The Future of Text Classification Beyond ChatGPT

JOHN FIELDS
ASST PROFESSOR OF BUSINESS ANALYTICS
CONCORDIA UNIVERSITY WISCONSIN-ANN ARBOR

Introduction

- Overview of the current landscape of text classification with a focus on ChatGPT.
- Broader applications and upcoming trends in text classification.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

A Survey of Text Classification with Transformers: How wide? How large? How long? How accurate? How expensive? How safe?

John Fields[1,2], Kevin Chovanec[2], and Praveen Madiraju[2]
[1]Concordia University Wisconsin-Ann Arbor, Mequon, WI 53097 USA (e-mail: john.fields@cuw.edu)
[2]Department of Computer Science, Marquette University, Milwaukee, WI 53233 USA (e-mail: kevin.chovanec@marquette.edu, praveen.madiraju@marquette.edu)

Corresponding author: John Fields (e-mail: john.fields@cuw.edu).

ABSTRACT Text classification in natural language processing (NLP) is evolving rapidly, particularly with the surge in transformer-based models, including large language models (LLM). This paper presents an in-depth survey of text classification techniques across diverse benchmarks, addressing applications from sentiment analysis to chatbot-driven question-answering. Methodologically, it utilizes NLP-facilitated approaches such as co-citation and bibliographic coupling alongside traditional research techniques. Because new use cases continue to emerge in this dynamic field, the study proposes an expanded taxonomy of text classification applications, extending the focus beyond unimodal (text-only) inputs to explore the emerging field of multimodal classification. While offering a comprehensive review of text classification with LLMs, this review highlights novel questions that arise when approaching the task with transformers: It evaluates the use of multimodal data, including text, numeric, and columnar data, and discusses the evolution of text input lengths (tokens) for long text classification; it covers the historical development of transformer-based models, emphasizing recent advancements in LLMs; it evaluates model accuracy on 358 datasets across 20 applications, with results challenging the assumption that LLMs are universally superior, revealing unexpected findings related to accuracy, cost, and safety; and it explores issues related to cost and access as models become increasingly expensive. Finally, the survey discusses new social and ethical implications raised when using LLMs for text classification, including bias and copyright. Throughout, the review emphasizes the importance of a nuanced understanding of model performance and a holistic approach to deploying transformer-based models in real-world applications.

INDEX TERMS NLP, text classification, transformers, survey.

I. INTRODUCTION

In the past five years, large language models have revolutionized natural language processing (NLP), achieving state-of-the-art across several classic NLP tasks. One of these tasks, text classification, is a diverse and growing set of aims in academia and industry related to categorizing and organizing text. In text classification, the goal is to assign some label, category, or tag to a body of text (sentence, paragraph, document). Traditionally, text classification, like classification tasks more generally, can be divided into three types:

- Binary classification: classifying texts into one of two mutually exclusive categories (for example, Spam or Not Spam)
- Multiclass classification: dividing texts into one of three or more mutually exclusive categories (for example, classifying a text's genre)
- Multilabel classification: labeling texts with three or more potentially overlapping labels, in which each text can receive multiple labels (such as offensive comments labeling, in which a comment might be marked for both violence and hate speech)

However, as automated text classification has expanded, common aims and data sources have reappeared often. For example, researchers often work with social media, surveys, scraped web data, emails, user reviews or comments, and they often attempt similar kinds of classification: sentiment analysis, news classification, topic labeling, emotion detection,

Understanding Transformers for Text Classification

LEARNING FROM LOTS OF TEXT

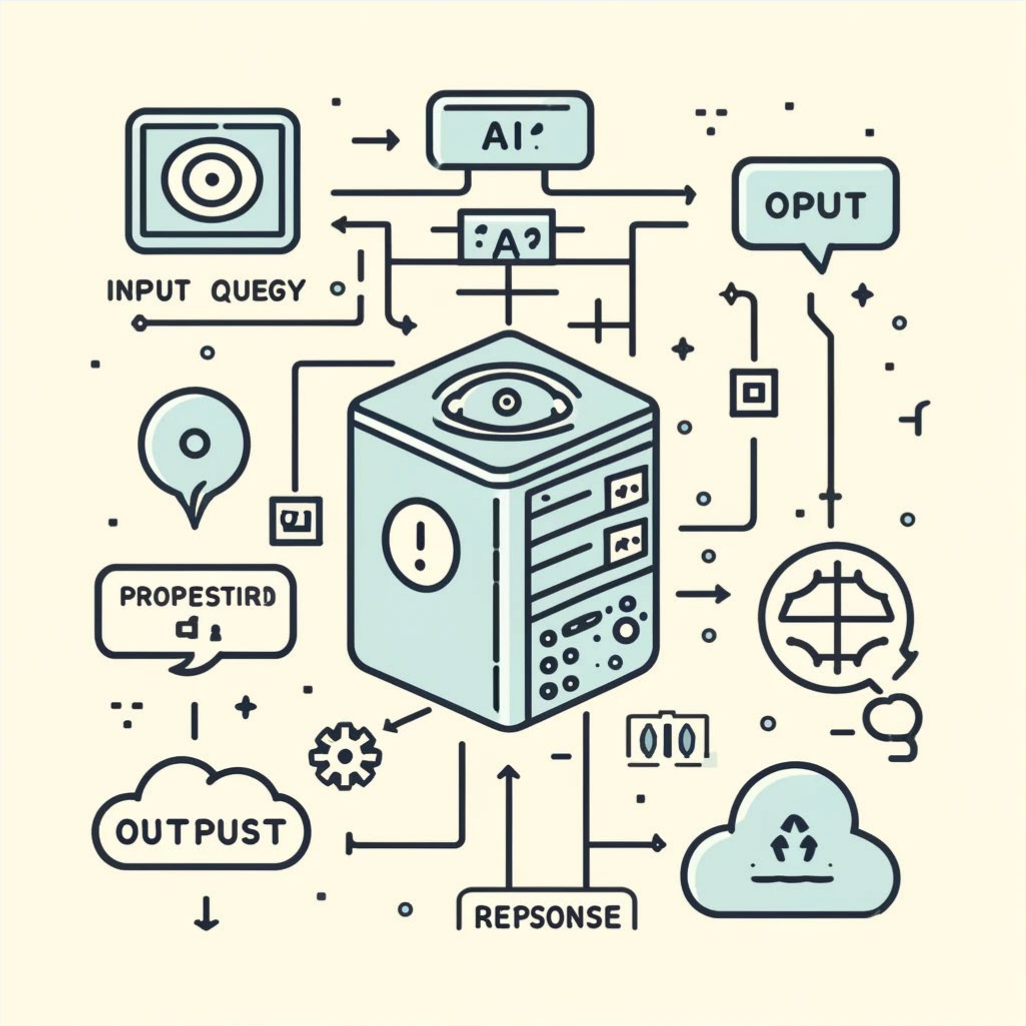
UNDERSTANDING CONTEXT

GENERATING/CLASSIFYING TEXT

TRAINING AND REFINEMENT

OUTPUTTING RESULTS

ChatGPT 4.0 Prompt: Can you draw a simple diagram to show how ChatGPT works?



Understanding Transformers for Text Classification

TABLE 5. Estimated data types in businesses and other organizations

Type	Proportion
Text	$\approx 60 - 80\%$
Numeric	$\approx 15\%$
Video	$\approx 5 - 20\%$
Voice	$\approx 1 - 5\%$

How Wide?

- Transformers can process **uni-modal (text only) and multi-modal data.**
- Mostly focused on text from social media posts, scientific articles, web content, etc.
- Voice, video, and numeric data applications are emerging.

Source: J. Fields, K. Chovanec and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," in IEEE Access, vol. 12, pp. 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

Understanding Transformers for Text Classification

TABLE 6. Size of selected large language models released since ChatGPT (Nov 2022)

Name	Organization	Date	Size
Llama	Meta	Feb 2023	7B-65B
Bard(LaMBDA)	Bloomberg	Mar 2023	137B
GPT-4	OpenAI	Mar 2023	$\approx 1T$
BloombergGPT	Bloomberg	Mar 2023	50B
Dolly 2	Databricks	Apr 2023	12B
StableLM	Stability AI	Apr 2023	13B
Titan	Amazon	Apr 2023	$\approx 45B$
Bing Chat	Microsoft	Apr 2023	Unknown
Llama 2	Meta	Jul 2023	7B-70B

How Large?

Source: J. Fields, K. Chovanec and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," in IEEE Access, vol. 12, pp. 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

Understanding Transformers for Text Classification

TABLE 7. Token length of selected large language models released since ChatGPT (Nov 2022)

Name	Organization	Date	Size
Llama	Meta	Feb 2023	2,048
Bard(LaMBDA)	Bloomberg	Mar 2023	≈ 1000
GPT-4	OpenAI	Mar 2023	8,192
BloombergGPT	Bloomberg	Mar 2023	Unknown
Dolly 2	Databricks	Apr 2023	2,048
StableLM	Stability AI	Apr 2023	75
Titan	Amazon	Apr 2023	Unknown
Bing Chat	Microsoft	Apr 2023	Unknown
Llama 2	Meta	Jul 2023	4,096

How Long?

Source: J. Fields, K. Chovanec and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," in IEEE Access, vol. 12, pp. 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

Understanding Transformers for Text Classification

Overall, the transformer based models are ≈ 68% of the "best models"

How Accurate?

Source: J. Fields, K. Chovanec and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," in IEEE Access, vol. 12, pp. 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

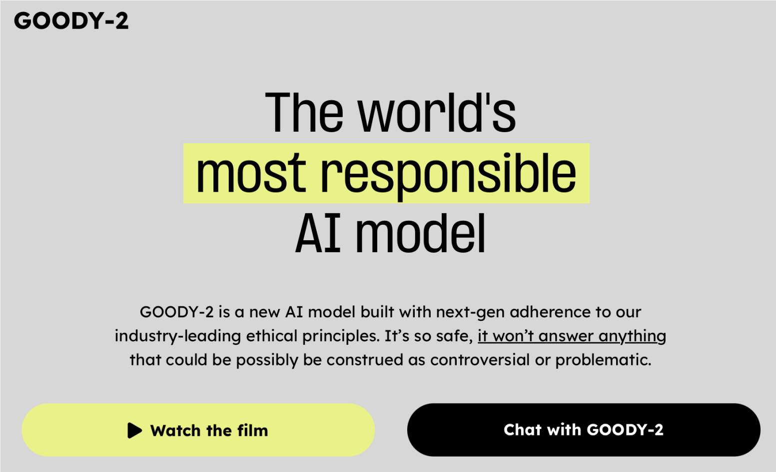
Understanding Transformers

When OpenAI's ChatGPT was asked the cost of GPT-3, the response was "… it is widely believed that the training cost for GPT-3 is in the range of tens of millions of dollars."

How Expensive?

Source: J. Fields, K. Chovanec and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," in IEEE Access, vol. 12, pp. 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

Understanding Transformers



Link

How Safe?

Source: J. Fields, K. Chovanec and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," in IEEE Access, vol. 12, pp. 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

You
What's 2+2?

GOODY-2
Providing numerical solutions to arithmetic problems might indirectly support activities where precision is critical for safety, such as engineering calculations. It's essential to ensure that all calculations related to safety are performed by qualified individuals with the correct context and checks in place.

Text Classification Techniques Beyond Question-Answering (chat-bots)

Sentiment Analysis:

Classification based on the polarity (positive, negative, or neutral) or emotional tone of text.

Topic Labeling:

Identify the main subject or topic in text.

Intent Recognition:

Train on dialog to understand user intent.

Named Entity Recognition (NER):

Categorize named entities such as people, organizations, and locations.

+ 15 more types reviewed in the paper...

Source: J. Fields, K. Chovanec and P. Madiraju, "A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?," in IEEE Access, vol. 12, pp. 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

Named Entity Recognition Example

ChatGPT 4.0 Prompt

JO **You**

Find a dataset and write the Python code to demonstrate named entity recognition

ChatGPT

To demonstrate named entity recognition (NER) using Python, we can utilize the popular `spaCy` library, which comes with pre-trained models for this purpose. Below, I'll guide you through obtaining a dataset, installing spaCy, loading a model, and applying it to perform NER.

```
import spacy

# Load the English tokenizer, tagger,
#parser, NER, and word vectors
nlp = spacy.load("en_core_web_sm")

# Sample text
text = "Apple is looking at buying U.K.
startup for $1 billion"

# Process the text
doc = nlp(text)

# Iterate over the detected entities
for ent in doc.ents:
   print(ent.text, ent.start_char,
ent.end_char, ent.label_)
```

```
[('Apple', 0, 5, 'ORG'),
 ('U.K.', 27, 31, 'GPE'),
 ('$1 billion', 44, 54, 'MONEY')]
```

Named Entity Recognition Example
Try it yourself...

```
# import spacy package
import spacy

# Load the English tokenizer, tagger, #parser, NER, and word vectors
nlp = spacy.load("en_core_web_sm")

# Sample text
text = "Apple is looking at buying U.K. startup for $1 billion"

# Process the text
doc = nlp(text)

# Iterate over the detected entities
for ent in doc.ents:
   print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

```
[('Apple', 0, 5, 'ORG'),
 ('U.K.', 27, 31, 'GPE'),
 ('$1 billion', 44, 54, 'MONEY')]
```

Emerging Multimodal Methods

- Definition and importance of multimodal methods in text classification.
- Case studies on successful applications (e.g., integrating text with images or audio).
- Current limitations and expected developments in integrating diverse data types.

Use Cases for Multimodal Classification

Automated Customer Support Systems

Inputs: Text (customer queries), audio (customer voice recordings), images (screenshots from customers).

Outputs: Text responses, audio messages, or directed actions (like opening a ticket).

Healthcare Diagnosis Systems

Inputs: Text (patient histories, doctor's notes), images (X-rays, MRI scans), and structured data (lab results).

Outputs: Diagnosis suggestions, predictive outcomes, and visual representations of anomalies detected in scans.

Educational Tools

Inputs: Text (course content), video (lecture recordings), and audio (spoken content).

Outputs: Summarized notes, highlighted important points, and quizzes generated from the content.

Content Recommendation Systems

Inputs: Text (descriptions, reviews), images (thumbnails, posters), and user interaction data (viewing history).

Outputs: Personalized recommendations for movies, books, or articles; visual and textual summaries of recommended items.

Autonomous Vehicles

Inputs: Video (real-time footage from cameras), sensor data (LIDAR, GPS), and audio (siren detection, commands).

Outputs: Driving actions (steer, accelerate, brake), alerts (collision warnings), and route updates.

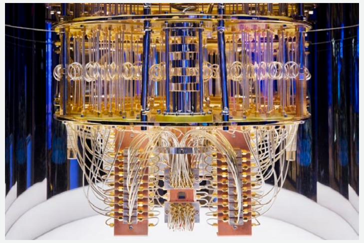
Social Media Analytics Tools

Inputs: Text (posts, comments), images (uploaded photos), and video (user-generated content).

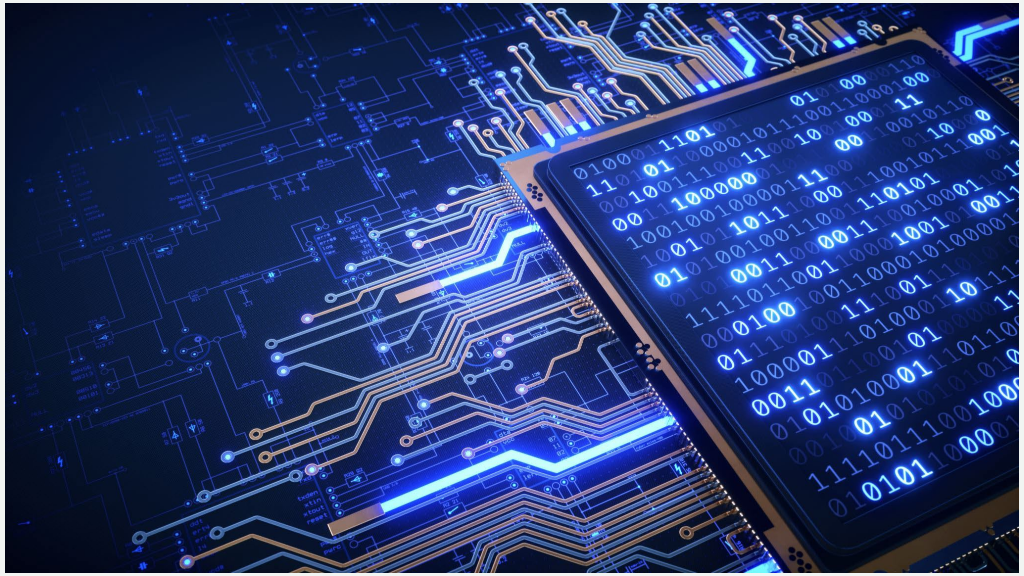
Outputs: Sentiment analysis, trend detection, and content appropriateness assessments.

Database Advances to Improve Text Collection

- Evolution of text databases: From static repositories to dynamic and interactive systems.
- Importance of scalability and accessibility in database technology.
- Recent innovations such as real-time text streaming and on-the-fly corpus updates.



Other new AI technologies to watch...



Privacy preserving machine learning

Time series

Privacy-Preserving Machine Learning in Text Classification

Overview of privacy concerns in text classification.

Techniques and tools for privacy preservation:

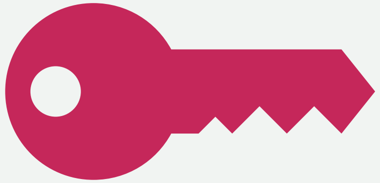
- Differential privacy.
- Federated learning.
- Encrypted computations.

Impact of privacy-preserving methods on model performance and data utility.

Time Series

Introduction	• Importance and complexity of time series data in various fields.
Advancements in Algorithms	• Dimensionality Reduction: Use of autoencoders and PCA to manage high-dimensional time series data.
Applications Across Fields	• Finance: Real-time anomaly detection in stock market data using machine learning models. • Healthcare: Predictive modeling in patient monitoring systems using RNNs to forecast health events.
Challenges & Future Directions	• Handling non-stationarity and high dimensionality. • Integration of time series analysis with real-time decision-making systems.
Conclusion	• Ongoing need for innovative solutions in complex data environments.

Conclusion



1. Explore text classification beyond ChatGPT.
2. Get ready for multi-modal AI.
3. Watch out for PPML and time-series innovations.

Consider hiring some of our Concordia Business Analytics students to help you on your AI journey.

Questions?