm.v8i1.14550.




BIOGRAPHIES OF AUTHORS

Kholoud Khaled    received my BSc in Computer Science in 2015 from Cairo University and is currently a teaching assistant at the Faculty of Computers and Artificial intelligence, Cairo University since 2016 I taught many subjects in programming, software engineering and artificial intelligence. I have been working in the software development industry for 6 years using a variety of tools and technologies. Her research interests include social network analysis, machine learning, deep learning, and big data. She can be contacted at email: k.khaled@fci-cu.edu.eg.



Abeer Elkorany    received Ph.D. in Electronics and Communications Engineering, Feb 2002, M.Sc. in Electronics and Communications Engineering (EE) May 1996, B.S. in EE '92 from Cairo University, faculty of engineering with honor degree. With more than 60 publications in local and international periodicals and conferences, which qualified her to obtain the scientific publication award many times from Cairo University. Research interests of Prof Abeer includes quality assurance and knowledge-based system measurement, semantic network, ontology development, knowledge management, social network analysis, and recommendation systems. Prof Abeer had supervised several international and national research projects. She can be contacted at email: a.korani@fci-cu.edu.eg.



Cherry A. Ezzat    is currently a full-time Assistant Professor at the Faculty of Computer Science and Artificial Intelligence, Cairo University. She obtained her PhD degree in Computer Science in 2016, from the Faculty of Computer Science and Artificial Intelligence (FCAI), Cairo University, in the field of social network analysis. She got her MSc and BSc from FCAI in 2006 and 2001, respectively. Her research interests are focused on artificial intelligence, social network analysis, and machine learning. She can be contacted at email: c.ahmed@fci-cu.edu.eg.