Using Language Models for assessment of users' satisfaction with their partner in Persian

Zahra Habibzadeh
School of Electrical
Computer Engineering,
College of Engineering,
University of Tehran
z.habibzadeh213@ut.ac.ir

Masoud Asadpour
School of Electrical
Computer Engineering,
College of Engineering,
University of Tehran
asadpour@ut.ac.ir

Abstract

Sentiment analysis, the process of gauging user attitudes and emotions through their textual data, including social media posts and other forms of communication, is a valuable tool for informed decision-making. In other words, by determining whether a statement conveys positivity, negativity, or neutrality, sentiment analysis offers insights into public sentiment regarding a product, individual, event, or other significant topics. This research focuses on the effectiveness of sentiment analysis techniques, using Machine Learning (ML) and Natural Language Processing (NLP) especially pre-trained language models for Persian, in assessing users' satisfaction with their partner, using data collected from X (formerly Twitter). Our motivation stems from traditional in-person surveys, which periodically analyze societal challenges in Iran. The limitations of these surveys led us to explore Artificial Intelligence (AI) as an alternative solution for addressing contemporary social issues. We collected Persian tweets and utilized data annotation techniques to label them according to our research question, forming the dataset. Our goal also was to provide a benchmark of Persian tweets on this specific topic. To evaluate our dataset, we employed several classification methods, including classical ML models, Deep Neural Networks, and pre-trained language models for Persian. Following a comprehensive evaluation, our results show that BERTweet-FA (one of the pre-trained language models for Persian) emerged as the best performer among the classifiers for assessing users' satisfaction. This point indicates the ability of language models to understand conversational Persian text and perform sentiment analysis, even in a low-resource language like Persian.

1 Introduction

Assessing people's sentiments, culture, social values, and attitudes is paramount in gaining insights into a society's collective mindset and functioning. Human societies are made up of individuals who interact and shape their environment based on shared beliefs and behaviors. Therefore, we can comprehend society's challenges, strengths, and weaknesses by investigating these factors.

Since 2000, traditional assessments have been conducted three times in Iran in provincial surveys under the supervision of experts in various fields. One of the critical topics in these surveys is community members' satisfaction in measuring their social and cultural status. The main challenges in this field can be identified by evaluating and understanding people's satisfaction with various factors in this extensive research and surveys. However, this traditional approach to data collection, such as questionnaires, was last conducted in 2015 in Iran, where families were interviewed and completed self-report lists to assess their satisfaction. Nevertheless, this method has shortcomings in conducting detailed investigations and analyzing surveys responses. One issue is the difficulty attracting many participants due to privacy concerns and disagreement with the research. Additionally, social satisfaction is a variable characteristic that changes over time, requiring regular surveys every few years to keep up with changing trends. Conducting surveys every few years is costly for the country regarding finances and human resources, and the reporting process is time-consuming, taking several weeks or even months to complete. Furthermore, there is a risk of human error in the reports, leading to lower accuracy and higher costs.

At the same time, the popularity of social networks has grown with the advancement of Internet technology and the widespread use of smartphones. They have become a crucial part of our lives, enabling people to communicate and access information. These platforms also offer a new way of sharing information, exchanging knowledge, and connecting people globally. Social networks serve

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 78–88
January 20, 2025. ©2025 Association for Computational Linguistics

more than just as a tool for users to document their lives and connect with others; they also provide an avenue for expressing personal thoughts and maintaining relationships. X (formerly Twitter) is a popular social networking platform with real-time and interactive features. Users can express their emotions through texts, emojis, photos, and videos, making it a suitable and essential platform to share happiness and sadness. Furthermore, tweets contain short emotional information that holds significant value in shaping public opinion and driving social impact. This feature reflects users' interests and preferences and can significantly influence the spread of online public opinion. Therefore, in this research, we decided to analyze social network data to address a crucial societal issue. Specifically, we aim to examine the satisfaction levels of individuals within their relationships. With the increasing divorce rates in our society, it is crucial to understand the factors that contribute to satisfaction and dissatisfaction in relationships.

Hence, we propose a methodology that utilizes Artificial Intelligence (AI) and Machine Learning (ML) techniques, especially Language Models (LMs), to minimize the challenges and costs associated with traditional surveys. By analyzing the tweets that users post on social media platforms, we can gather data on a large scale without requiring human resources. This will allow us to design an efficient model, saving time and resources while providing valuable insights into our society's social challenges. However, we acknowledge that this study is based on data collected from the Persian-speaking community on X, which may not fully represent the wider population. Therefore, this research serves as a preliminary case study, highlighting the potential of social network data to address societal issues, while also acknowledging the limitations of its specific user base.

Overall, in this paper, we make the following contributions. (1) We provided a new labeled dataset and a benchmark that explores user satisfaction with their partner, specifically targeting Persian tweets, to evaluate the performance of different classification models and LMs; (2) We designed a new framework to analyze social questions based on social networks using AI that can reduce the drawbacks of traditional surveys; (3) Following the results of our classification models, we testified the power of the transformer-based model for the Persian language in investigating social problems.

2 Related Work

Over the past decade, sentiment analysis has emerged as one of the main areas of research in both Data Mining and Natural Language Processing (NLP). Researchers have provided this approach in different applications like analyzing movie reviews (Ouyang et al., 2015), identifying hateful content on social media (Pitsilis et al., 2018), analyzing mobile reviews in Persian (Saraee and Bagheri, 2013), opinion analysis (Alimardani and Aghaie, 2015) and opinion mining (Alikarami et al., 2023). The mentioned projects are part of the intensive literature in this research area.

Furthermore, sentiment analysis is a valuable tool for examining and analyzing user characteristics on social networks. In (Quercia et al., 2011), researchers conducted a comprehensive analysis of the relationship between users' personalities, including popular users and influencers, using X data. They developed a model to estimate users' personalities based on follower data and used ML algorithms such as Support Vector Machine (SVM) (Stitson et al., 1996; Tuba and Stanimirovic, 2017) for prediction. Their research revealed that emotional stability and extroversion are common traits among all users, while popular users tend to be more imaginative, and influential users are typically more organized. These findings provide valuable insights that were previously difficult to quantify. By predicting user personalities from public data, we can gain important information for various applications. Bai et al. (2014) proposed a social satisfaction prediction model based on research in the field. They used APIs to collect micro-blogging data from social networks and conducted surveys to obtain user satisfaction scores. Their results showed that regional social satisfaction is linked to local economic indicators. This suggests that the prediction model can accurately identify social satisfaction through social media data. Also in (Liao et al., 2017), a novel technique for measuring user satisfaction with a product was introduced using Deep Neural Networks (DNNs) like Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; Bottou et al., 1994) instead of traditional ML algorithms. The CNN network achieved a higher accuracy rate than classical algorithms such as SVM. After noticing the good performance of neural networks, researchers applied different deep learning techniques and architectures such as word embeddings and Long short-term memory

(LSTM) (Hochreiter, 1997), even to develop sentiment analysis systems with higher performance (Ouyang et al., 2015; Pitsilis et al., 2018; Zhao et al., 2017; Hassan and Mahmood, 2017).

Previous research has primarily focused on individual-level opinions, such as those related to films and products. However, there remains a significant gap in the analysis of collective sentiments and opinions on pressing social issues. Recognizing this gap and building upon prior research, we propose models to predict users' satisfaction with their partner based on Persian tweets. By analyzing social media content related to users' satisfaction with their partner, we hope to better understand the real-life problems families face in our society. Furthermore, we believe that these models can serve as a valuable foundation for addressing other cultural and social issues in the future.

3 Method

In this section, we will discuss in detail the approach proposed for assessing satisfaction with their partner, and we will provide more information about our benchmark dataset. Our proposed framework is shown in Figure 1, and we will provide additional details in the following sections.

3.1 Data Collection

We collected data from the X based on specific keywords. In other words, to ensure relevance to our study on life partner satisfaction, we selected Persian keywords such as "wife", "my wife", "betrayal", "divorce", and others that are commonly associated with this topic. These keywords were chosen to capture many user sentiments concerning life partner relationships. From April 2021 to April 2022, we extracted 179,891 Persian tweets that matched our selected keywords. After initial data retrieval, we retained only the text column, discarding other irrelevant metadata such as user information and timestamps. This approach enabled us to focus on textual content relevant to our analysis of life partner satisfaction.

3.2 Dataset

3.2.1 Primary Preprocessing

We began by including general words (e.g., men, women) as criteria to align the dataset more closely with the research topic. Subsequently, we conducted further data filtering using more specific words (e.g., wife, marriage, relationship), collecting 16,499 tweets directly from our collection. The tweets were then subjected to primary preprocessing. During this preprocessing stage, we implemented several crucial steps to prepare the dataset for analysis (e.g., removing URLs, email addresses, Unicode characters, weird patterns, retweets, hashtags, usernames, links, and duplicate tweets). However, we kept punctuation marks and emojis that convey emotional expression, which is essential for accurate tagging. Upon completing this preprocessing stage, our dataset comprised 13,239 tweets, all set for the subsequent labeling stage.

3.2.2 Data Annotation

We developed a comprehensive guideline (Appendix A) based on extensive research within the field. This guideline was shared with our team of annotators to guide the data annotation process.

We introduced new columns for the data annotation process. In the following sections, we provide details on these additional columns and their role in our data annotation methodology.

- Relevance Label: In the Persian language, many common words can change the meaning of a tweet depending on the sentence's context, making relevance detection crucial. This label is used to determine whether the tweet is related to our research topic.

- InRelationship Label: This label indicates the user's relationship status, which can be Unknown, Single, or Married.

- General Comment Label: Some tweets may address the topic in general rather than based on personal experience. This label determines if the tweet is related to the research topic and whether it publicly expresses satisfaction or dissatisfaction towards a life partner. If the tweet discusses the topic generally, it is classified as Positive, Negative, or Neutral based on the emotional tone conveyed.

- Specific Comment Label: Finally, it is checked if the tweet pertains to our topic and whether it refers to the user's partner or not. In case it does refer to the user's partner, we analyze the emotional tone of the tweet and assign one of three labels - Positive, Negative, or Neutral - based on the sentiment.

During the data annotation process, annotators performed the process twice to minimize the error

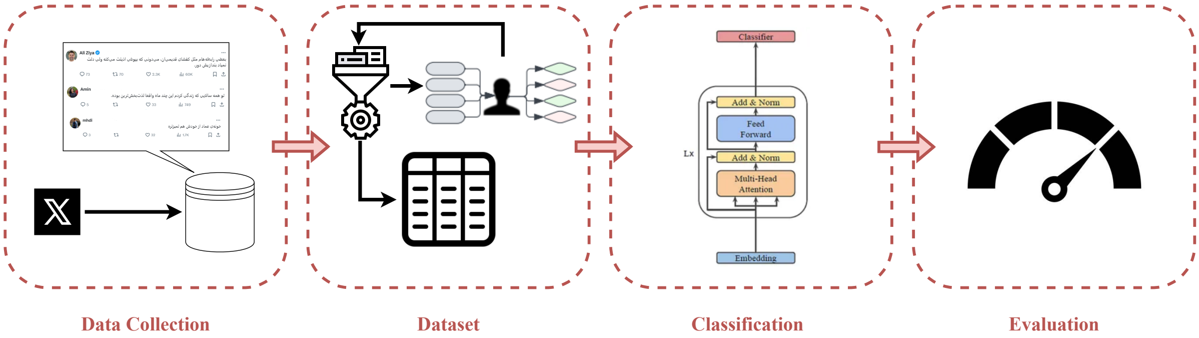

Figure 1: This figure shows our proposed framework. First, we collected data from X. After a primary preprocessing stage, we labeled the data using human annotators. Following a second round of preprocessing, we created our final dataset. Finally, we trained classifiers and evaluated their performance.

rate, ensuring the highest possible accuracy in the dataset through this validation step.

3.2.3 Secondary Preprocessing

Once the labeling stage is completed, the dataset undergoes another round of preprocessing to prepare it for classification. This involves removing punctuation marks, emoticons, Persian and English numerals, English words, and stopwords to optimize it for ML algorithms. However, for deep learning models, stopwords, emoticons, and punctuation marks are retained as they may contain valuable information for the models. The final dataset used for this research comprises 13,239 tweets, each labeled appropriately.[1]

3.3 Classification

In the context of this research, the dataset, along with its collection and data annotation process, was developed as a novel and unique contribution, so no established benchmark model existed for direct performance comparison. To assess the dataset, we initially employed a diverse set of ML models, such as K-Nearest Neighbors (KNN) (Shah et al., 2020), Random Forest (RF) (Pranckevičius and Marcinkevičius, 2017), Naive Bayes (NB) (Kim et al., 2006), SVM, and Logistic Regression (LR) (Peng et al., 2002), alongside CNNs and Bidirectional Long Short-Term Memory (BiLSTM) (Graves and Schmidhuber, 2005). We also utilized LMs such as the pre-trained Persian Bidirectional Encoder Representations from Transformers (ParsBERT) (Farahani et al., 2021) and BERTweet-FA (Malekzadeh, 2020), fine-tuning them as necessary for our specific task. Finally, we implemented a hybrid model that leveraged ParsBERT's embeddings in combination with DNNs.

In the following sections, we first explain the word embedding techniques we used and then comprehensively explain each model used in our research.

3.3.1 Word Embeddings

Before classification, textual data must be converted into numerical vectors to be processed by the classifier. To enable KNN, RF, NB, SVM, and LR algorithms to model the texts effectively, the TF-IDF (Term Frequency-Inverse Document Frequency) (Ramos et al., 2003) method was employed. We used the Hazm library (optimopium et al., 2023) to tokenize each tweet, enhancing the interpretability of numerical vectors for DNNs. This process generates a vector containing the indexed words for each tweet. Subsequently, we randomly selected the embedding matrix before feeding the data into the network's embedding layer. This step ensures that the input becomes more understandable for the intended network. Furthermore, we used pre-trained embeddings from two Persian language models: the ParsBERT and BERTweet-FA, to investigate the impact of pre-trained embeddings on our models' performance.

3.3.2 Machine Learning Models

After completing the previous steps and preparing the dataset for classifier training, we used LR, KNN, NB, RF, and SVM classification algorithms in the first step of classification to determine the best algorithm. We employed GridSearchCV and k-fold cross-validation (with $k = 5$) to optimize the hyperparameters for these models. The numerical vectorization of texts and labels was performed with all parameter combinations. Each classifier

[1] The dataset and codes are available at this link: https://github.com/zaha2020/UserSatisfactionSentiment

was then trained and tested on the dataset. See Appendix B for the hyperparameters of each model.

3.3.3 Deep Neural Networks

In this study, DNNs were implemented using the PyTorch framework (Paszke et al., 2019). The first implemented model was CNNs, which utilized a three-layered convolutional structure to extract local features. The network consisted of 36 filters with sizes of 3, 5, and 7. Furthermore, a max-pooling layer was incorporated to reduce dimensionality, followed by a fully connected layer to facilitate classification tasks. To prevent overfitting during model training, a dropout rate of 0.1 was applied within the network structure. The optimization process employed the Adam optimizer with a learning rate of 0.001, and the CrossEntropy error function was used as the loss function.

The other implemented model was a BiLSTM network. This model employed a bidirectional recurrent layer with 10 hidden units to learn dependencies between input units and retain word-level features. To prevent overfitting, a dropout rate of 0.5 was applied within the network structure. Similar to the CNN model, the optimization process used the Adam optimizer with a learning rate of 0.001, and the CrossEntropy loss function served as the objective function for training the BiLSTM model.

3.3.4 ParsBERT and BERTweet-FA Models

We also employed the ParsBERT model, a monolingual language model built upon Google's Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) architecture. In 2020, this model was pre-trained on a vast corpus of Persian text, containing diverse writing styles and a wide range of subjects, including scientific literature, novels, and news articles.

To evaluate the performance of ParsBERT models, we tested three distinct models:

- ParsBERT_I: In the first case, the ParsBERT-trained model without freezing the network parameters was used.

- ParsBERT_II: In the second case, all layers up to the 11th layer were frozen, while the last layer remained unfrozen.

- ParsBERT_III: All layers of the model in the third and final case to prevent updates were frozen.

BERTweet-FA is another transformer-based model trained on a dataset of 20,665,964 Persian tweets. Notably, this model was trained for only one epoch and included 322,906 training steps. Despite its relatively short training duration, the model reveals the ability to understand the meaning of a substantial portion of conversational sentences in the Persian language. It is essential to emphasize that the model's architecture closely follows the original BERT framework.

It is important to note that all models based on pre-trained language models in this study were trained across three to five epochs using the Adam optimizer with a learning rate of 0.00002.

3.3.5 ParsBERT with Deep Neural Networks

In our latest models, we have enhanced the input layer by replacing random word embeddings with pre-trained ParsBERT embeddings. This integration of pre-trained language models allows our CNN and BiLSTM layers to benefit from rich semantic and syntactic information.

Category	Train	Test
Relevant	5554	1227
Irrelevant	5302	1156

Table 1: Training and test data distribution for each Relevance label category.

Classifier	Accuracy	F1-score
RF	51.50	33.40
KNN	64.96	64.74
ParsBERT_III	67.10	66.89
ParsBERT_II	68.02	67.69
NB	69.66	68.81
BiLSTM	71.13	71.12
SVM	72.28	72.22
LR	72.41	72.36
ParsBERT-BiLSTM	73.61	73.42
CNN	73.48	73.47
ParsBERT-CNN	75.12	75.11
ParsBERT_I	78.10	78.09
BERTweet-FA	**80.53**	**80.51**

Table 2: Classifier performance on Relevance label with Accuracy and F1-score (%).

4 Results

To compare classifiers effectively, it is important to maintain a consistent dataset. To achieve this, we randomly selected 82% of the dataset as the training set, while the remaining 18% was assigned to the test set. These two datasets were then saved as separate CSV (comma-separated values) files, ensuring that a fixed dataset is used for all classifications. Different metrics were used to evaluate the proposed approaches. For the classification evaluation, we utilized accuracy and the macro F1 score (F1-score).

Category	**Train**	**Test**
Married	2324	524
Unknown	2953	650
Single	263	51

Table 3: Training and test data distribution for each InRelationship label category.

Classifier	**Accuracy**	**F1-score**
KNN	41.90	35.27
RF	58.44	42.96
ParsBERT_II	59.59	49.45
ParsBERT_III	62.45	51.49
SVM	74.57	52.72
NB	67.16	53.32
BiLSTM	73.31	53.58
ParsBERT-BiLSTM	73.55	54.71
LR	74.82	57.47
CNN	77.31	59.84
ParsBERT_I	77.55	60.93
ParsBERT-CNN	78.12	61.87
BERTweet-FA	**79.59**	**68.02**

Table 4: Classifier performance on InRelationship label with Accuracy and F1-score (%).

4.1 Relevance Label

Table 1 shows the train and test data for classification in each Relevance label category. Table 2 shows the performance results of the classifiers on the Relevance label. Based on the experimental results, the BERTweet-FA model achieved better performance compared to other models with an accuracy of 80.53% and F1-score of 80.51%. This indicates that the model can effectively recognize whether the new tweet is related to users' satisfaction topic with their partner or not.

4.2 InRelationship Label

Table 3 shows the train and test data for classification in each InRelationship label category. According to Table 4, the BERTweet-FA model has the best performance in detecting the users' relationship status in new tweets regarding satisfaction with a life partner, with an accuracy of 79.59% and an F1-score of 68.02%. We emphasize this by analyzing the obligation level of individuals whose tweets were identified as relevant during data annotation. As a result, the dataset used in this stage became more specific and reduced for classifiers.

4.3 General Comment Label

Table 5 shows the train and test data for each classification of the category of general comments labels. Also, table 6 shows the performance results of the classifiers on the General Comment label. In this label, we analyze the sentiment of tweets related to the research topic, categorizing them into three groups: Positive, Negative, and Neutral. We aim to test the accuracy of our classifiers in correctly categorizing tweets in the test dataset. As shown in Table 6, the BERTweet-FA model has achieved better performance, with an accuracy of 61.88% and an F1-score of 58.50%. In other words, this result indicates that the BERTweet-FA is effective in analyzing sentiment in Persian tweets.

Category	**Train**	**Test**
Positive	1244	269
Negative	2614	575
Neutral	1682	381

Table 5: Training and test data distribution for each General Comment label category.

Classifier	**Accuracy**	**F1-score**
NB	48.72	27.13
CNN	45.14	38.42
KNN	46.67	44.63
ParsBERT_II	54.51	45.02
LR	54.07	46.35
RF	54.40	46.72
SVM	53.25	47.60
BiLSTM	52.74	47.61
ParsBERT_III	57.23	50.49
ParsBERT-BiLSTM	53.80	51.59
ParsBERT-CNN	55.76	53.76
ParsBERT_I	60.65	55.97
BERTweet-FA	**61.88**	**58.50**

Table 6: Classifier Results on General Comment label with Accuracy and F1-score (%).

4.4 Specific Comment Label

Table 7 shows the train and test data for classification in each Specific Comment label category. In the final stage of our analysis, we evaluated the models' ability to predict the emotional tone of tweets related to users' relationships. Many users on X share personal experiences about their partners. By analyzing the emotional load of these tweets, categorized as positive, negative, or neutral, we can gain insights into users' satisfaction with their partners. According to the evaluation of the performance of the models in Table 8, the BERTweet-FA model, with an accuracy of 57.22% and an F1-score of 56.02%, has performed better than other models.

Category	Train	Test
Positive	797	195
Negative	799	181
Neutral	857	177

Table 7: Training and test data distribution for each Specific Comment label category.

Classifier	Accuracy	F1-score
ParsBERT-CNN	42.23	33.36
CNN	42.96	33.78
KNN	41.78	40.72
BiLSTM	42.96	42.69
ParsBERT_II	48.08	48.14
NB	46.95	46.51
SVM	46.95	46.67
LR	48.24	47.81
PBERT-BiLSTM	47.90	47.85
RF	49.91	49.58
ParsBERT_III	48.81	48.81
ParsBERT_I	56.31	55.16
BERTweet-FA	**57.22**	**56.02**

Table 8: Classifier Results on Specific Comment label with Accuracy and F1-score (%).

5 Conclusion

The motivation for this study stems from the drawbacks of the traditional surveys conducted in Iran every few years. Our research aims to enhance traditional survey methods by introducing a new approach to analyzing complex social issues by applying text classification methods and testing the performance of pre-trained language models for Persian. In particular, we leveraged ML and NLP techniques to classify the sentiment of tweets from X users regarding their satisfaction with their partner. Our data collection took place on the X social network, primarily in Persian, given its popularity among Persian-speaking individuals. Following data preprocessing, we employed human taggers to annotate the tweets according to our research question, forming a labeled dataset as a challenging benchmark for classification models and pre-trained LMs for Persian. As there was no existing foundational model for the subject under investigation, our research explored various classification algorithms, including SVM, KNN, NB, RF, LR, BiLSTM, CNN, ParsBERT, ParsBERT-BiLSTM, ParsBERT-CNN, and BERTweet-FA. Our comprehensive evaluation shows that BERTweet-FA, a pre-trained language model for Persian, outperformed the other classifiers in accurately classifying sentiment in Persian tweets. This result highlights the effectiveness of LMs in understanding conversational Persian text for sentiment analysis and challenging social problems.

In future research, we aim to explore semi-supervised learning techniques for data annotation and employ multilingual and large Language Models (LLMs) to enhance the dataset and classification models further, respectively. We also plan to investigate data augmentation methods to address the issue of data scarcity and improve the robustness of our models. Additionally, we will explore deeper linguistic insights, such as analyzing sentiment-bearing idioms and slang unique to Persian, to enhance the interpretability and performance of our models in Persian NLP.

6 Limitations

One main limitation of this study was the lack of data in Persian, as Persian remains a low-resource language in NLP research (Magueresse et al., 2020). This challenge was compounded by the specific social focus of our research topic, which further limited the availability of relevant data.

Furthermore, annotating tweets presented a significant bottleneck in establishing a benchmark for this study. In addition, a significant limitation of this study is the lack of specific user properties, such as age. Incorporating this information into future studies could provide more informative insights into the results.

References

Hossein Alikarami, Amir Massoud Bidgoli, and Hamid Haj Seyed Javadi. 2023. Belief mining in persian texts based on deep learning and users' opinions (revised december 2022). *IEEE Transactions on Affective Computing*.

Saeedeh Alimardani and Abdollah Aghaie. 2015. Opinion mining in persian language using supervised algorithms.

Shuotian Bai, Rui Gao, Bibo Hao, Sha Yuan, and Tingshao Zhu. 2014. Identifying social satisfaction from social media. *arXiv preprint arXiv:1407.3552*.

Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Larry D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. 1994. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, pages 77–82. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Abdalraouf Hassan and Ausif Mahmood. 2017. Deep learning approach for sentiment analysis of short texts. In *2017 3rd international conference on control, automation and robotics (ICCAR)*, pages 705–710. IEEE.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng. 2006. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, 18(11):1457–1466.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng. 2017. Cnn for situations understanding based on sentiment analysis of twitter data. *Procedia computer science*, 111:376–381.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Malekzadeh. 2020. Bertweet-fa: A pre-trained language model for persian (a.k.a farsi) tweets. *https://github.com/arm-on/BERTweet-FA*.

optimopium, sir kokabi, and mhdi707. 2023. Hazm 0.9.4: A persian text processing toolkit. In *GitHub repository*. Roshan Research.

Xi Ouyang, Pan Zhou, Cheng Hua Li, and Lijun Liu. 2015. Sentiment analysis using convolutional neural network. In *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing*, pages 2359–2364. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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

Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Tomas Pranckevičius and Virginijus Marcinkevičius. 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Mohamad Saraee and Ayoub Bagheri. 2013. Feature selection methods in persian sentiment analysis. In *Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings 18*, pages 303–308. Springer.

Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. 2020. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5:1–16.

MO Stitson, JAE Weston, A Gammerman, V Vovk, and V Vapnik. 1996. Theory of support vector machines. *University of London*, 117(827):188–191.

Eva Tuba and Zorica Stanimirovic. 2017. Elephant herding optimization algorithm for support vector machine parameters tuning. In *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE.

Wei Zhao, Ziyu Guan, Long Chen, Xiaofei He, Deng Cai, Beidou Wang, and Quan Wang. 2017. Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 30(1):185–197.

A Annotators Guidelines for Sentiment Classification

The team of annotators consisted of four graduate students (two male and two female) at the University of Tehran. We decided on the final label of each data point based on a majority vote of the annotators.

To ensure accurate and unbiased annotations, we provided our team with detailed guidelines. The annotators were instructed to label the tweets in a CSV file, strictly following these guidelines and setting aside any personal beliefs or biases. Additionally, we asked our annotators to write a brief comment about each tweet, explaining the reasons for their labels. This process helps reduce errors and biases in the dataset.

In this section, we will provide instructions and examples of our guidelines. It is notable to mention that translating these sentences from Persian to English may add ambiguity due to the linguistic properties of the Persian language. Feel free to contact authors if you want to get the original guidelines in Persian.

Relevance Label

This label is used to determine whether the tweet is related to our research topic. If a person does not have a partner, wants a partner, etc., they are not suitable for our problem and label all these tweets as irrelevant.

Below are examples of each Relevance label category.

- **Relevant**: *"I suggested to my husband we go to his mom's for a kebab, but he just laughed and called me a foodie."*
- **Irrelevant**: *"I want a husband now."*

InRelationship Label

The InRelationship label indicates the user's relationship status, which can be Unknown, Single or Married. The labels are assigned based as follows:

- **Single:**: The user is single.
- **Married**: The user has a life partner.
- **Unknown**: The status is unclear.

Below are representative examples of each InRelationship label category.

- **Single**: *"What more could I ask from life? A fat bank account, a partner like Kristen Stewart, and a family like Queen Elizabeth's."*
- **Married**: *"The beauty of my spouse is amazing. [Heart emoji]"*
- **Unknown**: *"Oh, they got married. Some people have all the luck with such good spouses."*

General Comment Label

This label determines if the tweet is related to the research topic and whether it publicly expresses satisfaction or dissatisfaction towards a life partner. Some tweets may address the topic in general rather than based on personal experience. If the tweet discusses the topic generally, it is classified as Positive, Negative, or Neutral based on the emotional tone conveyed.

- **Positive**: Tweets conveying happiness or satisfaction.
- **Negative**: Tweets expressing anger, dissatisfaction, or dislike.
- **Neutral**: Tweets with no emotional tone.

Below are examples of each General Comment label category.

- **Positive**: *"It was Eid al-Fitr that I received my wife as a gift from God."*
- **Negative**: *"Marriage is awful. You even have to visit your spouse's relatives."*
- **Neutral**: *"Did you give Eid gifts to your spouse or boyfriends yet?"*

Model	Parameters
KNN	n_neighbors=9
Random Forest	bootstrap=True, max_depth=80, max_features=2, min_samples_leaf=3, min_samples_split=8, n_estimators=100
Naive Bayes	Default parameters for MultinomialNB
SVM	decision_function_shape='ovo',degree=1, kernel='linear', C=1, gamma=1
Logistic Regression	max_iter=5000, multi_class='multinomial', penalty='l2', solver='newton-cg'

Table 9: Hyperparameters for Machine Learning Models related to the Relevance Label.

Model	Parameters
KNN	n_neighbors=1
Random Forest	bootstrap=True, max_depth=80, max_features=3, min_samples_leaf=3, min_samples_split=8, n_estimators=1000
Naive Bayes	Default parameters for MultinomialNB
SVM	decision_function_shape='ovo', degree=2, kernel='poly', C=5, gamma=1
Logistic Regression	max_iter=5000, multi_class='multinomial', penalty='l2', solver='saga'

Table 10: Hyperparameters for Machine Learning Models related to the InRelationship Label.

Model	Parameters
KNN	n_neighbors=9
Random Forest	Default parameters for RandomForestClassifier
Naive Bayes	Default parameters for MultinomialNB
SVM	kernel='linear', C=1, gamma=1
Logistic Regression	max_iter=5000, multi_class='multinomial'

Table 11: Hyperparameters for Machine Learning Models related to the General Comment Label.

Model	Parameters
KNN	n_neighbors=10
Random Forest	Default parameters for RandomForestClassifier
Naive Bayes	Default parameters for MultinomialNB
SVM	decision_function_shape='ovo', degree=1, kernel='linear', C=1, gamma=1
Logistic Regression	max_iter=5000, multi_class='multinomial', penalty='l2', solver='saga'

Table 12: Hyperparameters for Machine Learning Models related to the Specific Comment Label.

Specific Comment Label

In the Specific Comment Label, we want to determine whether the tweet is relevant to our topic and if it mentions the user's partner. If the tweet does reference the user's partner, we will analyze its emotional tone and assign one of three labels: Positive, Negative, or Neutral, based on the sentiment expressed. For this section, we will consider four labels:

- **Positive**: Expressing happiness or satisfaction from their life partner.
- **Negative**: Expressing dissatisfaction or anger

from their life partner.

- **Neutral**: Statements without emotional tone.

The following are examples of each Specific Comment label category.

- **Positive:** *"My husband bought our favorite pizza for dinner. Such a thoughtful gesture."*
- **Negative:** *"Marriage is awful. Visiting in-laws is such a chore."*
- **Neutral:** *"Should we visit my in-laws or stay with my family for the holidays?"*

B Configuration of Machine Learning Models

Tables 9, 10, 11 and 12 provide a detailed overview of the hyperparameters utilized for the ML models implemented in Python with the Scikit-learn library (Pedregosa et al., 2011). For more details about the implementation, please refer to the code.