# Developing a Machine Learning-Based Classifier for Identifying Exceptional Facts in Text

## 1. Introduction

In an era of information abundance, the ability to distinguish exceptional facts from the sea of textual data is paramount. Being able to recognize these unique and significant pieces of information can greatly impact fields like academia, journalism, and decision-making. This proposal outlines a comprehensive capstone project dedicated to developing a methodical and effective approach for identifying exceptional facts from diverse textual sources, including but not limited to Wikipedia articles and news websites. The project's multifaceted methodology involves systematic processes, human annotation, machine learning, and rigorous evaluation.

## 2. Objective

The primary objective of this capstone project is to construct a robust and reliable classifier capable of discerning exceptional facts embedded within textual content. By exceptional facts, the reference is to unique and significant pieces of information that stand out due to their rarity, relevance, or importance. These facts are defined based on their alignment with the criteria established by the Maverick platform, renowned for its exceptional fact generation. By addressing the real-world significance of identifying exceptional facts, this project aims to contribute meaningfully to academia, journalism, and decision-making processes.

## 3. Methodology

The project methodology is structured into three essential phases, each contributing to the creation of a proficient exceptional fact classifier.

### 3.1 Sentence Collection from Textual Sources

The initial phase of the project involves extraction of sentences from a wide spectrum of textual sources. This phase is guided by a combination of predefined templates and keywords. These templates can either be shaped by the project's specific requirements or derived from Maverick's existing exceptional facts. The emphasis lies in curating sentences that convey factual information while focusing on the distinctive exceptional facts akin to those recognized by Maverick.

### 3.2 Human Annotation of Sentences

Following the sentence collection, human annotators will engage in an annotation process. During this phase, sentences will be reviewed and labeled as either "Yes" or "No" in terms of their alignment with Maverick's definition of exceptional facts. This annotated dataset will play a pivotal role in training the machine learning classifier to accurately identify exceptional facts.

### 3.3 Classifier Development and Sentence Classification

The pivotal aspect of the project revolves around the creation of a machine learning classifier. This sophisticated classifier, empowered by the annotated dataset, initiates an intricate process of learning and recognition. Employing a combination of supervised learning and pattern recognition techniques, the classifier acquires the ability to distinguish sentences aligning with exceptional fact patterns from those that do not. The training dataset, enriched with labeled sentences, serves as a guiding beacon for the classifier's evolution, enabling it to comprehend and identify the core attributes that differentiate an exceptional fact. This final step involves the construction of a machine learning classifier, leveraging the annotated dataset obtained in the previous phase. The classifier's primary objective is to discern sentences that adhere to the distinctive patterns of exceptional facts from those that deviate. Through training on the annotated sentences, the classifier acquires the proficiency to identify the defining traits that characterize an exceptional fact.

## 4. Evaluation and Fine-tuning

The project's reliability and effectiveness are upheld through a meticulous evaluation process. This process involves using key measures such as precision, recall, and F1-score to gauge the performance of the classifier. These measures serve as standards for quantifying the accuracy and proficiency of the classifier in identifying exceptional facts.

Furthermore, a commitment to refining the classifier's performance is demonstrated through an iterative approach. A continuous improvement process is maintained through feedback loops. These loops involve systematically gathering feedback from the outcomes of the classifier. This feedback is then utilized to fine-tune the classifier. The purpose of this fine-tuning is to enhance accuracy, mitigate potential biases, and strengthen the classifier's capacity to identify a wide range of exceptional facts within various forms of textual content.

Ultimately, this comprehensive evaluation and iterative fine-tuning process ensure that the classifier not only meets but surpasses performance expectations. It achieves this by effectively uncovering exceptional facts across diverse textual contexts.

**5. Significance and Implications**

The results of this project have implications that go beyond just its initial use. By following a structured approach and successfully putting it into practice, the project contributes to improving how we understand and identify exceptional facts. This concept is crucial in many different areas. For instance, it matters in academia, where researchers want to get more meaningful information, and in journalism, where being able to spot exceptional facts helps create more influential stories. This project essentially empowers experts in various fields to gain deeper insights and generate more significant content by accurately recognizing exceptional facts.

**6. Future Prospects**

This project establishes a robust foundation for potential future innovations. The creation of a systematic framework for identifying exceptional facts lays the groundwork for the development of automated tools. These tools are designed to efficiently extract critical insights from extensive textual datasets, streamlining the process of extracting valuable knowledge.

The project's outcomes have a resonance that extends beyond its immediate applications. They hold relevance within the broader landscape of information discovery. By enhancing the identification of exceptional facts, this project contributes to a deeper understanding across various subjects. This, in turn, charts a trajectory for innovative advancements that amplify the effectiveness of informed decision-making and the enrichment of knowledge across diverse sectors and industries. Ultimately, this project's influence reaches beyond its present boundaries, setting the stage for a future that is characterized by well-informed decision-making and a broader knowledge base.

## 7. Expected Deliverables

The project's anticipated outcomes encompass a set of valuable deliverables that reflect the comprehensive nature of its objectives and contributions. These deliverables include:

### 7.1 A Fully Functional and Well-Tuned Machine Learning Classifier

The project will yield a classifier that is not only fully operational but has also undergone thorough tuning to ensure its accuracy and effectiveness. This classifier will serve as a practical tool for identifying exceptional facts within textual sources.

### 7.2 An Annotated Dataset of Exceptional Facts

A carefully curated dataset will be created, consisting of sentences accurately categorized as exceptional facts or otherwise. This dataset will serve as a crucial resource for training the classifier and assessing its performance, enhancing its ability to distinguish exceptional facts within diverse textual content.

### 7.3 Detailed Evaluation Reports

The project will generate comprehensive reports that provide in-depth insights into the classifier's performance. These reports will utilize metrics such as precision, recall, and F1-score to measure the classifier's effectiveness in identifying exceptional facts.

### 7.4 Comprehensive Methodology Documentation

A detailed documentation of the project's methodology will be provided, outlining the step-by-step approach taken during the project's execution. This documentation will not only cover technical aspects but also highlight the broader implications and significance of the project's outcomes.

These deliverables collectively represent the tangible results of the project, ensuring that its findings are transparent, well-documented, and ready to be applied in practical contexts and future research endeavors.

## 8. Conclusion

In summary, this proposal encapsulates a multifaceted project with a central objective of crafting an advanced classifier capable of recognizing exceptional facts within diverse textual sources. The meticulously structured approach, encompassing stages such as sentence collection, human annotation, and classifier development, underpins the potential to significantly elevate the realms of knowledge extraction and informed decision-making across a spectrum of domains. The anticipation of future advancements and the far-reaching implications of this endeavor underscore its profound significance within the dynamic landscape of information exploration and discovery. As such, this project stands as a testament to the pivotal role that systematic exceptional fact identification can play in advancing various fields of study and application.

**Student Name:** Alishbah Fahad
**Student ID:** 1001924185
**Project Advisor:** Dr. Chengkai Li
**Professor's Email:** cli@uta.edu
**Date:** 09/03/2023