# CKingCoder at SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis

Harish Kumar B, Naveen D, Prem B and Aarthi S

(harishkumarb0711@gmail.com, dnaveen0502@gmail.com prembala1710@gmail.com, aarthigopinath.msec@gmail.com) Computer Science and Engineering, Meenakshi Sundararajan Engineering College

### Abstract

The SemEval 2023 Task 9 Multilingual Tweet Intimacy Analysis, (Pei et al., 2023) is a shared task for analysing the intimacy in the tweets posted on Twitter. The dataset was provided by Pei and Jurgens (2020), who are part of the task organizers, for this task consists of tweets in various languages, such as Chinese, English, French, Italian, Portuguese, and Spanish. The testing dataset also had unseen languages such as Hindi, Arabic, Dutch and Korean. The tweets may or may not be related to intimacy. The task provided was to score the intimacy in tweets and place it in the range of 0-5based on the level of intimacy in the tweet using the dataset provided which consisted of tweets along with its scores. The intimacy score is used to indicate whether a tweet is intimate or not. Our team participated in the task and proposed the ROBERTa model, Liu et al. (2019) to analyse the intimacy of the tweets.

# 1 Introduction

Intimacy is the feeling that makes people feel close to each other. Intimacy is formed as a result of sharing knowledge and experience with each other. Intimacy is developed through a process of four phases, as per the Levinger and Snoek Interdependence Model. The Four Phases of Intimacy are the No Contact Phase, the Awareness Phase, the Surface Contact Phase, and the Coexistence Phase. The no-contact phase is when the people involved don't know each other. The awareness phase is when the people involved just know about each other but don't have any superficial contact. The surface contact phase is when the people involved know each other and have superficial contact with each other. The coexistence phase is where the people involved have deep links and are mutually dependent on each other. Social media is now on an enormous upward curve. Intimacy is commonly referred to as physical intimacy. It has four major forms, among which one is physical intimacy. The

other forms of intimacy are emotional intimacy, where intimacy forms personal bonds such as love or marriage; spiritual intimacy, where intimacy is formed due to the sharing of spiritual ideas or opinions; and intellectual intimacy, where intimacy is formed by shared common opinions between induividuals.

In modern times, there has been a notable shift in the way individuals express their emotions in real-time. With the proliferation of social media, communication has largely moved online, with individuals utilizing platforms such as Twitter, chat applications, and virtual meetings. Twitter, in particular, has experienced tremendous growth, with over 500 million tweets being sent per day as of March 1, 2023. Users are able to share a wide range of emotions on this platform, from news of personal victories to concerns and anxieties. This shift towards virtual communication has led to a society that is increasingly introverted and reliant on digital forms of interaction, with physical communication becoming increasingly rare.

Social media platforms are used for both professional and personal communication, which including intimate tweets. Through the process of Multilingual Tweet Intimacy Analysis, The task at hand involves identifying tweets that are classified as intimate and those that are not. The difficulty here is not only the anonymity of the tweeter but also the limited information provided in the tweet. Since the tweet string length is limited to 280 characters, you should extract the score of intimacy from the provided tweet of 280 characters maximum alone. It takes its complications to yet another level, and last but not least, it is a multilingual task where you may not be provided with data in a particular language alone. Few Transformer Models were used for the task such as XLM-R (Conneau et al., 2020), XLM-T (Barbieri et al., 2022), DistilBERT (Sanh et al., 2020) and miniLM (Wang et al., 2020) and tried to identify which provided

the highest accuracy for the task. XLM-R achieved a slightly higher accuracy for the Task.

# 2 Background

Starting with the task, The task is to predict the intimacy score of the tweets in the dataset. The training dataset consists of 9491 tweets. The languages that are provided in the training dataset are considered seen languages. This dataset was used to train the model. This indicates that there are fewer samples to predict highly intimate content. The tweets hide external data, such as The user names that have been tagged in the tweet have been replaced with "@user, and the hyperlinks in the tweets have been replaced with 'http://' by the organiser itself. To avoid models using clickbait spoiler strategies to find out if the tweet is intimate or not since the task is about the analysis of intimacy in multilingual tweets alone and no external data is allowed to be used apart from text. Dataset was used in each of the models such as XLM-R (Conneau et al., 2020), XLM-T (Barbieri et al., 2022), DistilBERT (Sanh et al., 2020), miniLM (Wang et al., 2020) and the best of them i.e. RoBERTa was used.

The Transformer model is a deep learning architecture for natural language processing that was introduced by Vaswani et al. (2017). The model consists of an encoder and a decoder, each with multiple layers of self-attention and feed-forward neural networks. The self-attention mechanism allows the model to focus on different parts of the input sequence when generating its output, while the feed-forward layers provide nonlinear transformations that enable the model to capture complex patterns in the data. The Transformer has been widely used for a variety of natural language processing tasks, including language translation, text generation, and sentiment analysis, and has achieved state-of-the-art performance in many of these tasks. Its success has led to the development of several variants and extensions, such as BERT (Devlin et al., 2019), GPT (Radford et al., 2018), and RoBERTa (Liu et al., 2019), which have further improved the performance of the model on various language tasks.

# 3 System Overview

This work describes the Multilingual Tweet Intimacy Analysis approach taken by CKingCoder, which is based on the RoBERTa (Robustly Optimized BERT) approach (Liu et al., 2019), in the

Language	Number of Tweets
Portuguese	1596
Chinese	1596
Spanish	1592
French	1588
English	1587
Italian	1532

Table 1: Language Distribution in Tweets of TrainingData.

Score Range	No. of Tweets
1 - 2	4630 tweets
2 - 3	2970 tweets
3 - 4	2475 tweets
4 - 5	416 tweets
Total Tweets	9491 tweets

Table 2: Intimacy Score Distribution in Tweets of Training Data.

SemEval Task 9 conducted in 2023. RoBERTa is an adaptation of the Bidirectional Encoder Representations from Transformers (BERT) model developed by Jacob Devlin and his colleagues in 2019 (Devlin et al., 2019). Since its introduction, BERT has become the most popular transformer model in natural language processing, with over 150 publications utilizing the model. RoBERTa, along with other transformers such as DistilBERT (Sanh et al., 2020) and XLM-R (Conneau et al., 2020), have been developed to improve upon BERT's performance. When BERT (Devlin et al., 2019) was initially developed, it was implemented in two model sizes. One of BERT-BASE with 12 encoders and 12 bidirectional self-attention heads adding up to 110 million parameters, and another of BERT-LARGE with 24 encoders and 16 bidirectional self-attention heads adding up to 340 million parameters. These high-performance models were pre-trained with Wikipedia English and the Toronto Books Corpus, which were the largest available encyclopedias at the time, which made it the largest pretrained language model up to date. After a stage, it became as popular as Google's implementation of BERT in English language search queries, and by October 2020, all the English and English based queries were processed by a BERT model.

BERT is pre-trained simultaneously for two tasks: one is language modelling and the other is next sentence prediction. Since it is already pre-trained with larger datasets, it can be fine-tuned with less computational resources on smaller data based on the requirements of the Natural Language Processing Task. It was considered that BERT is superior to Deep Learning Models because Deep Learning Models try to learn information from the text using all possible methods of neural networks, whereas BERT tries to learn the information from the words surrounding the text and concluding with the exact required information for the text.

As time progressed, many modified versions of BERT were released that were suitable for certain circumstances, such as miniLM (Wang et al., 2020) and DistilBERT (Sanh et al., 2020), which were simpler implementations of BERT that could produce a similar outcome with less training time, and similarly, XLM-RoBERTa (Conneau et al., 2020) was a transformer based on BERT that was pretrained on two terabytes of data that was obtained through the web crawlers, which have been collecting the data for over 10 to 12 years and are available at the Common Crawl Corpus. Although the XLM-T Transformer (Barbieri et al., 2022) was tested, it was found to be pre-trained with only up to 30 languages, while the XLM-R (Conneau et al., 2020) was pre-trained with approximately 100 languages. Consequently, to maintain the project's global inclusivity, it was deemed necessary to employ the XLM-R Transformer instead. It is good to note that during its release, the BERT was considered state-of-the-art due to its exceptional performance in specific Natural Language Processing Tasks such as GLUE (General Language Understanding Evaluation) by Wang et al. (2019), SQuAD (Stanford Question Answering Dataset) by Rajpurkar et al. (2016) and SWAG (Situations With Adversarial Generations) by Zellers et al. (2018).

The BERT Model helps create the required model for the Multilingual Tweet Intimacy Analysis since it is not a use-case that can take a long time to process as done with Deep Learning Models and in the meantime, it should not compromise the accuracy of the intimacy prediction since higher false positives or true negatives could affect the prediction results.

The RoBERTa, i.e., Robustly optimised BERT approach, (Liu et al., 2019) is trained with dynamic masking, full sentences without NSP loss, large mini-baches, and a larger byte-level BPE. Additionally, it has been investigated with two other important factors that have been emphasised in previous models: the data used for pretraining and the

number of training passes through the data. It uses an additional dataset, CC-NEWS, along with the existing datasets of BERT, i.e., Toronto BookCorpus, and Wikipedia English. Instead of pretraining the BERT model for 1 million steps, RoBERTa pretrains it just one hundred thousand times, additionally pretraining with additional data and pretraining the data for a little longer in series. It was found that RoBERTa achieved state-of-the-art excellent results, outperforming all existing BERT models.

### 4 Experimental Setup

Since no other data is available, the training data was split up into two sets, one for training and the other for testing, in the ratio of 70:30. As the data needed to be preprocessed, unwanted elements such as the username handles and urls were removed. Different BERT models are trained with the Split Training Data and are experimented with the Split Testing Data among which the RoBERTa model was exceptionally better than other models. Hence, the RoBERTa model was finalised and used for testing the final test data provided by the organiser.

The test data had additional languages such as Hindi, Arabic, Dutch, and Korean for which training data were not provided. So it was a separate challenge to overcome. But since the RoBERTa is already pre-trained in hundreds of languages, it was able to overcome the shortcomings a bit.

Analysing the Tweets of Individual Languages, more than 90% of the tweets have less than 20 tokens. Chinese and Korean languages are relatively shorter than other languages because the languages use ideographs instead of alphabets like other languages. Tweets in Hindi have approximately 15 tokens.

The metrics that were considered for the evaluation were Pearson's correlation coefficient, which is often referred to as Pearson's R. The Pearson's R or correlation coefficient, is defined as the ratio of the covariance of the two variables in question in the numerical dataset normalised to the square roots of their variances. Mathematically, it is the division of the covariance of the two variables by the product of their standard deviations. The Pearson correlation ranges between +1 and -1.

A Pearson Correlation Value of +1 indicates that it is perfectly positive correlated, i.e., Person A's age and Person B's age exactly increase over the same duration.

Language of Tweets	Test data consist
Korean Tweets	1410
English Tweets	1396
Spanish Tweets	1396
Portuguese Tweets	1390
Dutch Tweets	1389
French Tweets	1382
Arabic Tweets	1368
Chinese Tweets	1354
Italian Tweets	1352
Hindi Tweets	1260

Table 3: Language Distribution of Test Data.

A Pearson Correlation value of -1 indicates that it is perfectly negatively correlated, i.e., like the Hour of the Day and the Remaining Hours of the Day, where if one increases, the other will decrease in the same quantity.

A positive Pearson The coefficient value indicates that the value is positively correlated but not perfectly, i.e., when the quantity of food eaten is increased and the feeling of having eaten well is increased, after a certain point the feeling disappears. A negative Pearson The correlation value indicates that the value is negatively correlated, like the speed of the vehicle and the time taken to reach the destination, but the values don't increase or decrease in exact quantity.

Lastly, if the Pearson value is 0, then the value is said to be independent, i.e., the quantity of food eaten and the score obtained in exams.

Therefore, the closer the value is to +1 or -1, the greater is their dependence.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=0}^{n} (y_i - \bar{y})^2}}$$

Formula of Pearson Correlation Coefficient.

# 5 Result

The results were initially tested with different metrics, such as Mean Absolute Error (MAE) which is the sum of all differences with average to the total number of values; Root Mean Squared Error (RMSE) which is the square root of the average squared difference between actual and predicted value; Mean Squared Error (MSE) which is the squared difference between actual and predicted

Language	Pearson Correlation Coefficient
Spanish	0.3845
Portuguese	0.299
Italian	0.298
English	0.2836
French	0.2667
Chinese	-0.1232
Dutch	0.2116
Korean	0.1445
Hindi	0.0478
Arabic	0.1

Table 4: Pearson Correlation Obtained

value. But later on, It was found that Pearson Coefficient Correlation has to be used. On obtaining test results, The model faced a major drawback with a low overall Pearson correlation coefficient of 0.1325. We were ranked 43rd on the leaderboard with an Pearson Coefficient of 0.224 among seen languages.

Since languages such as Hindi, Arabic, Korean, and Dutch were introduced in the test data, The model had a lower Pearson coefficient compared to other languages. This model is now able to score the intimacy score in the tweet.

### 6 Error Analysis

During the development of the Multilingual Tweet Intimacy Analysis model, external data was utilized for testing purposes. However, it would have been advantageous to incorporate a manually created bulk external dataset to ensure the model's robustness prior to publication. Unfortunately, due to a last-minute error, the submission was limited to a single submit, which constrained testing opportunities. To enhance the model's performance, black box testing could have been conducted to provide additional insights into the model's behavior. Furthermore, data augmentation techniques could have been explored as a means of improving the model's performance by generating additional training data. Implementing these measures could have potentially resulted in the Multilingual Tweet Intimacy Analysis model achieving higher extremes than its current iteration. It is notable that this model exhibits a lower Pearson correlation coefficient when tasked with languages that incorporate characters that are not analogous to those found in the English language. The intention is to enhance the language model to achieve the highest possible accuracy.

## 7 Conclusion

The XLM-T Transformer (Barbieri et al., 2022) was considered as a baseline model recommended in the dataset paper, and it could have improved the metrics better than the current model. It is a pre-trained model on millions of tweets specifically in over thirty languages and provides a Twitter Sentiment Analysis dataset in over eight languages in which it was trained, giving it a slight advantage over other models. Despite the short period, the accuracy was increased to the maximum possible level. The dataset will be improved further to maximize the Pearson coefficient to at least 0.90 and develop a web application to deploy the model to end-users, which could detect intimate tweets just by providing the tweet's URL and a single click. Additionally, the application could be extended to report a Twitter user when there are frequent tweets of highly intimate nature in real-time without any delays, making it easier for users to access and utilize the application.

The scope of the analyzer will be extended to audio files, enabling it to be implemented in podcasts and live broadcasts to censor or report the contents of intimate conversations. Furthermore, media files sent along with the tweet will also be captured to gain additional insights into whether the tweet is intimate or not. This will involve stepping into computer vision and enhancing the features of the application from that point on. Since Twitter doesn't currently open their API Access, The approach currently used is Web Scraping to get the tweet context from the Twitter using the snscrape, a Social Networking Service Scraper (JustAnotherArchivist, 2018) and analyse the tweet using this Model. The data collected is not used for development, training or commercial purposes and the data collected is neither saved nor displayed anywhere in the system. Once The Twitter API is open to public, It was be enhanced in future instead of the web scraping technique of snscrape.

### 8 References

#### References

- Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.
- JustAnotherArchivist. 2018. snscrape: A social networking service scraper.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.
- Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.
- Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. Semeval 2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep selfattention distillation for task-agnostic compression of pre-trained transformers.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference.