Boosting Sentiment Analysis in Persian through a GAN-Based Synthetic Data Augmentation Method

Masoumeh Mohammadi
Fordham University, USA
mm256@fordham.edu

Shadi Tavakoli
Pars Tourism Card, Tehran, Iran
tavakoli.shadii@gmail.com

Mohammad Ruhul Amin
Fordham University, USA
mamin17@fordham.edu

Abstract

This paper presents a novel Sentiment Analysis (SA) dataset in the low-resource Persian language, including a data augmentation technique using Generative Adversarial Networks (GANs) to generate synthetic data, boosting the volume and variety of data for achieving state-of-the-art performance. We propose a novel annotated SA dataset, Senti-Persian, made of 67,743 public comments on movie reviews from Iranian websites (Namava, Filimo, and Aparat) and social media (YouTube, Twitter and Instagram). These reviews are labeled with one of the polarity labels, namely positive, negative, and neutral, by humans and later augmented. Our study includes a novel text augmentation model based on GANs. The generator was designed following the linguistic properties of Persian linguistics. In contrast, the discriminator was developed based on the cosine similarity of the vectorized original and generated sentences, i.e., using CLS-embeddings of BERT. An SA task was applied on both collected and augmented datasets, for which we observed a significant improvement in accuracy from 88.4% for the original dataset to 96% when augmented with synthetic data. The senti-Parsian dataset, including the original and the augmented ones, can be accessed on GitHub.[1].

1 Introduction

Using the World Wide Web allows us to access the languages we encounter daily. Even though the Web began as an overwhelmingly English phenomenon, it now contains texts in thousands of languages (Usa, 2021) (Int, 2012). The ability to combine prior knowledge with updated information across thousands of languages and to generate new patterns based on those languages is the most compelling reason for advancing language processing (van Kessel et al., 2019).

[1]https://github.com/engmahsa/Senti-Persian-Dataset

There is a unique opportunity for computational linguists now, as this field has unprecedented access to low-resource languages. However, researchers must act swiftly, as every few days, we lose another language from the face of the Earth due to the lack of native speakers. This loss is driven by complex political, social, racial, and economic factors. Thus, we must gather online resources and develop advanced language models to preserve these disappearing languages. By doing so, we can safeguard linguistic diversity and ensure that even endangered languages remain accessible and celebrated in the digital age (Her and Kruschwitz, 2024) (Tatineni, 2020).

Natural language processing (NLP) and computational linguistics (CL) primarily focus on languages with large text corpora. Machine learning (ML) techniques are usually used to train NLP tools, and lots of languages lack large annotated corpora for training (Hauer et al.) (Xu et al., 2022) (ImaniGooghari et al., 2023) (Zhao, 2022). Using natural language to mine opinions and sentiments is extremely challenging as it involves understanding how language structures convey explicit and implicit information in individual words or entire text (Bhatia et al., 2018) (Liu and Zhang, 2012).

The necessity of this article lies in addressing the challenges faced by NLP when dealing with low-resource languages. These challenges arise due to limited supervised data availability and a scarcity of native speakers or expert contributions. To overcome this obstacle, this paper introduces a data augmentation technique that leverages GANs to generate synthetic data. Doing so enhances the volume and variety of available data, which is particularly advantageous in fields where data acquisition is costly, such as low-resource languages like Persian.

This research significantly enhances the capabilities of NLP models for low-resource languages by introducing innovative methods and datasets. The

Processings of the 1st Workshop on NLP for Languages Using Arabic Script (AbjadNLP), pages 54–63

significant challenges we addressed while working for the low-resourced Persian language are mentioned below:

- Increased Data Diversity: This technique generates new comments by applying transformations (e.g., synonym replacement, paraphrasing) to existing movie reviews. This diversifies the dataset, making the model more robust to variations in language and context.

- Mitigation of Overfitting: By introducing synthetic examples, data augmentation helps prevent overfitting. It exposes the model to different linguistic patterns, reducing its reliance on specific training instances.

- Improved Generalization: Augmented data provides additional context and linguistic variations. Consequently, NLP models learn more generalized features, leading to better performance on unseen data.

- Addressing Low-Resource Scenarios: In languages with limited labeled data, augmentation generates synthetic samples, enabling practical training even when native speaker contributions are scarce.

- Enhanced Performance: Empirical results often show improved accuracy and robustness when applying data augmentation.

This paper contributes the following:

1. A labeled dataset for SA in Persian, Senti-Persian comprises three types of movie reviews: positive, negative, and neutral. This marks the first representation of user movie reviews in Persian within a dataset of 67,743 entries.

2. A cutting-edge GAN-based text generator is implemented to augment the comments.

3. In order to determine how accurate the models can be, resampling techniques are used on the set for balancing, and then evaluation metrics are compared.

4. A number of data augmentation methods are applied, including random insertion, synonym replacements, and random swaps, which also affect model accuracy.

Following is the organization of this paper: The summary of the related articles is included in Section 2. The structure of the proposed approach is described in Section 3. Section 4 presents the methodology. Section 5 discusses the results of our research and our plans for the future.

2 Related Work

The ParsiNLU (Khashabi et al., 2021) NLI database contains 2,700 instances, primarily written by native speakers, with some translated from the MultiNLI dataset (Williams et al., 2018). The FarsTail dataset, in comparison, has four times more native sentences than ParsiNLU. FarsTail uses fewer task-specific human-generated texts to create more natural-looking sentences. Methods for transferring knowledge across resource-limited languages are often employed. Studies like those by Dashtipour et al. (Dashtipour et al., 2021) have compared approaches to multilingual SA. Balahur and Turchi (Balahur and Turchi, 2012) found that translating training data between languages from the same family (Italian, French, Spanish) improves results.

Devlin et al. introduces Text AutoAugment (TAA), a data augmentation framework for text classification that uses Bayesian Optimization to find optimal augmentation policies. TAA outperforms manual methods, improving classification accuracy, especially in low-resource and imbalanced datasets, while reducing the need for prior knowledge and manual tuning. The paper (Karimi et al., 2021) introduces AEDA, using punctuation insertion, which improves text classification accuracy and outperforms previous methods like EDA across multiple datasets.

The article "DeepSentiPers" introduces two deep learning models, bidirectional LSTM and CNN, for Persian SA, using three data augmentation techniques to improve classification accuracy in both binary and multi-class tasks, advancing SA in low-resource languages (PourMostafa et al., 2020) (Sartakhti et al., 2022) enhances Persian relation extraction on the PERLEX dataset using text preprocessing and augmentation techniques, significantly improving accuracy with ParsBERT (Farahani et al., 2021) and Multilingual BERT models, addressing the resource scarcity in Persian NLP.

Mi et al. introduces a method using SMT and RNN to generate target-side paraphrases, significantly improving translation quality for low-

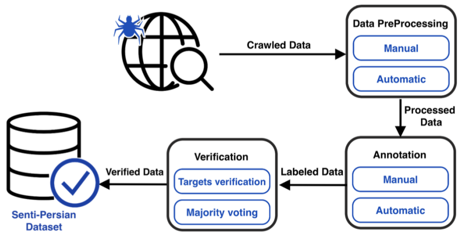

Figure 1: A flow diagram shows the four major phases of Senti-Persian's development: data crawling, preprocessing, data annotation, and label verification.

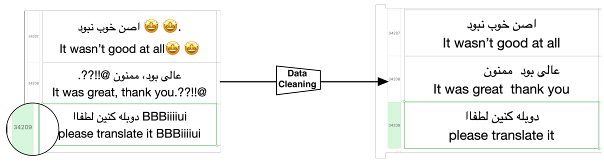

Figure 2: This figure presents the final results of data cleaning

resource languages tested on various language pairs (Bornea et al., 2021) introduces machine translation and adversarial training to enhance multilingual QA systems, considerably improving cross-lingual performance over zero-shot baselines by aligning language-specific embeddings.

The work (Shorten et al., 2021) surveys various text augmentation techniques, highlighting their impact on model generalization and performance in NLP tasks, particularly for limited labeled data, and emphasizes the need for task-specific strategies to maximize augmentation's potential. The article "BnPC: A Gold Standard Corpus for Paraphrase Detection in Bangla, and its Evaluation" (Sen, 2023) introduces BnPC, a benchmark Bangla corpus for paraphrase detection, showing its effectiveness in improving detection accuracy and advancing Bangla NLP research.

3 Senti-Persian Dataset

Creating a corpus involves several key steps: gathering, cleaning, annotating, and analyzing data, each influencing the others (McEnery and Brookes, 2022), (Ste, 2016). For example, analysis can reveal issues with annotations or sampling, leading to improvements and additional data collection. These steps are often recursive, as adjustments to annotations and dataset selection may be needed even after model training. Figure 1 provides an overview of the process we followed for Senti-Persian.

3.1 Data Collection

Senti-Persian corpora are built by sampling and filtering based on specific criteria using keywords and metadata to track sentiment. Among many choices, we collected data considering factors like time, location, and user demographics who posted or commented on movies (Moreno-Ortiz and García-Gámez, 2023) (Hu, 2016). Furthermore, our text selection approaches relied on movie genre, subjectivity, and popularity (Rheindorf, 2019) (Nandwani and Verma, 2021). Finally, the text selection process was constrained using Persian linguistic features, such as positive/negative words, intensifiers, negations, sentiment-laden adjectives, and emojis.

3.2 Text Cleaning

Unlike the Latin alphabet, the Persian alphabet does not have uppercase or lowercase letters, and the text is written from right to left. Furthermore, punctuation in Persian is limited, and many users need clarification on their proper use in text. Therefore, the first step in preprocessing is the removal of punctuation, as it often doesn't carry essential semantic information. The second step involves eliminating numbers, which may not add meaning to the sentiment depending on the context. In the third step, emojis that don't necessarily contribute to the core meaning of the content are removed. The fourth step includes the omission of extra spaces between words or sentences. Finally, as the data is sourced from web pages, we also observe HTML tags that are removed. Exceptionally, in this case, stop words are not removed as every word plays a pivotal role in preserving the original meaning of the contents (Lee et al., 2021) (Aut, 2022). Figure 2 presents the details.

3.3 Preprocessing Text Data

Both automatic and manual preprocessing are performed. During the manual phase, 'typos' are eliminated. To discover the appropriate form of a word, we used the Persian Accessible Dictionary Database (PD). Input texts containing a word not appearing in PD were considered typos. The corrected word was substituted for the typo in PD. For example, in the text تضویر بذ, the bolded letters indicate typo errors that must be corrected. By replacing the particles, it became تصویر بد. Preprocessing also includes null value imputation and removing unwanted data.

Algorithm 1: Majority Voting & Final Labeling

```
Begin
Corpus ← Collection of crawled and cleaned texts
Defined_labels ← [-1,0,1]
Final_Matrix()
For text in Corpus:
    tmpLabel = Select From Defined_labels
    Final_Matrix.append(text, tmpLabel)
End
```

3.4 Annotation Process

Labels for the entire corpus were manually assigned based on a majority vote. This involved defining an annotation scheme, markers, and granularity. In Opinion Mining (OM) and SA, labeling is challenging due to the need for a standard model.

Ten annotators categorized The collected data into Positive, Negative, and Neutral. Categorical and dimensional methods helped define emotions by grading polarity (positive/negative/neutral) and arousal. (active/passive). Algorithm 1 outlines the labeling process.

3.4.1 Guidelines and Process of Marking

This phase involved ten annotators, project managers, and expert reviewers. Annotators labeled sentiment polarities (positive, neutral, or negative) for predefined aspects of each sentence, following the methodology of (Chakravarthi et al., 2020). Native Persian annotators received training to ensure consistency.The annotation process had three rounds:

Data was split among five teams for independent annotation. Results were divided into Sub-Agree (consistent labels) and Sub-Disagree (disagreements). Sub-Agree data was reviewed, while Sub-Disagree cases were re-evaluated by the project manager. Complex cases were handed to expert evaluators for final decisions.

3.4.2 Annotation Validation

We recruited Persian university students as volunteers to handle the tagging process. They reviewed labels using Google Forms on their computers. Information about their gender, educational background, and schooling medium was collected for diversity. Reviewers were warned about potential hostile language in the comments and instructed

Figure 3: Google form for data annotations by volunteers.

to remain unbiased. Each Google Form contained 100 comments (10 per page). Annotators had to confirm their understanding of the scheme before proceeding. Figure 3 shows a portion of the Google Form.

3.4.3 Analysis and Exploitation

OM and SA-labeled datasets are crucial for training and testing ML tools for emotion classification, where data quality and quantity considerably impact results. Quality control techniques help detect errors, and comparing automated and human classification improves reliability.

Reusable, portable datasets are essential for emotion-oriented systems, and defining annotation standards is critical in OM and SA. The manual annotations were analyzed to understand Senti-Persian labeling distribution, highlighting polarity and emotional expressions. The chart in Figure 5 shows a sample distribution of movie reviews.

3.5 Balancing Techniques

A significant way to improve Deep Learning(DL) models is by behaving with categorical imbalanced datasets. Unbalanced collections can be handled in a variety of ways; there are two popular ways: "oversampling" and "undersampling" (Chawla, 2009) (He and Garcia, 2009). We observed in our previous paper that under-sampling yields better performance for all DL methods we

Comments	
Positive	11,583
Negative	7,578
Neutral	3,420
Total comments	**22,581**

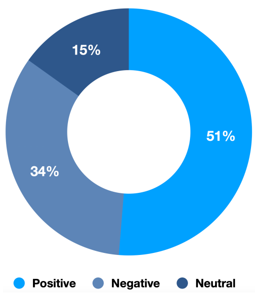

Figure 4: Comments Distribution **before** Augmentation

Comments	
Positive	22,581
Negative	22,581
Neutral	22,581
Total comments	**67,743**

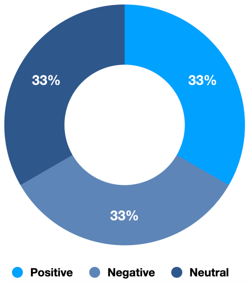

Figure 5: Comments Distribution **after** Augmentation

used (Mohammadi and Tavakoli, 2020).

3.6 Data Augmentation

Generative models enhance NLP quality, especially for low-resource languages (Chen et al., 2024). An essential contribution of this paper is the implementation of a GAN-based text generator for augmenting datasets, which will be detailed in the next section.

4 Methodology

This study collected limited movie reviews with positive, negative, and neutral sentiments. Each sentence consists of 'n' tokens. HAZM[2] Library was used to tag parts of speech (POS) in the corpus, and the chart in Figure 6 shows the frequency distribution of various POS, like verbs, adj, and nouns.

In Persian text augmentation, random masking for insertion, swapping, or synonym generation presents different linguistic challenges. We can augment most POSs, except verbs, which risk altering the sentence sentiment, a linguistic issue. For example, in فيلم بدي ني, which means "not a bad movie," if we change the verb position, the sentiment of the original sentence may change. for instance it may become فيلم بدي ه that means "it's a bad movie". Thus, in this study, tokens fall into two categories:

- Tokens that can change during the augmentation process, such as nouns, adjectives, and adverbs.
- Tokens that cannot change, primarily verbs.

Therefore, the applicability of the augmentation method on the samples depends on the specific characteristics, such as the use of subject, object, or modifiers in the text and their relative positions.

[2] https://github.com/roshan-research/hazm

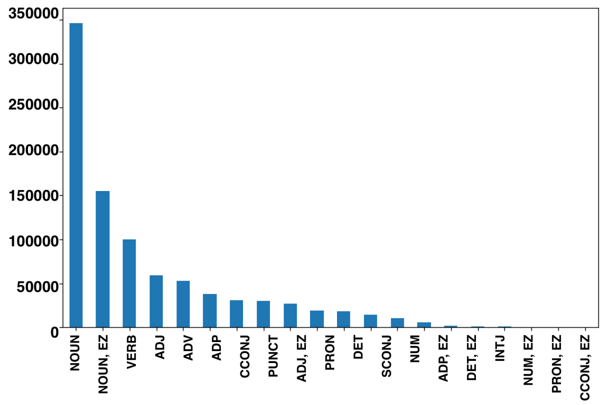

Figure 6: distribution of various parts of speech in the whole population

These tokens are masked for generating diverse but contextually similar samples. On the other hand, the method avoids masking tokens in the verb position.

4.1 GAN

GAN, commonly used in computer vision, also plays a key role in NLP (Goodfellow et al., 2014) (Chollet, 2017). In this study, GAN-based models generate new sentences by paraphrasing limited data. GAN has two components: a generator (based on ParsBERT) and a discriminator (Goodfellow et al., 2014). The generator produces new phrases, and the discriminator classifies them as fake or real (Farahani et al., 2021).

The Transformers pipeline simplifies this process through APIs for text augmentation. Initially, Random Replacement yielded the best results. For example, in the sentence یک جوري بود این قسمت اصلا به دلم ننشست، خوشم نیوم مسخره بود, the word قسمت (meaning "part") is rearranged using BERT (Devlin et al., 2018) to یک جوري بود این بخش اصلا به دلم ننشست، خوشم نیوم مسخره بود, maintaining the same meaning but with different words. The process is shown in Figure **??** and 8.

4.1.1 Generator

This paper implements a technique using transformers and the "fill-mask" pipeline to augment sentences through random insertion, synonym insertion, and random swapping. In this approach, sentences are generated by randomly masking the *Nth* token of a source sentence. For example, in این فيلم عالي بود ("It was a great movie"), each token can be masked and replaced using the unmasker pipeline. However, masking verbs may change the sentiment, so careful selection of masked tokens is needed. Nouns and pronouns are more suitable for masking to preserve sentiment. A list of sentences with varying masked positions is created, and the discriminator evaluates each one. Algorithm 2 outlines this process.

4.1.2 Discriminator

The discriminator model classifies the output from the generator as either DIFFERENT or SIMILAR. It evaluates whether the generated sentences, modified through insertion, swapping, etc., retain the semantically similar context of the source sample. A SIMILAR label means the sentiment is preserved, while DIFFERENT indicates a deviation from the source meaning. Algorithm 3 outlines this classification process.

In BERT, the CLS token is a unique token added at the start of a sentence to capture its overall meaning. The CLS embedding represents the entire sentence and is helpful for sentence-level tasks. The similarity between two CLS embeddings, typically calculated with cosine similarity, indicates how much the augmented text resembles the source. Cosine similarity ranges from -1 (opposed) to 1 (identical) (Choi et al.). Therefore, using the measures of TP, FP, TN, and FN, we compute the performance of Algorithm 3 compared to the ground truth of human annotation. According to the Figure 7, the cosine similarity of 0.8 results in the best discriminator performance.

5 Experiments and Results

5.1 Experimental Setup

We use 80% of the data for training and equally divide the rest for evaluation and testing. We preprocess the data by removing punctuation, emojis, duplicates, and html tags and transferring digits from English to Farsi. As simple baselines, we compare our results against a majority and random baseline. Our performance metrics include accuracy, precision, recall, and the F1 score. We use thundersvm for SVM; ThunderSVM exploits GPUs and multi-core CPUs to achieve high efficiency. For the pre-trained language models, we fine-tune ($\lambda = 2 \times 10^{-5}$, batch size 32) the models for 3 epochs with early stopping.

Algorithm 2: Generator

Begin
Dataframe ← Reads data from a CSV file
Do POS tagging and filter the verbs
Unmasker ← creates a fill-mask pipeline using the ParsBERT model
Inserts the '[MASK]' token at the randomly chosen index
Uses the unmasker pipeline to predict the most likely completion for the masked token.
Evaluate the generated sentences
End

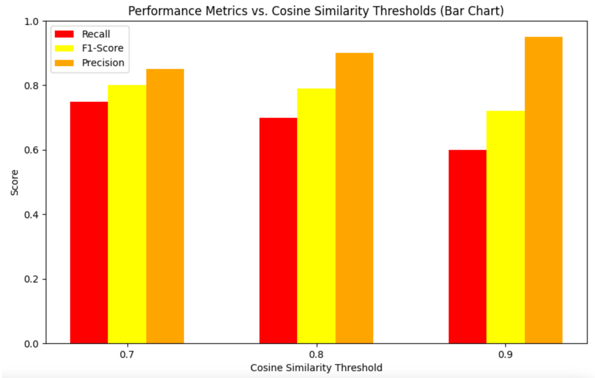

Figure 7: Performance Metrics comparison, to find the best threshold.

Algorithm 3: Discriminator

Begin
Sentence1 ← CLS embedding of source sentence before augmentation
Sentence2 ← CLS embedding of augmented sentence
Score ← cosine similarity between Sentence1 and Sentence2
If Score > 0.8:
    return "DIFFERENT"
else:
    return "SIMILAR"
End

5.2 Results & Analysis

In Tables 1 and 2, we present the performance of different models on the augmented and non-augmented datasets. By comparing the F1 scores of the two tables, we observe that all models show higher accuracy with augmented data than non-augmented data. On our dataset, the best-performing model is found to be WASSBERT (Mohammadi and Tavakoli, 2020), which was pre-trained on the highest volume of Farsi data.

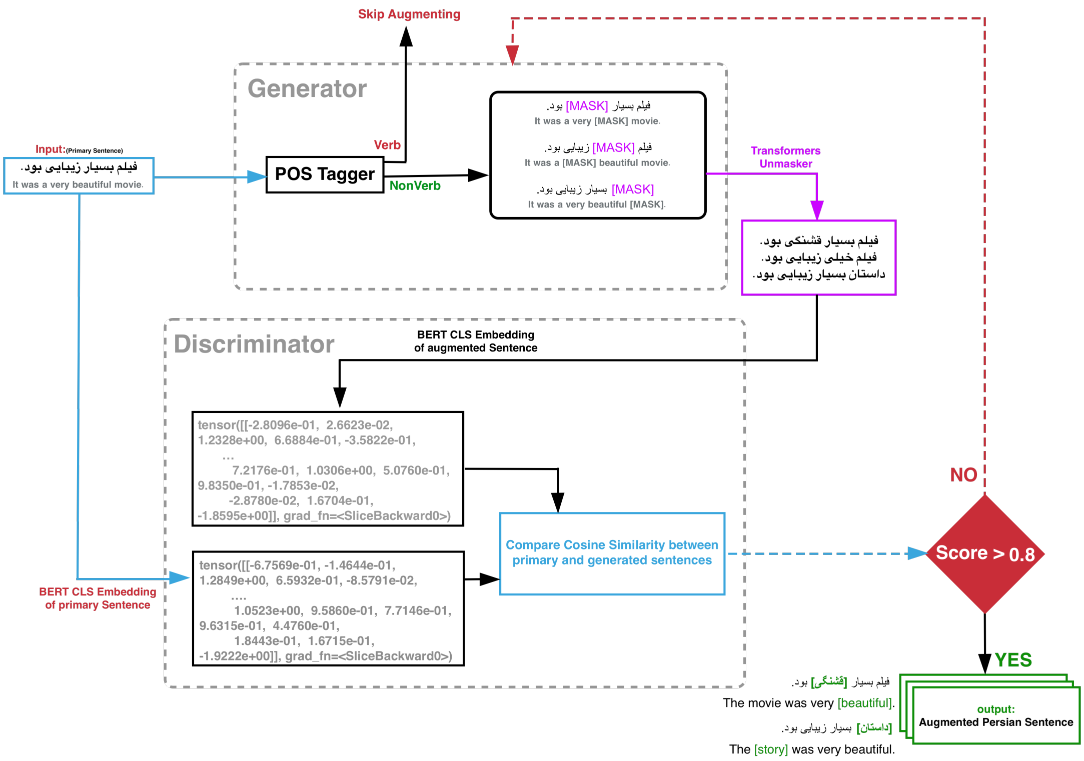

Figure 8: The GANs based model in detail

Model	Augmented Data			
	Accuracy	Precision	Recall	F1 Score
CNN	83.38%	83%	80%	81%
SVM	76%	80%	75.5%	75.5%
LSTM	72%	72%	72%	72%
CNN+LSTM	81%	81%	81%	81%
Bi-LSTM	87.07%	82%	85%	82%
Stacked Bi-LSTM	42.08%	42%	42%	42%
mBERT	90%	93.4%	90%	91%
XLM-RoBERTa	91%	90.01%	90%	90%
WassBERT	96%	95%	95%	95%

Table 1: Performance of different language models for the SA on the human-annotated movie reviews.

Model	Non-Augmented Data			
	Accuracy	Precision	Recall	F1 Score
CNN	77.33%	77%	70%	71%
SVM	70%	71.5%	70%	70%
LSTM	72%	72%	72%	72%
CNN+LSTM	81%	81%	81%	81%
Bi-LSTM	80%	79%	79%	75%
Stacked Bi-LSTM	38%	40%	37.5%	36%
mBERT	82%	84%	81%	82%
XLM-RoBERTa	83%	80%	81.3%	80%
WassBERT	90%	89%	89%	89%

Table 2: Presenting the improvement in the different language models after using augmented dataset.

6 Discussion

6.1 Diversity and Balance of Senti-Persian

We ensured diversity and balance in the Senti-Persian dataset by collecting data from various sources (social media, movie reviews), including formal, informal, and regional dialects (e.g., Shirazi, Isfahani). Gender, age considerations, and quality control were applied. After manual annotation, each sentiment category (positive, negative, neutral) was input into a GAN-based model to generate additional sentences. The synthetic data was manually reviewed for linguistic accuracy and sentiment relevance, resulting in a final corpus of 67,743 balanced comments.

6.2 Application on Other Arabic Languages

Our approach can be adapted for Arabic-script languages like Dari, Pashto, Urdu, Uyghur, Sindhi, Arabic, and Kurdish (Sorani), which share right-to-left writing, similar scripts, and word order but have unique features. Challenges include orthographic issues, vowel ambiguity, dialects, data imbalance, and complex morphology. Translating the primary dataset and applying GAN-based techniques can address these challenges and generate synthetic data.

6.3 Limitations

Persian has several linguistic characteristics that can influence the augmentation process we followed in this work. Following are a few aspects of Persian that may require specific adaptations:

1. Free word order: Changing word order for emphasis doesn't affect sentence sentiment, so models don't need to accurately prioritize capturing word arrangement or dependencies.

2. Morphology: Persian's inflectional nature, using prefixes and suffixes, doesn't affect sentence sentiment but poses challenges for tokenization. For example, کتاب (book) becomes کتاب خانه (library). The Hazm tokenizer handles these complexities accurately.

3. Postpositions and Case Marking: Persian uses postpositions (e.g., "in," "on" after nouns) instead of prepositions, affecting syntax but not sentiment.

4. Clitics and Compounds: Persian uses clitics and compound words, complicating tokenization. The Hazm tokenizer, designed for Persians, handles this effectively. For example, the word, دانش - "knowledge" and گاه - "place" or "house" together دانشگاه Translation: "University."

5. Lack of Capitalization: Persian lacks capitalization, impacting Named Entity Recognition (NER) models but not SA.

7 Conclusion and Future Works

This study presents a collection of 22,581 human-annotated data samples, which is later augmented using GANs, making it a total of 67,743 movie reviews annotated for SA. Our augmentation process resulted in achieving 96% accuracy, producing a boost of 7.6% in accuracy over the previous results. In the future, we aim to propose an approach that combines Reinforcement Learning (RL) with GANs to enhance the generation of long, coherent, and contextually appropriate text. We envision that the hybrid strategy would be able to refine GAN training mechanisms, improving the generated text's realism and linguistic quality. By combining the generative capabilities of GANs with the goal-oriented optimization of RL, we anticipate significant advancements in NLP, pushing the boundaries of current AI-driven text generation technologies.

References

2012. *Number of Internet Users by Language*. Archived from the original on 26 April 2012. Retrieved 10 May 2020.

2016. Steps for Creating a Specialized Corpus and Developing an Annotated Frequency-Based Vocabulary List. *TESL Canada Journal/Revue TESL du Canada*, 34(11):87–105.

2021. *Usage statistics of content languages for websites*. Archived from the original on 12 November 2021. Retrieved 12 November 2021.

2022. Automated rule-based data cleaning using nlp. In *2022 32nd Conference of Open Innovations Association (FRUCT)*, volume 32, pages 162–168.

2023. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. Presented in [Conference/Journal Name].

A. Balahur and M. Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60.

Surbhi Bhatia, Manisha Sharma, and Komal Kumar Bhatia. 2018. *Sentiment Analysis and Mining of Opinions*, volume 30.

M. Bornea, L. Pan, S. Rosenthal, R. Florian, and A. Sil. 2021. Multilingual transfer learning for qa using translation as data augmentation. In *AAAI Conference on Artificial Intelligence*. IBM Research AI, Thomas J. Watson Research Center, Yorktown Heights, NY.

B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae. 2020. Corpora for sentiment analysis of dravidian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 1–9. European Language Resources Association (ELRA).

N.V. Chawla. 2009. *Data Mining for Imbalanced Datasets: An Overview*. Springer, Boston, MA.

Y. Chen, Z. Yan, and Y. Zhu. 2024. A unified framework for generative data augmentation: A comprehensive survey. *arXiv preprint arXiv:2310.00277*.

H. Choi, J. Kim, S. Joe, and Y. Gwon. Evaluation of bert and albert sentence embedding performance on downstream nlp tasks. Unpublished manuscript.

F. Chollet. 2017. *Deep Learning with Python*, first edition. Manning Publications.

K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, A. Hussain, and Z. Ling. 2021. A hybrid persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks. *Neurocomputing*, 445:241–252.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2021. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:2109.00523*.

M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53:3831–3847.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial networks. Manuscript submitted for publication.

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. Semi-supervised and unsupervised sense annotation via translations. Alberta Machine Intelligence Institute, Department of Computing Science, University of Alberta, Edmonton, Canada.

H. He and E. A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Wan-Hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using bavarian as a case study. Preprint accepted at SIGUL 2024. Information Science, University of Regensburg, Germany.

K. Hu. 2016. *Compilation of Corpora for Translation Studies*. New Frontiers in Translation Studies. Springer, Berlin, Heidelberg.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. CIS, LMU Munich, Germany; Munich Center for Machine Learning (MCML), Germany; Instituto Superior Técnico (Lisbon ELLIS Unit); Instituto de Telecomunicações; Unbabel; Sorbonne Université, CNRS, ISIR, France.

A. Karimi, L. Rossi, and A. Prati. 2021. Aeda: An easier data augmentation technique for text classification. *arXiv preprint arXiv:2108.13230v1 [cs.CL]*.

Daniel Khashabi, Arman Cohan, Shima Shakeri, Payam Hosseini, Pouya Pezeshkpour, Mahsa Alikhani, Mohammad Aminnaseri, Mohammad Bitaab, Fatemeh Brahman, Sahand Ghazarian, Mohammad Gheini, Amir Kabiri, Ramin Karimi Mahabadi, Omid Memarrast, Alireza Mosallanezhad, Ehsan Noury, Shima Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Elham Sadeqi Azer, Nafise Sadat Safi Samghabadi, Mohsen Shafaei, Sina Sheybani, Asieh Tazarv, and Yadollah Yaghoobzadeh. 2021. Parsinlu: A suite of language understanding challenges for persian. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3538–3555.

G. Y. Lee, L. Alzamil, B. Doskenov, and A. Termehchy. 2021. A survey on data cleaning methods for improved machine learning model performance. Submitted on 15 Sep 2021.

Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*. Springer, Boston, MA.

T. McEnery and G. Brookes. 2022. *Building a written corpus: what are the basics?*, 2nd edition, page 13. EBook ISBN: 9780367076399.

C. Mi, L. Xie, and Y. Zhang. 2022. Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. *Neural Networks*, 148:194–205.

M. Mohammadi and S. Tavakoli. 2020. Wassbert: High-performance bert-based persian sentiment analyzer and comparison to other state-of-the-art approaches. *Journal Name*, 12:209–220.

A. Moreno-Ortiz and M. García-Gámez. 2023. Strategies for the analysis of large social media corpora: Sampling and keyword extraction methods. *Corpus Pragmatics*, 7:241–265.

P. Nandwani and R. Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11:81.

J. PourMostafa, R. Sharami, P. Abbasi Sarabestani, and S. A. Mirroshandel. 2020. Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus. *arXiv preprint arXiv:2004.05328v1 [cs.CL]*.

M. Rheindorf. 2019. *Working with Corpora Small and Large: Qualitative and Quantitative Methods*. Postdisciplinary Studies in Discourse. Palgrave Macmillan, Cham.

M. Salimi Sartakhti, R. Etezadi, and M. Shamsfard. 2022. Improving persian relation extraction models by data augmentation. *arXiv preprint arXiv:2203.15323v1 [cs.CL]*.

C. Shorten, T. M. Khoshgoftaar, and B. Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data*, 8(101):209–220.

Sumanth Tatineni. 2020. Deep learning for natural language processing in low-resource languages. *International Journal of Advanced Research in Engineering & Technology*, 11(5):1301–1311.

Patrick van Kessel, Skye Toor, and Aaron Smith. 2019. Popular youtube channels produced a vast amount of content, much of it in languages other than english. Pew Research Center. Retrieved 2 May 2022.

A. Williams, N. Nangia, and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2022. A survey on multilingual large language models: Corpora, alignment, and bias. (No.2022JJ006).

Q. Zhao. 2022. Review of natural language processing for corpus linguistics. *Corpus Pragmatics*, 6:311–314.