GreekPolitics: Sentiment Analysis on Greek Politically Charged Tweets

1st Emmanouil Patsiouras
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
emmanoup@csd.auth.gr

2nd Ioanna Koroni
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
koroniioanna@csd.auth.gr

3rd Ioannis Mademlis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
imademlis@csd.auth.gr

4th Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
pitas@csd.auth.gr

Abstract—The rapid growth of on-line social media platforms has rendered opinion mining/sentiment analysis a critical area of Natural Language Processing (NLP) research. This paper focuses on analyzing Twitter posts (tweets), written in the Greek language and politically charged in content. This is a rather underexplored topic, due to the scarcity of publicly available annotated datasets. Thus, we present and release "GreekPolitics", i.e., a dataset of Greek tweets with politically charged content, independently annotated across four different sentiment dimensions: polarity, figurativeness, aggressiveness and bias. GreekPolitics has been evaluated comprehensively in a classification setting, separately for each sentiment, using state-of-the-art Deep Neural Networks (DNNs) and data augmentation methods. This paper details the dataset, the evaluation process and the experimental results. Based on these, best practices are identified for achieving the highest classification accuracy at the test stage.

Index Terms—Sentiment analysis, natural language processing, social media analysis, deep neural networks, opinion mining, data augmentation, computational politics

I. INTRODUCTION

The broad adoption of on-line social media platforms has made it easy for a great number of people to express their opinion upon a certain matter. The vast amount of textual content generated daily can be exploited by Natural Language Processing (NLP) algorithms, in order to model and predict user behavior or preferences [1] [2]. A critical relevant task is *sentiment analysis* or *opinion mining* [3] [4]: the classification of a text with respect to a certain type of sentiment it expresses, (e.g., 3-class classification into "positive", "negative" or "neutral"). Twitter is a social media platform that offers abundant user sentiments in textual form. Thanks to the spread of smartphones, people utilize Twitter to express their opinions for any number of socially relevant topics. Twitter posts ("tweets") are very short texts and typically involve no more than a sentence.

One topic popular in Twitter is politics. Sentiment analysis algorithms applied on politically charged tweets can automatically draw conclusions, such as sympathy or likeness indices towards certain political parties, or may even predict the outcome of forthcoming elections [5] [6]. Modern sentiment analysis is performed using machine learning methods, with the state-of-the-art being Deep Neural Networks (DNNs) [7].

Assuming that each tweet is a different text, the simplest approach is to equate sentiment with opinion and simplify opinion as falling within a "positive"-to-"negative" spectrum (i.e., *polarity*). However, alternative dimensions of sentiment can also be extracted, potentially overlapping or complementary. For instance, a text can be characterized as "ironic"/"sarcastic" or "literal" [8]; as "offensive"/"racist"/"abusive" or not [9], etc. Machine learning models, such as DNNs, can be trained on datasets annotated with similar labels and, subsequently, be employed for analyzing novel tweets.

Sophisticated sentiment analysis would arguably require *all* of these properties (e.g., polarity, sarcasm, etc.) to be extracted. This is not typically possible because popular relevant datasets for training machine learning models are annotated only along one or two of these dimensions. This issue is even more pronounced in languages other than English, where the available annotated data are much smaller in size. Thus, this paper proposes a sentiment definition as a combination of four individual sentiment dimensions: **polarity**, **figurativeness**, **aggressiveness** and **bias**. A sample of Greek-language tweets with political content has been collected, annotated according to our definition and employed for training DNN models on sentiment analysis. This is doubly important, since: a) *there is a relative lack of similar datasets in Greek*, and b) *most published opinion mining results in Greek have been obtained with outdated machine learning methods*.

Thus, this paper contributes a new dataset consisting of 2,578 unique tweets, gathered through the official Twitter API, with politically charged content in Greek and annotated according to the four employed sentiment dimensions. The

results of extensive evaluation with state-of-the-art DNNs on the introduced dataset are thoroughly discussed, along with identified best practices for maximizing test classification accuracy. The generated GreekPolitics dataset is freely available on-line at https://aiia.csd.auth.gr/auth-greekpolitics-dataset/.

II. RELATED WORK

The vast majority of sentiment analysis research, including tweet analysis, concerns English. In contrast, this work focuses on sentiment analysis of Greek tweets. An early relevant dataset [10] was composed of over four million tweets, spanning a number of different topics, but only 0.00015% of them were annotated for the sentiments of anger, disgust, fear, happiness, sadness, and surprise. Recently, [11] was exploited for extracting user characteristics, a task partially unconnected to conventional sentiment analysis.

Studies [12] and [13] both focused solely on the polarity dimension and utilized traditional/outdated Natural Language Processing (NLP) tools, combined with classic machine learning algorithms (e.g., random forest, decision trees, Support Vector Machine classifiers). In contrast, [14] focused on offensiveness and exploited DNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. Finally, pioneering research concerning irony identification in politically charged Greek tweets was conducted in [15] and [16], but only employing outdated probabilistic classifiers.

III. A NEW GREEK POLITICAL-RELATED DATASET FOR SENTIMENT ANALYSIS

This Section presents in detail the GreekPolitics dataset.

A. Data gathering

GreekPolitics was designed as a large-scale dataset of politically charged Greek tweets, accompanied by full ground-truth labels along the proposed 4 sentiment dimensions. This multidimensional sentiment definition is expected to allow fuller semantic characterization of a tweet.

GreekPolitics Twitter posts were collected based on specific query hashtags related to the Greek political scene, using the official Twitter API. These hashtags are mainly related to the names of the various political parties and popular politicians represented in the Greek parliament over the past decade, while variants of them (in both the Greek and the Latin alphabet) were also exploited. Indicative query hashtags are presented in Table I. Collected tweets span a large time scale, from January 2014 up to March 2021.

After an initial data cleaning stage, we ended up with over 8,000 tweets. After removing retweets, duplicates or poorly written posts, the final dataset contained 2,578 unique tweets.

B. Data annotation

Based on the proposed 4-dimensional sentiment definition, each individual tweet was independently annotated for classification with respect to polarity, figurativeness, aggressiveness and bias. As far as *polarity* is concerned, each tweet was assigned one out of three possible class labels: "positive", "negative" or "neutral". For *figurativeness*, each tweet was assigned

TABLE I
INDICATIVE QUERY HASHTAGS USED TO COLLECT THE CONTENT OF GREEKPOLITICS.

Greek political parties names:
#Νεα_Δημοκρατια, #ΣΥΡΙΖΑ, #ΠΑΣΟΚ, #ΚΙΝΑΛ, #ΚΚΕ, #Χρυση_Αυγη, #Ελληνικη_Λυση, #Μερα25 #Ανεξαρτητοι_Ελληνες, #Ενωση_Κεντρωων
Greek politician names:
#Μητσοτακης, #Αντωνης_Σαμαρας, #Τσιπρας, #Γιωργος_Παπανδρεου, #Κουτσουμπας, #Καμμενος, #Φωφη_Γεννηματα, #Παπαρηγα, #Μιχαλολιακος, #Βελοπουλος, #Λεβεντης
Greek politically charged words:
#Δημοψηφισμα, #Εκλογες, #Μνημονιο, #Κυβερνηση

the ground-truth label of either "figurative" (ironic, sarcastic or figurative in general) or "normal" (i.e., non-figurative, literal). Regarding *aggressiveness*, each tweet was annotated with either an "aggressive" (offensive, abusive, racist or aggressive in general) or "normal" (i.e., non-aggressive) label. Finally, for *bias*, each tweet was annotated as either a "partisan" (if its expressed a strong, supportive and adamant opinion) or a "neutral" (i.e., non-partisan) one.

Therefore, each tweet was manually and independently annotated with: i) 1 label for a 3-class classification task, and ii) 3 different labels for 3 binary classification tasks (one per task).

TABLE II
NUMBER OF ANNOTATED TWEETS PER SENTIMENT CLASS.

Polarity			
Positive	Negative	Neutral	Total
156	589	1833	2578
Figurativeness			
Figurative	Non-figurative		Total
1257	1321		2578
Aggressiveness			
Aggressive	Non-aggressive		Total
438	2140		2578
Bias			
Partisan	Non-partisan		Total
1562	1016		2578

Manually annotating a tweet with the employed four labels (one per dimension) may potentially lead to different results for each individual human annotator. In order to provide as

objective ground-truth annotations as possible, a team of three volunteers was asked to classify each tweet in the dataset with respect to the classes of each different sentiment dimension. Inter-annotator agreement was subsequently calculated and annotations with majority agreement were selected as the actual annotations of the tweets in question. The resulting class split for each sentiment is presented in Table II

Table II shows a huge inter-class imbalance with regard to the ground-truth class size: in the case of polarity, positive tweets are significantly less than the negative or neutral ones. This is because most Twitter users tend to express rather negative or neutral opinions regarding politics. This effect may cause great issues when training and evaluating machine learning models, since a classifier could undesirably learn to favour at the test stage those classes that were the most large-sized during training. Certain measures taken to compensate for this issue are described in Section IV.

C. Data preprocessing

After collection and annotation of tweets, they were pre-processed in order to produce the final GreekPolitics dataset. While reading and comprehending a tweet's content is done effortlessly by a human, providing raw input to machine learning models is not ideal. Tweets may include undesirable content, such as hashtags, URLs or emojis, which could impede successful sentiment analysis of their actual text. Thus, the following preprocessing steps were followed:

- Remove all mentions (i.e., text starting with @), URLs, emojis or any other special character. Since hashtags may be quite meaningful, they were retained: Table VII presents the top forty most frequently appearing ones.
- Strip accents. Greek words usually contain accents which were striped.
- Remove all punctuation. Punctuation marks do not help us discriminate between different text sentiments.
- Remove multiple spaces and line breakers, so that each tweet is expressed in one line and each word is separated by a single space.
- Convert all words to lower case in order to achieve data uniformity and to avoid ambiguity.
- Tokenize the sentences. Tokenization is the process of splitting a piece of text into smaller units called tokens. For GreekPolitics, each text was tokenized by considering the words as splitting tokens. For example, the sentence ["βολες αχτσιογλου κατα της κυβερνυσης"] is tokenized as ["βολες","αχτσιογλου","κατα", "της", "κυβερνησης"].

IV. EXPERIMENTS

This Section presents and discusses the experimental evaluation conducted on the introduced GreekPolitics dataset.

A. Model selection

Four different classifiers were trained separately for the four classification tasks (i.e., sentiment dimensions). Polarity analysis was addressed as a 3-class classification task, while binary classifiers were employed for the remaining three sentiment dimensions. Two different DNNs were investigated as alternative options: i) a Convolutional Neural Network (CNN) adopted from [17], and ii) a Transformer adopted from [18]. The CNN has 5 1D convolutional layers, each one with an increasing amount of output filters. The Transformer has an encoder module composed of a multi-head self-attention layer, a normalization layer with a residual connection, two fully-connected layers and a final normalization layer with a residual connection. The output of both the CNN and the Transformer is fed to a fully-connected classification layer with as many neurons as the number of classes.

A pretrained FastText [19] word embedding DNN was first utilized for transforming a given word into a unique 300-dimensional vector[1]. Subsequently, the resulting vector representations of all words in a tweet were concatenated into a matrix $T \in \mathbb{R}^{M \times 300}$, where M is chosen as the maximum sentence length in terms of word count. If a sentence does not exceed the maximum length T is appropriately padded with zero value vectors. Given the text of the specific tweets in GreekPolitics, M was set to 60.

80%/20% of the GreekPolitics tweets were used for the training/test set, respectively. DNNs were trained separately for each sentiment dimension, using random parameter initialization, the Adam optimizer, a categorical cross-entropy loss, a learning rate of 0.001, a mini-batch size of 64 and a total of 60 epochs.

B. Results

This Subsection presents: a) initial performance comparisons in the original dataset split (Table II), and b) refined experimental results, after employing common tricks for increasing classification accuracy on certain under-performing tasks.

1) Polarity: Table III reports results on 3-class polarity classification, both on the overall test set and on each individual class. The huge class imbalance (i.e., the "positive" contains only 6.1% of the GreekPolitics tweets) leads unsurprisingly to poor class-specific performance. Thus, Table III reports evaluation results based on the best measured performance on the minority classes, as wells as precision and recall metrics. Initial comparisons in the original dataset split are showcased in the top section of Table III (i.e., the models with the suffix 1). The superiority of the Transformer against the CNN and the unacceptable accuracy in the minority classes are evident.

In order to improve accuracy in the minority classes, data augmentation was applied. Ideally, new tweets should be generated based on the existing ones, that would maintain identical ground-truth sentiment but would be expressed in a different manner. Two text augmentation methods were applied:

1) Back-translation [20]. A sentence is translated on another language, e.g., from Greek to English, and then back to the source language. If the new generated sentence is different but holds the exact same meaning, it is used as an augmented version of the original one.

[1]https://fasttext.cc/docs/en/crawl-vectors.html

TABLE III
INDIVIDUAL CLASS ACCURACIES, OVERALL ACCURACY AND OVERALL PRECISION/RECALL METRICS FOR POLARITY CLASSIFICATION. THE ✓ SYMBOL INDICATES THAT THIS METHOD WAS APPLIED FOR TRAINING THE RESPECTIVE MODEL.

Model	Augment.	Pruning	Acc. positive	Acc. negative	Acc. neutral	Overall Acc.	Precision	Recall
CNN-1	-	-	0.24	0.42	0.90	0.76	0.75	0.73
Transformer-1	-	-	**0.39**	**0.67**	**0.89**	**0.81**	**0.805**	**0.806**
CNN-2	✓	-	0.57	0.61	0.81	0.78	0.77	0.73
Transformer-2	✓	-	**0.62**	**0.74**	**0.89**	**0.83**	**0.84**	**0.83**
CNN-3	✓	✓	0.74	0.67	0.79	0.74	0.75	0.74
Transformer-3	✓	✓	**0.83**	**0.78**	**0.82**	**0.81**	**0.83**	**0.82**

TABLE IV
INDIVIDUAL CLASS ACCURACIES, OVERALL ACCURACY AND OVERALL PRECISION/RECALL METRICS FOR FIGURATIVENESS CLASSIFICATION.

Model	Acc. figurative	Acc. non-figurative	Overall Acc.	Prec.	Recall
CNN	0.60	0.71	0.66	0.64	0.68
Transf.	**0.70**	**0.75**	**0.72**	**0.70**	**0.72**

TABLE V
INDIVIDUAL CLASS ACCURACIES, OVERALL ACCURACY AND OVERALL PRECISION/RECALL METRICS FOR AGGRESSIVENESS CLASSIFICATION.

Model	Aug.	Acc. aggressive	Acc. non-aggressive	Overall Acc.	Pre.	Rec.
CNN-1	-	0.52	0.86	0.69	0.66	0.68
Transf.-1	-	**0.55**	**0.96**	**0.88**	**0.86**	**0.88**
CNN-2	✓	0.64	0.85	0.75	0.74	0.74
Transf.-2	✓	**0.69**	**0.91**	**0.86**	**0.85**	**0.86**

TABLE VI
INDIVIDUAL CLASS ACCURACIES, OVERALL ACCURACY AND OVERALL PRECISION/RECALL METRICS FOR BIAS CLASSIFICATION.

Model	Aug.	Acc. partisan	Acc. non-partisan	Overall Acc.	Pre.	Rec.
CNN-1	-	0.83	0.44	0.68	0.65	0.66
Transf.-1	-	**0.85**	**0.67**	**0.77**	**0.70**	**0.72**
CNN-2	✓	0.84	0.52	0.70	0.69	0.69
Transf.-2	✓	**0.86**	**0.77**	**0.81**	**0.79**	**0.81**

TABLE VII
THE 40 MOST FREQUENTLY APPEARING HASHTAGS (APPEARANCE FREQUENCY IN PARENTHESES).

#συριζα(879), #νδ(504), #μητσοτακης(439), #τσιπρας(367), #κυβερνηση_μητσοτακη(156), #ανελ(122), #ελλαδα(118), #νδ_ξεφτιλες(105), #κκε(101), #πολιτικη(93), #syriza(82), #συριζα_ξεφτιλες(76), βουλη(76), #κυβερνηση_τσιρκο(73), #πασοκ(73), #παρακρατος_τσιπρα(67), #αποστασια(66), #νδ_απατεωνες(62), #φουρθιωτης(59), #νδ_the_end(58), #παρακρατος(55), #tsipras(52), #νδ_χουντα(51), πασοκ(73), #εκλογες(50), #παππας(46), #μακεδονια(45), #κιναλ(45), #κυβερνηση_συμμορια(40), #vouli(39), #συριζα_βουρκος(37), #νδ_θελατε(32), #μμε_ξεφτιλες(31), #κυβερνηση(31), #σκοπιανο(30), #κυβερνηση_φουρθιωτη(29), #μιωνη(28), #ελληνωνσυνελευσις (29), #κορωνοιος(28), #καμμενος(28)

2) Synonym substitution [21][22]. In synonym substitution, given a certain probability, each word in a sentence is replaced with a synonym one acquired from an external vocabulary of synonyms, e.g., Thesaurus. The new sentence is again used as an augmented version of the original one.

Augmentation was applied only on 80% of the "positive" samples, i.e., the samples used for training, with the test set remaining untouched. Thus, out of the 156 "positive" training samples, additional 111 tweets were generated. After retraining from scratch on the augmented dataset for polarity classification, the results reported in the middle section of Table III (i.e., the models with the suffix 2) were obtained. As it can be seen, augmentation resulted in a huge increase in test accuracy for the "positive" class. However, the inter-class accuracy gap remains considerable (0.62/0.74/0.89 for "positive"/"negative"/"neutral", in the Transformer model).

To further reduce the inter-class accuracy gap, the "neutral" training samples were pruned, while the augmented "positive" ones were retained. Thus, out of the 1833 "neutral" samples, 737 were removed. The results are reported in the bottom section of Table III (i.e., the models with the suffix 3). It can be seen that the test accuracy for the minority classes increased significantly, with a small drop in the accuracy of the "neutral" class. This inter-class balancing effect under the presence of ground-truth class imbalances is relevant in application domains where it is more important to estimate the comparative/relative volumes of "positive", "neutral" and "negative" tweets, rather than accurately characterize a specific tweet.

2) Figurativeness: Figurativeness classification results are reported on Table IV. The ground-truth classes are rather balanced and the obtained accuracy is acceptable, so no training dataset processing was applied to boost test performance. The Transformer again surpasses the CNN.

3) Aggressiveness: Aggressiveness classification results are reported on Table V. A huge class imbalance was again observed, as the "offensive" tweets are few (i.e., 20% of the dataset). The augmentation strategies previously described for polarity classification were adapted and independently/separately applied: out of the 438 "offensive" training samples, an additional 256 augmented samples were generated. Initial comparison results for the original dataset are showcased in the top section of Table V (i.e., the models with the suffix 1). Results produced by retraining the employed models on the augmented scenario are reported in the bottom section of Table V (i.e., the models with the suffix 2). As before, augmentation leads to a massive increase in model accuracy for the minority class, while maintaining acceptable accuracy for the initially dominant class.

4) Bias: Bias classification results are reported on Table VI. Moderate data imbalance was observed here for the two classes (39% and 61%). To counter it, the augmented tweets already generated for the polarity and aggressiveness tasks (annotated as "non-partisan") were employed for retraining from scratch. Thus, 202 augmented training tweets were added to the original 1016 "non-partisan" ones. Initial comparison results in the original dataset are showcased in the top section of Table VI (i.e., the models with the suffix 1), while results for the augmented dataset are reported in the bottom section of Table VI (i.e., the models with the suffix 2).

V. CONCLUSIONS

This paper introduced the recently captured/annotated "GreekPolitics" dataset for sentiment analysis of politically charged Twitter posts in the Greek language. The tweets have been manually labelled for classification across 4 independent sentiment dimensions (polarity, figurativeness, aggressiveness and bias). A thorough experimental study was conducted by utilizing state-of-the-art Deep Neural Networks (DNNs), which yielded promising results. The domain-specific problem of data class imbalance (too few positive tweets) was tackled in the experimental evaluation by employing standard data augmentation tricks, based on generating novel tweet samples from the existing ones. Such methods can generate a limited amount of new training samples and should be validated by humans. Interesting future work could explore more sophisticated data augmentation for natural language text, which could ideally be applied without any human supervision.

REFERENCES

[1] D. Tsirmpas, I. Gkionis, and I. Mademlis, "Neural Natural Language Processing for long texts: A survey of the state-of-the-art," *arXiv preprint arXiv:2305.16259*, 2023.

[2] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–41, 2016.

[3] W. Medhat, A. Hassan, and Korashy H., "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering Journal*, vol. 5, pp. 1093–1113, 2014.

[4] R.-M. Karampatsis, J. Pavlopoulos, and P. Malakasiotis, "AUEB: Two stage sentiment analysis of social network messages," in *Proceedings of the International Workshop on Semantic Evaluation (SemEval 2014)*, 2014.

[5] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "Election result prediction using Twitter sentiment analysis," in *Proceedings of the International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2016, vol. 1.

[6] D. Karamouzas, I. Mademlis, and I. Pitas, "Public opinion monitoring through collective semantic analysis of tweets," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–21, 2022.

[7] C. Tzogka, N. Passalis, A. Iosifidis, M. Gabbouj, and A. Tefas, "Less is more: deep learning using subjective annotations for sentiment analysis from social media," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[8] D. Karamouzas, I. Mademlis, and I. Pitas, "Neural knowledge transfer for sentiment analysis in texts with figurative language," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2022.

[9] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," *arXiv preprint arXiv:1903.08983*, 2019.

[10] G. Kalamatianos, D. Mallis, S. Symeonidis, and A. Arampatzis, "Sentiment analysis of greek tweets and hashtags using a sentiment lexicon," in *Proceedings of the Panhellenic Conference on Informatics*, 2015.

[11] S. Baxevanakis, S. Gavras, D. Mouratidis, and K. L. Kermanidis, "A machine learning approach for gender identification of Greek tweet authors," in *Proceedings of the ACM International Conference on Pervasive Technologies Related to Assistive Environments*, 2020.

[12] A. Tsakalidis, S. Papadopoulos, R. Voskaki, K. Ioannidou, C. Boididou, A. I. Cristea, M. Liakata, and Y. Kompatsiaris, "Building and evaluating resources for sentiment analysis in Greek language," *Language resources and evaluation*, vol. 52, pp. 1021–1044, 2018.

[13] D Beleveslis, C Tjorjis, D. Psaradelis, and D. Nikoglou, "A hybrid method for sentiment analysis of election related tweets," in *South-East Europe Design, Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*. IEEE, 2019.

[14] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive language identification in Greek," *arXiv preprint arXiv:2003.07459*, 2020.

[15] B. Charalampakis, D. Spathis, E. Kouslis, and K. Kermanidis, "A comparison between semi-supervised and supervised text mining techniques on detecting irony in Greek political tweets," *Engineering Applications of Artificial Intelligence*, vol. 51, pp. 50–57, 2016.

[16] B. Charalampakis, D. Spathis, E. Kouslis, and K. Kermanidis, "Detecting irony on Greek political tweets: a text mining approach," in *Proceedings of the International Conference on Engineering Applications of Neural Networks (INNS)*, 2021.

[17] L. Zewen, L. Fan, Y. Wenjie, P. Shouheng, and P. Jun, "A survey of Convolutional Neural Networks: analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 6999–7019, 2021.

[18] A. Gillioz, J. Casas, E. Mugellini, and O. Abou K., "Overview of the Transformer-based models for NLP tasks," in *Proceedings of the Conference on Computer Science and Information Systems (FedCSIS)*. 2020, IEEE.

[19] P. Bojanowski, E. Grave, A. Joulin, and Mikolov. T, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[20] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2020.

[21] X Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2015.

[22] J. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.