sCambiaMenti at ACTI: Ensemble model with majority voting for Automatic Conspiracy Detection

Selene Bianco[1], Daniela Salusso[2]

[1]AizoOn Consulting S.r.l.
[2]University of Turin

Abstract

English. In this paper we describe the methodology that we have implemented to solve the subtask A of the Automatic Conspiracy Detection (ACTI) task (EVALITA 2023). We have developed different classifiers and then used a majority voting approach to obtain the final prediction. The implemented classifiers can be distinguished into three different types: Machine Learning models trained on a document-term matrix (Support Vector Machines, Random Forest and Multinomial Näive Bayes), Neural Network models trained on the text sequences (Long Short Term Memory), and Machine Learning models trained on a set of linguistic features derived from the text. While the single models were prone to overfitting, the classification obtained with the voting approach appears to be more stable and showed an adequate performance on the official test set of the contest.

Italian. *In questo articolo descriviamo la metodologia implementata per risolvere il subtask A del task Automatic Conspiracy Detection (ACTI) task (EVALITA 2023). Abbiamo sviluppato diversi classificatori, per poi ottenere una classificazione finale tramite un meccanismo di voto a maggioranza. I classificatori implementati possono essere distinti in tre gruppi: modelli di tipo Machine Learning allenati su una rappresentazione documento-termine del testo (Support Vector Machines, Random Forest and Multinomial Näive Bayes), modelli di tipo Neural Network allenati sulle sequenze di testo (Long Short Term Memory) e modelli di tipo Machine Learning allenati su un insieme di caratteristiche linguistiche ricavate dal testo. Mentre i singoli modelli sono molto suscettibili all'overfitting, la classificazione ottenuta con il meccanismo di voto sembra essere più stabile ed ha mostrato una performance adeguata sul test set ufficiale del contest.*

Keywords
Conspiracy Theory, Content Moderation, Large Language Models, Computational Social Science

1. Introduction

With the proliferation of misinformation across multiple platforms and channels, detecting conspiracy theories and fake news has become a crucial task to ensure public safety and preserve democratic discourse. There have been efforts to develop Natural Language Processing (NLP) techniques to identify whether a given text contains fake news content [1] or not. The Automatic Conspiracy Theory Identification (ACTI) [2] task A in the EVALITA 2023 competition [3] aims to investigate different approaches to detect conspiracy theories from messages shared in platforms with lax moderation policies (like Telegram, 4chan, and Parler).

Different kinds of models are able to capture different aspects of a text and define suitable predictors for conspiracy messages.

Machine Learning (ML) models based on a matrix representation of terms in documents (document-term matrix) only "understand" which terms are used in a sentence. Therefore, they are not able to capture all features that can characterize a conspiratorial text.

Moreover, Artificial Neural Network (ANN) models does not only consider the presence of single words, but they are also able to estimate patterns and relationships among sequences of tokens in a sentence.

ML classifiers can also be trained on sets of linguistic features that can be predominantly present in fake content [4]. In literature, conspiratorial texts tends to display some particular linguistic properties [5] like both shortest and longest sentence, higher volume of punctuation marks, more frequent use of exclamation and question marks, a prominent use of adverbs [5, 6], predicative and attribute adjectives [7], capitalization and interjections [8, 9]. Emoji also seems to be more popular within fake news compared to true news [10], since they are used to increase persuasiveness of electronic communication [11].

Fake news corpus also tends to display also grammatical mistakes and inconsistencies, problems with sentence structure and an incorrect use of punctuation [5]. However, these latter features are more likely to be detected by a qualitative rather than quantitative analysis.

2. Related Work

Moderation of Fringe Communities. Recently, moderation of online fringe communities has become a pressing concern for mainstream platforms. Numerous works

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
selene.bianco@aizoongroup.com (S. Bianco);
daniela.salusso@unito.it (D. Salusso)

focused on the the efficacy of "deplatforming" these online communities [12, 13] as a way to limit the spreading of toxic ideas on mainstream platforms. However, when content moderation policies are applied users often migrate to alternative *fringe platforms*, sometimes created exclusively to host the banned community [14, 15]. Banning, in that context, would not only strengthen the infrastructure hosting these fringe platforms [12] but allow these communities to become more toxic elsewhere [16]. Indeed, such moderation policies do not avoid these fringe communities to grow and spread misinformation, for instance, about conspiratorial content. Therefore developing automated models to identify the spreading of such dangerous ideas is of the utmost importance.

Automated Methods for Identification of Toxic Content. Recent advances of natural language processing [17, 18, 19, 20] provides tools to address questions related to the identification of troublesome content online. Clearly traditional methods utilizing surface level features [21, 22] have opened the path to automatic detection of hate speech and fake news. Word embeddings [23] and recurrent neural networks [24, 25] significantly increased the prediction abilities of models to identify troublesome content. Finally, more recent systems based on transformers architectures [17, 18] have improved prediction accuracy among numerous tasks in different fields spanning from politics [26, 27], conflict prediction [28], and, of course, hate speech detection [29, 30, 31].

3. Description of the System

In order to evaluate different aspects of the text, different models have to been tested on a training-validation (70%-30%) split on the labeled training dataset. Since its dimension for task A is 1842 records, models shown a high tendency to overfit. Once the more performing models and their parameters were selected, the final models were trained again on the entire training set in order to predict labels on the test set (460 records).

In order to limit this problem and build a more stable system, we have defined a voting ensemble machine learning model combining the predictions from different models able to capture different features from the text. Therefore, the final ensemble classification represents the majority voting of the single predictions from each model.

3.1. ML models

Several ML models have been tested. The ones that performed better and have been selected as part of the voting approach were: Support Vector Machines (SVM) with radial basis function (rbf), Random Forest (RF) with maximum depth equal to 50 and Multinomial Näive Bayes (MNB).

In order to fit ML models, the training data has to be preprocessed and returned in the form of a document-term matrix. After cleaning the text from punctuation and stopwords, we tried to fit each model both on lemmatized and original tokens, once with only words and another with emoji too.

3.2. ANN model

The ANN model implemented is a Long Short-Term Memory (LSTM) with the structure as described in table 1.

3.3. ML model based on syntactical features

The latter model is trained on a set of syntactical features that have been extracted from the texts. The defined features were: the number of tokens and the percentages of punctuation, emoji, uppercase words, adjectives, adverbs, and interjections in the text. All the percentages had as denominators the sum of the number of tokens, punctuation signs, and emoji. The emoji were identified in the text with the emoji library. The Italian flag emoji was not recognized which this library and then identified in a second step. The Part Of Speech (POS) tagging used to count the number of adjectives, adverbs and interjections was performed with the Spacy's pipeline "it_core_news_sm".

We have also tried to identify the percentage of misspelled words with the phunspell library. However, the result was not entirely accurate and we decided to discard this feature for this reason.

We have first tested the significance of each feature with a probit logistic regression model and then used the significant ones as predictors in a RF model with maximum depth equal to 13.

4. Results and Discussion

The syntactical features used in the RF model are the ones that were significant in the logistic model (see table 2).

The logistic regression showed that a rise in the number of tokens and the proportion of punctuation and of uppercase words in the text is slightly increasing the probability of the text being of conspiracy type. These associations were in line with what was found in the literature [5]. On the contrary, the negative association of the proportion of emojis [10], adjectives [6], adverbs [5] and interjections [9] seems to be in contrast with what previously found for fake news in the English language.

Layer (type)	Output Shape	Param
embedding (Embedding)	(None, 645, 70)	869960
lstm (LSTM)	(None, 645, 100)	68400
pooling (GlobalMaxPooling1D)	(None, 100)	0
dense (Dense)	(None, 2)	202

Table 1
LSTM model summary.

feature	coef	std	z	P> $\|z\|$	[0.025, 0.975]
number of tokens	0.0036	0.000	7.700	0.000	0.003, 0.005
% punct	0.4955	0.240	2.068	0.039	0.026, 0.965
% uppercase words	1.8740	0.427	4.385	0.000	1.036, 2.712
% emoji	-3.2503	0.875	-3.716	0.000	-4.964, -1.536
% adjectives	-1.3083	0.460	-2.844	0.004	-2.210, -0.407
% adverbs	-2.1999	0.500	-4.400	0.000	-3.180, -1.220
% interjections	-11.7131	4.887	-2.397	0.017	-21.292, -2.134

Table 2
Significant syntactical features in the logistic regression.

Fitting the model on the text containing the emojis does not seem to give better results than fitting the model on the text alone. SVM and MNB models gave better results with the original tokens, while RF works better with lemmatized tokens. On the train-validation split, RF and SVM obtained higher precisions, whereas MNB higher recall.

The LSTM model was highly prone to overfitting starting from the second epoch. The f-1 scores of the tested models on the final test set are reported in table 3.

Limitations and further developments

Although multilingual frameworks for fake news detection are starting to be developed, a significant efficiency gap in modeling exists between English and languages other than English [32].

The model based on linguistic features should be improved by further detailing of the POS categories (adjectives, adverbs and interjections) in subcategories to better identify key predictors for fake news. A deeper analysis should be performed on emojis too, since they are known to play a role in the spread of misinformation by appealing to emotions [10]. The proportion of emojis in the text seems to be associated with an increased probability of the text not being of conspiracy type probably because emoji can express different kinds of feelings and should be divided into subtypes too.

Moreover, explainable AI (XAI) techniques could be further identified to underline specific tokens or other elements in the text influencing the model to classify a text as conspiracy or not.

The corpus provided for the ACTI task would have benefited from a data augmentation technique [33] or an integration from other corpora. However, we have not worked on this side, nor tested techniques of bootstrap or boosting. In addition, we have not tried to use more sophisticated models like the ones exploiting attention mechanism, both because of the size of the training set and because the fine-tuning of large language models would have required large GPU resources.

Ethics Statement

As we explore the potential benefits and limitations of using artificial intelligence (AI) to detect and categorize online conspiracy theories, it is important to consider the broader implications and risks involved in this area of research. While AI holds great promise for helping us better understand the spread of misinformation online and assist human moderators in identifying problematic content, it must not be deployed without considering its limitations in terms of accuracy of the results. Moreover, we need to remember that models trained on textual data can be potentially at risk of privacy leakage by adversarial attacks [34].

References

[1] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp.

Model	F1-score
RF	0.7538
MNB	0.7147
SVM rbf	0.7573
LSTM	0.7422
RF with linguistic features	0.7007
Majority voting	0.7918

Table 3
F1-score of the single models and of the ensemble on the test data.

6086–6093. URL: https://aclanthology.org/2020.lrec-1.747.

[2] G. Russo, N. Stoehr, M. H. Ribeiro, Acti at evalita 2023: Overview of the conspiracy theory identification task, 2023. arXiv:2307.06954.

[3] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3391–3401. URL: https://aclanthology.org/C18-1287.

[5] R. Sousa-Silva, Fighting the fake: A forensic linguistic analysis to fake news detection, International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique 35 (2022) 2409 – 2433. doi:10.1145/322234.322243.

[6] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://aclanthology.org/D17-1317. doi:10.18653/v1/D17-1317.

[7] J. Grieve, H. Woodfield, The Language of Fake News, Elements in Forensic Linguistics, Cambridge University Press, 2023. doi:10.1017/9781009349161.

[8] A. Aich, S. Bhattacharya, N. Parde, Demystifying neural fake news via linguistic feature-based interpretation, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6586–6599. URL: https://aclanthology.org/2022.coling-1.573.

[9] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of Economic Perspectives 31 (2017) 211–36. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211. doi:10.1257/jep.31.2.211.

[10] S. Suntwal, L. Brandimarte, S. A. Brown, Understanding the role of nonverbal tokens in the spread of online information, Proceedings of the 56th Hawaii International Conference on System Sciences (2023) 5484–5493. URL: https://hdl.handle.net/10125/103303.

[11] T. Maiberger, D. Schindler, N. Koschate-Fischer, Let's face it: When and how facial emojis increase the persuasiveness of electronic word of mouth, Journal of the Academy of Marketing Science (2023). doi:10.1007/s11747-023-00932-8.

[12] E. Zuckerman, C. Rajendra-Nicolucci, Deplatforming our way to the alt-tech ecosystem, Knight First Amendment Institute at Columbia University, January 11 (2021).

[13] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[14] C. Dewey, Washington Post — These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy, and why it banned them, https://wapo.st/3AO7pbl, 2016.

[15] G. Russo, M. Horta Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, in: Proceedings of the 15th ACM Web Science Conference 2023, WebSci '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 251–259. URL: https://doi.org/10.1145/3578503.3583608. doi:10.1145/3578503.3583608.

[16] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, R. West, Do platform migrations compromise content moder-

ation? evidence from r/the_donald and r/incels, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–24.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[19] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, arXiv preprint arXiv:1906.08976 (2019).

[20] G. Russo, N. Hollenstein, C. C. Musat, C. Zhang, Control, generate, augment: A scalable framework for multi-attribute text generation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 351–366. URL: https://aclanthology.org/2020.findings-emnlp.33. doi:10.18653/v1/2020.findings-emnlp.33.

[21] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[22] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, S. Mishra, Analyzing labeled cyberbullying incidents on the instagram social network, in: Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7, Springer, 2015, pp. 49–66.

[23] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 302–308.

[24] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.

[25] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[26] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Disentangling active and passive cosponsorship in the u.s. congress, ArXiv abs/2205.09674 (2022).

[27] J. Valvoda, T. Pimentel, N. Stoehr, R. Cotterell, S. Teufel, What about the precedent: An information-theoretic analysis of common law, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2275–2288. URL: https://aclanthology.org/2021.naacl-main.181. doi:10.18653/v1/2021.naacl-main.181.

[28] M. Zhong, S. Dhuliawala, N. Stoehr, Extracting victim counts from text, arXiv preprint arXiv:2302.12367 (2023).

[29] P. Alonso, R. Saini, G. Kovács, Hate speech detection using transformer ensembles on the hasoc dataset, in: Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings, Springer, 2020, pp. 13–21.

[30] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020).

[31] L. Stappen, F. Brunn, B. Schuller, Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel, arXiv preprint arXiv:2004.13850 (2020).

[32] R. Mohawesh, S. Maqsood, A. Qutaibah, Multilingual deep learning framework for fake news detection using capsule neural network, Journal of Intelligent Information Systems (2023) 1–17. doi:10.1007/s10844-023-00788-y.

[33] A. Keya, M. A. Wadud, M. Ph. D., M. Alatiyyah, M. A. Hamid, Augfake-bert: Handling imbalance through augmentation of fake news using bert to enhance the performance of fake news classification, Applied Sciences 12 (2022) 8398. URL: https://www.mdpi.com/2076-3417/12/17/8398. doi:10.3390/app12178398.

[34] L. Song, X. Yu, H.-T. Peng, K. Narasimhan, Universal adversarial attacks with natural triggers for text classification, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3724–3733. URL: https://aclanthology.org/2021.naacl-main.291. doi:10.18653/v1/2021.naacl-main.291.