Clasificación de tuits humorísticos: un enfoque eficiente basado en aprendizaje automático y ensamble

Brian Miguel Escalona Maldonado

Laboratorio de Ingeniera de Software

Universidad Autónoma de la Ciudad de México (UACM)

brian.escalona@alumnos.uacm.edu.mx

Referencia de este artículo [Maldonado(2026)].

RESUMEN

Este trabajo presenta los resultados obtenidos en la competencia de clasificación de tuits de humor, donde se evaluaron diversos modelos de aprendizaje automático. El modelo **WW_GG_Robertuito_LL_V1** alcanzó el tercer lugar, con un desempeño muy cercano al de los dos primeros puestos, mostrando diferencias menores a un punto porcentual. Los resultados evidencian que arquitecturas optimizadas y técnicas de ensamble pueden competir de manera efectiva contra modelos de gran escala como *Gemini* y *DeepSeek*, manteniendo un equilibrio entre *Precisión*, *Recall* y *F1-score*. Asimismo, se confirma una mejora significativa respecto a los *baselines* tradicionales, lo que valida la pertinencia del enfoque propuesto y su potencial aplicación en el análisis automático de contenido humorístico en redes sociales.

Palabras clave: Clasificación de tuits humorísticos, Modelos de machine learning, Técnicas de ensamble, Optimización de arquitecturas, Gemini, DeepSeek, Precisión, Recall, F1-score, Baselines, Análisis de contenido en redes sociales.

ABSTRACT

This work presents the results obtained in the humor tweet classification competition, where various machine learning models were evaluated. **The model WW_GG_Robertuito_LL_V1** achieved third place, with performance very close to the top two positions, showing differences of less than one percentage point. The findings demonstrate that optimized architectures and ensemble techniques can effectively compete against large-scale models such as Gemini and DeepSeek, while maintaining a balanced performance across Precision, Recall, and F1-score. Furthermore, a significant improvement over traditional baselines is confirmed, validating the relevance of the proposed approach and its potential application in the automatic analysis of humorous content on social media.

Keywords: Humor tweet classification, Machine learning models, Ensemble techniques, Optimized architectures, Gemini, DeepSeek, Precision, Recall, F1-score, Baselines, Social media content analysis.

MSC 2020: 35R11

Introducción

En el marco de la competencia de clasificación de humor del curso de Computación evolutiva 2026-I, la presente investigación tiene como finalidad mostrar el trabajo realizado, con el que obtuvo al final de la competencia un **F1-Score** de $0{,}8923$ logrando así el tercer lugar dentro de la competencia. El objetivo del concurso era determinar si un tuit pretendía ser humoristico o no, el desempeño de la tarea se mide utilizando la puntuación Macro F1. La clasificación es binaria, por lo que se compone de dos clases:

- Clase 0 (No humorístico): El texto no contiene elementos humorísticos evidentes.
- Clase 1 (Humorístico): El texto contiene características que indican humor.

0.1. Conjunto de datos

El conjunto de datos proporcionado está basado en la competencia HaHa de Iberlef [Chiruzzo et al.(2020)Chiruzzo, Castro, and Rosá]. Este es un conjunto de datos con un tamaño de $16,000$ tuits anotados manualmente y separados en dos subconjuntos:

- Entrenamiento: $10,400$ tuits.
- Evaluación: $5,600$ tuits.

Cada linea del archivo contiene un json, que consta de los campos:

- ïd": Define el identificador del texto.
- "klass": Define la clase del texto: 1 para humor; 0 no contiene humor.
- "text": Contiene el texto con el tuit.

0.2. Humor

El humor es un concepto emocional complejo y ambiguo, propio del lenguaje natural [Li et al.(2022)Li, Liu, and Wang]. El lenguaje humorístico no puede existir de forma independiente, ya que adquiere significado solo cuando se acompaña de contexto, situación y trasfondo cultural El análisis del discurso permite interpretar el humor [Liang(2011)]. El reconocimiento del humor es un desafió en el procesamiento del lenguaje natural (NLP), esto debido a diferentes razones, en primer lugar el humor suele utilizar el lenguaje figurado, como la ironía y el sarcasmo. Además, el sentido del humor varia entre diferentes grupos sociales, culturales y geográficos. Las personas con diferentes conocimientos previos reaccionan de manera distinta ante el mismo chiste, esta variabilidad es lo que dificulta la detección de contenido humorístico [Wu et al.(2024)Wu, Huang, and Lau].

0.3. Large Language Models

Los Modelos de Lenguaje de Gran Escala (LLMs), tales como Gemma 3, Qwen 2.5 y Llama 3 [et al.(2025c)] [Team(2025)] [et al.(2024)], representan la vanguardia en el procesamiento del lenguaje natural y la inteligencia artificial generativa. Fundamentados primordialmente en la arquitectura de redes neuronales autorregresivas basadas en Transformers, estos modelos sustituyen los antiguos enfoques recurrentes mediante el uso intensivo de mecanismos de autoatención (**"self-attention"**). Esta característica les permite procesar secuencias de texto de manera no lineal y capturar dependencias contextuales de largo alcance de forma altamente eficiente en paralelo.

La robustez de los LLMs subyace en sus dimensiones vectoriales y en su masivo conteo de parámetros, los cuales son ajustados durante una etapa inicial de preentrenamiento autoregresivo utilizando corpus masivos de texto no estructurado. Computacionalmente, el proceso de inferencia de un LLM consiste en predecir la probabilidad de distribución del siguiente fragmento de texto o **"token"** (t_i) condicionado por la secuencia de **"tokens"** precedentes ($t_1, t_2, \ldots, t_{i-1}$), formulado matemáticamente como:

$$P(t_i \mid t_1, t_2, \ldots, t_{i-1}) \tag{1}$$

Esta enorme parametrización dota a las redes de capacidades emergentes, destacando el "aprendizaje en contexto" (**"In-Context Learning"**) [et al.(2020)]. Mediante este fenómeno, el modelo aprovecha el conocimiento latente codificado en sus pesos base para adaptarse de forma inmediata a nuevas tareas o dominios sin requerir modificaciones en sus gradientes, operando exclusivamente bajo las restricciones semánticas proporcionadas dentro del límite de su ventana de contexto activa.

0.4. Prompt Engineering

La ingeniería de prompts (**"Prompt Engineering"**) constituye una disciplina metodológica fundamental dentro de la interacción hombre-máquina, actuando como la interfaz abstracta de comunicación y control entre los LLMs y los sistemas de software periféricos. Lejos de ser una simple redacción de instrucciones, esta técnica representa la estructuración matemática y semántica del espacio de entrada del modelo, influyendo de manera directa en las matrices de atención y, en consecuencia, en la entropía y calidad de las distribuciones probabilísticas de salida.

El objetivo central de la ingeniería de prompts es mitigar la ambigüedad inherente al lenguaje natural, guiando los mecanismos de decodificación de la red hacia un espacio de soluciones óptimo para la ejecución de tareas especializadas. Metodologías avanzadas tales como el **"Few-Shot Prompting"** (provisión de ejemplos estructurados de entrada-salida) o el **"Chain-of-Thought"** (instar al modelo a desglosar una lógica secuencial de razonamiento antes de emitir un veredicto) optimizan el rendimiento del LLM. En escenarios hiperespecíficos como el análisis de moda, un prompt rigurosamente estructurado previene sesgos semánticos y reduce drásticamente las alucinaciones fenotípicas de la IA, asegurando que las respuestas se ciñan estrictamente a formatos legibles y estructurados por código, como esquemas JSON o clasificaciones taxonómicas discretas [White et al.(2023)White, Fu, Hays, Sandborn, Olea, Gilbert, Elnashar, Spencer-Smith, and Schmidt].

0.5. Fine-Tuning

El ajuste fino (**"Fine-Tuning"**) es un proceso metodológico de optimización supervisada enmarcado dentro del paradigma del Aprendizaje por Transferencia (**"Transfer Learning"**). A través de esta técnica, un modelo fundacional genérico previamente entrenado con vastos recursos de conocimiento general es sometido a un segundo ciclo de optimización utilizando un conjunto de datos acotado, de alta calidad y representativo de un dominio especializado [Parthasarathy et al.(2024)Parthasarathy, Zafar, Khan, and Shahid].

Desde una perspectiva matemática, mientras que el preentrenamiento busca generalizar el conocimiento minimizando la pérdida sobre un corpus web global, el fine-tuning altera la superficie de pérdida de la red para adecuar los pesos sinápticos (θ) a las sutilezas estadísticas de un espacio particular.

Dado el costo computacional prohibitivo de actualizar la totalidad de los miles de millones de parámetros de un LLM moderno, esta investigación se apoya en técnicas de Ajuste Fino Eficiente en Parámetros (PEFT), específicamente mediante la Adaptación de Bajo Rango (**LoRA, "Low-Rank Adaptation"**) [Hu et al.(2021)Hu, Shen, Wallis, Allen-Zhu, Li, Wang, Wang, and Chen]. Esta metodología inyecta matrices de descomposición de rangos menores en las capas de atención del Transformer, permitiendo actualizar únicamente una fracción minúscula de los pesos (menor al 1 %) mientras se congela el modelo base. Ello viabiliza el reentrenamiento local de modelos avanzados como Gemma o Qwen, dotándolos de un criterio estético profundo sin incurrir en la degradación catastrófica del conocimiento general previo.

Desarrollo

Durante el desarrollo de la competencia, se utilizaron diferentes enfoques para garantizar el desempeño de los modelos y aumentar la métrica de evaluación. Esta tarea fue complicada ya

que el conjunto de datos cuenta con un desbalance de clases bastante marcado (60|30) en la clase objetivo (clase 1: Humor), por lo que elevar la métrica fue un reto bastante complicado. En esta investigación, se destacan 4 modelos principales:

- **RoBERTuito**
- **Gemma 3:4B**
- **Qwen 2.5:3B**
- **LLama 3.2:3B**

0.6. Modelos de DeepLearning

0.6.1. RoBERTuito

RoBERTuito es un checkpoint de RoBERTa preentrenado específicamente para el español latinoamericano y el lenguaje utilizado en redes sociales (en especifico, la red social **X**) [Pérez et al.(2022)Pérez, Furman, Alemany, and Luque]. Este modelo fue el modelo ganador en la competencia celebrada en el curso de Redes Neuronales 2025-II con un **F1-Score** de $0{,}8744$ [Maldonado and Rodriguez(2025)]. Por lo qué fue el primer modelo con el que se empezó a experimentar.

Primeramente se realizo un cambio en la arquitectura para la optimización en el entrenamiento, esto con la finalidad de emplear el menor tiempo posible en el entrenamiento del modelo y aprovechar para realizar más experimentos. Una vez optimizado se realizo un cambio completo en los parámetros utilizados y se obtuvo un nuevo mejor resultado con un **F1-Score** de $0{,}8758$. La *Tabla:* 1 muestra el cambio de arquitectura utilizada.

Cuadro 1: Comparación de la arquitectura de los modelos RoBERTuito

	RB_MM_V3	Robertuito_V2
Neuronas	4096, 1024	764, 256
Batch size	16	16
Dropout	0.3, 0.2	0.25, 0.13
Max lenght	100	124
Epocas	4	15
F1	0.8744	0.8758

Con este cambio, el nuevo modelo **Robertuito_V2** se posicionaba por encima de su antecesor; la *Tabla:* 2 muestra la comparación detallada de las métricas evaluadas. Este modelo, con los pocos cambios y optimizaciones, logró obtener un F1 considerable, por lo que el principal objetivo de superar a la arquitectura del año pasado ya se había alcanzado.

Cuadro 2: Comparación de las métricas de los modelos RoBERTuito

Modelo	Presicion	Recall	F1	Acurracy
RB_MM_V3	0.8734	0.8755	0.8744	0.883
Robertuito_V2	0.8731	0.879	0.8758	0.8838

Sin embargo, pesé a los resultados positivos, este **F1** está cercano al límite del modelo, por lo que fue necesario probar nuevos modelos para incrementar aún más la métrica objetivo.

0.6.2. Gemma 3-4B

Gemma es una familia de modelos abiertos ligeros de Google; los modelos Gemma 3 son modelos multimodales con una ventana de contexto de $128K$ tokens y soporte para más de 140 idiomas [et al.(2025c)]. Para la competencia se utilizó el modelo de 4 billones de parámetros. Para poder realizar un proceso de fine-tuning y especializar al modelo en el contexto de nuestra

tarea, se utilizo el servicio de Google Colab para el entrenamiento del modelo. Se entrenaron cerca de $65k$ parámetros de los $4B$ disponibles.

Se modelo de manera inicial un prompt bastante sencillo con el que nos comunicaremos con el LLM para la inferencia.

> Eres un experto en lingüística computacional especializado en humor en español. Responde ÚNICAMENTE con SI o NO.

Con la inferencia se realizó una primera prueba en validación y se obtuvo un **F1** de $0{,}8459$ por lo que se buscó optimizar el umbral de decisión para obtener una mejor métrica a la hora de evaluar con el test, el *Código:* **??**. Con esta calibración se obtuvo un nuevo **F1** mejorado de $0{,}8732$ lo que supone un aumento de $0{,}0273$ puntos, lo cual es bastante bueno. Se probó el modelo con el conjunto de prueba y finalmente se obtuvo un **F1-score** de $0{,}8818$ una mejora significativa respecto al modelo **Robertuito_V2**. La *Figura:* 3 muestra la comparación de las métricas de los modelos probados.

Cuadro 3: Comparación de las métricas de los modelos

Modelo	Presicion	Recall	F1	Acurracy
RB_MM_V3	0.8734	0.8755	0.8744	0.8830
Robertuito_V2	0.8731	0.8790	0.8758	0.8838
GG_V3	0.8799	0.8840	0.8818	0.8896

0.6.3. Qwen 2.5-3B

Qwen 2.5 es una serie de modelos de lenguaje grande con una ventana de contexto de hasta $128K$ tokens, pudiendo generar hasta $8K$ tokens y con capacidades mejoradas en programación y matemáticas. Además, tiene soporte para más de 29 idiomas, incluyendo el español [Team(2025)]. Con este modelo, nuevamente se realizó un proceso de fine-tuning en el que se tomaron $29k$ parámetros de los $3B$ disponibles. Dado que se necesitaba una visión diferente para integrarse más adelante en un ensamble de modelos, Qwen se entreno con un enfoque especializado en la teoría del humor; para esto, se utilizo la teoría de Victor Raskin: **Humor theory: What is and what is not** [Raskin(2017)]. Con la teoría de Raskin, se obtuvieron algunas reglas que fueron incluidas en el prompt de entrenamiento.

> Eres un experto en lingüística computacional del humor especializado en la Teoría Semántica Ontológica del Humor (OSTH). Para clasificar un tweet como humorístico, busca: - Un cambio de guion semántico (script switch): el texto empieza en un contexto y termina en uno opuesto o inesperado. - Un punto de disparo (punch line): palabra o frase que activa el cambio. - Una oposición de tipo normal/anormal, real/irreal o posible/imposible. Si detectas estos mecanismos, el tweet contiene humor. Responde ÚNICAMENTE con la palabra 'Sí' (contiene humor) o 'No' (no contiene humor).

Después del entrenamiento se realizó nuevamente la validación y el ajuste de umbral, finalmente obteniendo un **F1** en validación de $0{,}8777$, realizando la inferencia en el conjunto de prueba el modelo **WW_V1** obtuvo un **F1-score** de $0{,}88$, este resultado es sumamente importante ya que solo es una baja porcentual de $0{,}0018$ respecto al modelo **GG_V3**. La *Tabla:* 4 muestra la comparación de los tres modelos probados.

Cuadro 4: Comparación de las métricas de los modelos

Modelo	Presicion	Recall	F1	Acurracy
RB_MM_V3	0.8734	0.8755	0.8744	0.8830
Robertuito_V2	0.8731	0.8790	0.8758	0.8838
GG_V3	0.8799	0.8840	0.8818	0.8896
WW_V1	0.8778	0.8824	0.88	0.8879

0.6.4. Llama3-3B

La colección Llama 3.2 de modelos de lenguaje grandes multilingües es un conjunto de modelos generativos preentrenados, estos modelos están optimizado para tareas de recuperación y resumen de agentes. Tiene soporte para 8 idiomas oficiales, incluyendo el español. Se utilizo el modelo de 3 billones de parámetros para el entrenamiento y fine-tuning, nuevamente se entrenaron $24K$ parámetros de los $3B$ disponibles. Nuevamente se utilizo la teoría del humor de Victor Raskin; sin embargo el enfoque de este modelo fue entrenarlo para reconocer que cosa **NO es humor** esto para brindarle una perspectiva distinta a un ensamble, por lo que el prompt de entrenamiento cambio un poco.

> Eres un experto en linguistica computacional del humor especializado en la Teoria Semantica Ontologica del Humor (OSTH). Tu tarea es identificar tweets que NO son humoristicos, es decir, texto serio, literal o neutro. Un tweet es SERIO (no humoristico) cuando: - No presenta cambio de guion semantico (script switch): el contexto es coherente de principio a fin. - No hay punto de disparo (punch line): no existe palabra o frase que active un giro inesperado. - No hay oposicion del tipo normal/anormal, real/irreal o posible/imposible. - El mensaje es directo, literal y su interpretacion principal no cambia al releerlo. Responde UNICAMENTE con la palabra 'Si' (el tweet es serio, no contiene humor) o 'No' (el tweet contiene humor).

Para la inferencia, se necesitaba hacer un ajuste, ya que nuestro modelo aprendió a predecir que cosa NO es humor. Entonces, para la clasificación, se realiza el ajuste: $score = 1 - p\_serio$. Con este ajuste se realizo la inferencia en validación más el ajuste y se obtuvo un **F1** de $0{,}8781$. Después de encontrar el umbral de decisión, se realizo la prueba en el conjunto de datos de la competencia y el modelo **LL_Anti_V1** obtuvo un **F1-score** de $0{,}8727$. Esto puede parecer un poco bajo comparado con los modelos anteriores, pero considerando que el modelo está diseñado para la tarea contraria, este resultado es bastante alto comparado con arquitecturas probadas anteriormente.

Cuadro 5: Comparación de las métricas de los modelos

Modelo	Presicion	Recall	F1	Acurracy
RB_MM_V3	0.8734	0.8755	0.8744	0.8830
Robertuito_V2	0.8731	0.8790	0.8758	0.8838
GG_V3	0.8799	0.8840	0.8818	0.8896
WW_V1	0.8778	0.8824	0.88	0.8879
LL_Anti_V1	0.8845	0.8644	0.8727	0.8846

0.7. Ensamble

Por ultimo para elevar la métrica objetivo se utilizo diferentes ensambles; sin embargo, el que mejor obtuvo resultados fue un ensamble en el que se incluyeron los 4 modelos presentados anteriormente. Para lograr esto, se generaron las probabilidades de cada uno de los modelos, está probabilidad era un numero decimal entre 0 y 1 que denota la probabilidad de que un tuit contenga humor o no. Para obtener la predicción se hizo una búsqueda de pesos para obtener el mejor resultado posible, el *Código:* **??**, muestra la implementación de la búsqueda e optimización de parámetros, así como la búsqueda del mejor umbral de decisión, con esto se obtuvo en el conjunto de validación un **F1** de $0{,}8940$ esto es una mejora bastante elevada respecto a los modelos anteriores, al probar en el conjunto de prueba el modelo **WW_GG_Robertuito_LL_V1** se obtuvo un **F1-score** de $0{,}8923$ esto es una mejora de $0{,}0105$ esto es un punto porcentual por lo que el ensamble exprime lo mejor de cada uno de los modelos. La *Tabla:* 6 muestra la comparación final de todos los modelos probados.

0.8. Resultados finales

El resultado final de la competencia, el modelo **WW_GG_Robertuito_LL_V1** obtuvo el tercer lugar de la competencia, la *Tabla:* 7 muestra la tabla con las clasificaciones finales.

Cuadro 6: Comparación de las métricas de los modelos

Modelo	Presicion	Recall	F1	Acurracy
RB_MM_V3	0.8734	0.8755	0.8744	0.8830
Robertuito_V2	0.8731	0.8790	0.8758	0.8838
GG_V3	0.8799	0.8840	0.8818	0.8896
WW_V1	0.8778	0.8824	0.88	0.8879
LL_Anti_V1	0.8845	0.8644	0.8727	0.8846
WW_GG_Robertuito_LL_V1	0.8939	0.8908	0.8923	0.9004

Cuadro 7: Clasificación final concurso CE 2026-I

Modelo	Precisión	Recall	F1	Accuracy
SPINOCORNIOSAURUS_1053	0.8909	0.9086	0.8974	0.9023
LYON	0.8939	0.8939	0.8939	0.9014
WW_GG_Robertuito_LL_V1	0.8939	0.8908	0.8923	0.9004
Prueba_Dieciseis_staking	0.8729	0.8824	0.8771	0.8843
DGNP	0.8708	0.8798	0.8747	0.8821
RB_MM_V3	0.8734	0.8755	0.8744	0.8830
GEMINI 3.5 Flash	0.8399	0.8659	0.8349	0.8375
LLM DeepSeek V3	0.8298	0.8550	0.8246	0.8273
BASELINE: SVM + TF-IDF	0.7914	0.7724	0.7796	0.8014
BASELINE: SVM + word embeddings fasttext	0.7545	0.7265	0.7350	0.7659
BASELINE: clasificación random	0.5039	0.5042	0.4957	0.5055

A pesar de haber obtenido el tercer lugar en la competencia, los modelos presentados en esta investigación son altamente competentes para la clasificación de tuits de humor. Por otro lado, el ensamble propuesto se encuentra únicamente por debajo del mejor modelo por $0{,}0051$, lo que no representa ni siquiera un punto porcentual, demostrando una alta competitividad frente a arquitecturas más grandes como *Gemini* [et al.(2025b)] o *DeepSeek* [et al.(2025a)].

Es importante destacar que, en comparación con los *baselines*, el desempeño de los modelos desarrollados supera ampliamente las métricas obtenidas por métodos tradicionales como SVM con TF-IDF o embeddings de *fastText*, alcanzando mejoras superiores al 10 % en *Accuracy* y más de 15 % en *F1-score*. Esto evidencia que la incorporación de arquitecturas modernas y técnicas de ensamble resulta fundamental para abordar tareas complejas de clasificación semántica.

Además, la consistencia entre las métricas de *Precisión*, *Recall* y *F1* sugiere que los modelos no solo logran un buen desempeño global, sino que también mantienen un equilibrio entre la detección de tuits humorísticos y no humorísticos, evitando sesgos hacia una sola clase. Este balance es crucial en aplicaciones prácticas donde la interpretación errónea de un mensaje puede afectar la calidad del análisis lingüístico.

Finalmente, los resultados obtenidos permiten concluir que los modelos desarrollados son competitivos frente a alternativas de mayor tamaño y costo computacional, ofreciendo una solución eficiente y escalable para la clasificación automática de humor en redes sociales.

Conclusiones

El desarrollo de esta investigación permitió demostrar que los modelos propuestos son altamente competitivos en la tarea de clasificación de tuits de humor. A pesar de que el modelo **WW_GG_Robertuito_LL_V1** obtuvo el tercer lugar en la competencia, su desempeño se encuentra prácticamente al nivel de los dos primeros lugares, con una diferencia menor a un punto porcentual. Este resultado evidencia que arquitecturas más ligeras y optimizadas pueden alcanzar rendimientos similares a modelos de mayor tamaño y complejidad, como *Gemini* o *DeepSeek*, pero con un costo computacional reducido.

Asimismo, la comparación con los *baselines* confirma la efectividad de las técnicas modernas de aprendizaje automático y ensamble, logrando mejoras significativas en métricas como *Accuracy* y *F1-score*. El equilibrio observado entre *Precisión* y *Recall* refuerza la robustez del sistema, evitando sesgos hacia una clase específica y garantizando un desempeño confiable en escenarios prácticos.

En conclusión, los resultados obtenidos validan la pertinencia de los modelos desarrollados y su potencial aplicación en el análisis automático de contenido humorístico en redes sociales.

Además, se abre la posibilidad de extender este enfoque hacia otras tareas de clasificación semántica, explorando futuras líneas de investigación como el uso de técnicas de *transfer learning*, la integración de representaciones multimodales.

Referencias

[Chiruzzo et al.(2020)Chiruzzo, Castro, and Rosá] Luis Chiruzzo, Santiago Castro, and Aiala Rosá. HAHA 2019 dataset: A corpus for humor analysis in Spanish. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5106–5112, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.628/.

[et al.(2025a)] DeepSeek-AI et al. Deepseek-v3 technical report, 2025a. URL https://arxiv.org/abs/2412.19437.

[et al.(2025b)] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2025b. URL https://arxiv.org/abs/2312.11805.

[et al.(2025c)] Gemma Team et al. Gemma 3 technical report, 2025c. URL https://arxiv.org/abs/2503.19786.

[et al.(2024)] Meta AI et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[et al.(2020)] Tom B. Brown et al. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[Hu et al.(2021)Hu, Shen, Wallis, Allen-Zhu, Li, Wang, Wang, and Chen] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

[Li et al.(2022)Li, Liu, and Wang] Zhuohang Li, Jiashuo Liu, and Yuci Wang. Performance analysis on deep learning models in humor detection task. In *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, pages 93–97, 2022. doi: 10.1109/MLKE55170.2022.00023.

[Liang(2011)] Ping Liang. Discourse analysis on humor. In *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, pages 5002–5005, 2011. doi: 10.1109/AIMSEC.2011.6011180.

[Maldonado and Rodriguez(2025)] Brian Maldonado and Leonardo Rodriguez. Clasificación de tweets de humor con redes neuronales. *UACM*, 2025.

[Maldonado(2026)] Brian Miguel Escalona Maldonado. Clasificación de tuits humorísticos: un enfoque eficiente basado en aprendizaje automático y ensamble. *Boletín UPIITA*, 21(114), Jul-Ago 2026. ISSN 20076150.

[Parthasarathy et al.(2024)Parthasarathy, Zafar, Khan, and Shahid] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities, 2024. URL https://arxiv.org/abs/2408.13296.

[Pérez et al.(2022)Pérez, Furman, Alemany, and Luque] Juan Manuel Pérez, Damián A. Furman, Laura Alonso Alemany, and Franco Luque. Robertuito: a pre-trained language model for social media text in spanish, 2022. URL https://arxiv.org/abs/2111.09453.

[Raskin(2017)] Victor Raskin. *Humor theory: What is and what is not*. Mouton de Gruyter, 2017.

[Team(2025)] Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[White et al.(2023)White, Fu, Hays, Sandborn, Olea, Gilbert, Elnashar, Spencer-Smith, and Schmidt] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023. URL https://arxiv.org/abs/2302.11382.

[Wu et al.(2024)Wu, Huang, and Lau] Shih-Hung Wu, Yu-Feng Huang, and Tsz-Yeung Lau. Humour classification by fine-tuning llms: Cyut at clef 2024 joker lab subtask humour classification according to genre and technique. In *Conference and Labs of the Evaluation Forum*, 2024. URL `https://api.semanticscholar.org/CorpusID:271793686`.