UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Astronomical Data Features Extraction and Citation Prediction

Author:
Vladyslav KUTSURUK

Supervisor:
Oleksii IGNATENKO

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2023

Declaration of Authorship

I, Vladyslav KUTSURUK, declare that this thesis titled, "Astronomical Data Features Extraction and Citation Prediction" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"He who is fixed to a star does not change his mind. "

Leonardo da Vinci

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Astronomical Data Features Extraction and Citation Prediction

by Vladyslav KUTSURUK

Abstract

Natural Language Processing methods present promising opportunities for analyzing astronomical data, enabling the extraction of essential information from vast amounts of observations. Yet, applying these techniques to astronomical data presents notable challenges, including the difficulty of astronomical terminology and the diverse range of data sources. In this research, we leverage multiple Natural Language Processing techniques to extract information from astronomical observations with a specific focus on predicting the future citation rate of astronomical telegrams. To achieve this, we create a comprehensive dataset gathering astronomical messages from various sources and utilize techniques such as Named Entity Recognition, doc2vec, word2vec, and topic extraction. Along with this, we enhance the extracted information by incorporating manually created features that capture the characteristics of astronomical telegrams beyond their direct context. These features aim to provide a comprehensive representation of the messages. We then use all the extracted information to predict the future impact of the telegrams, as indicated by their citation counts, using multiple Machine Learning techniques.

Acknowledgements

I would like to express my gratitude to my parents for their encouragement and support. I would also like to extend my appreciation to my supervisor, Oleksii Ignatenko, for his guidance and mentorship through this project.

Contents

List of Figures

List of Tables

List of Abbreviations

ML	**M**achine **L**earning
NLP	**N**atural **L**anguage **P**rocessing
NN	**N**eural **N**etwork
IE	**I**nformation **E**xtraction
IR	**I**nformation **R**etrieval
NER	**N**amed **E**ntity **R**ecognition
LLM	**L**arge **L**anguage **M**odel
ATL	(the) **A**stronomer's **Te**l**e**gram
GCN	**G**amma-ray **C**oordinates **N**etwork
GRB	**G**amma **R**ay **B**urst
TDAC	**T**ime-**D**omain **A**strophysics **C**orpus
CRF	**C**onditional **R**andom **F**ields
QA	**Q**uestion **A**nswering
LGBM	**L**ight **G**radient-**b**oosting **M**achine
RF	**R**andom **F**orest
BERT	**B**idirectional **E**ncoder **R**epresentation (from) **T**ransformer
RB	**R**ule-**B**ased
CBOW	**C**ontinuous **B**ag-o**f**-**W**ords
UMAP	**U**niform **M**anifold **A**proximation and **P**rojection
t-SNE	**t**-distributed **S**tochastic **N**eighbor **E**mbedding
SVM	**S**upport **V**ector **M**achine

Chapter 1

Introduction

1.1 Context

The Astronomer's Telegram (ATL) is a free online service that rapidly communicates new astronomical observations and discoveries to the professional astronomical community. It was founded in 1997 by a group of astronomers who recognized the need for a more efficient way of sharing information about transient astronomical events, such as supernovae, gamma-ray bursts, and other phenomena that appear suddenly in the sky.

ATL is a web-based bulletin board where astronomers can submit short messages describing their observations and discoveries in real-time. One of the key advantages of the ATL is its speed. The service is designed to provide rapid communication of new discoveries, with messages typically posted within hours or even minutes of the observations being made. This allows astronomers worldwide to quickly follow up on new discoveries and conduct further observations and analyses. But the other side of this feature is volume - the number of such messages is significant and increasing every day.

The Astronomer's Telegram has become an essential tool for professional astronomers and has contributed to many significant discoveries in the field of astronomy. It has also helped to foster collaboration and communication among astronomers and has enabled the rapid dissemination of new information about the universe to the broader scientific community.

Another source of data is GCN Circulars. It is a dataset of astronomical circulars released by the Gamma-ray Coordinates Network (GCN). The GCN system distributes two types of messages: notices containing information about the location of gamma-ray bursts (GRBs) and other transients obtained from various spacecraft and circulars containing information about follow-up observations made by optical, radio, X-ray, and TeV observers from ground-based and space-based sources. ATels are typically released within hours or days of observation and provide a quick overview of the discovery and initial analysis. GCN circulars are more comprehensive reports released after a more thorough analysis of the observations. Both report types are essential sources of information for astronomers and astrophysicists studying transient phenomena such as supernovae, gamma-ray bursts, and others.

These short messages contain much information hidden in enormous volumes of unstructured (and semi-structured) data, making it difficult to get meaningful insights reading them as is, even for scientists. We aim to apply NLP techniques to these datasets to solve several significant problems, interesting for researchers in the field.

1.2 Goals of the Master Thesis

Our goals in this research are the following:

1. Our goal is to create a clean and tagged dataset that is suitable for analysis using the mentioned data sources. Additionally, we will develop the required functionality that enables the population of this dataset with new data, ensuring its continuous updates and relevance.

2. We aim to develop ML solutions that can extract interesting information from astronomical telegrams. This includes identifying the telescopes and sources (the observed astronomical objects or events) mentioned in the observations and capturing the related topics discussed in the telegrams. In addition to the ML-based solutions, we seek to create comprehensive features that provide a holistic description of the telegrams, going beyond their content.

3. We want to solve the citation prediction problem. The thing is that number of reported observations is growing, and usually, only a few of them discover new events. These few telegrams then are cited in others or (later) in papers. Hence, we aim to use the mentioned above information to predict the informational value of astronomical telegrams, which is expressed as the citation rate of the latter.

Chapter 2

Related Work

2.1 Related Researches Overview

Machine learning is a powerful tool that has become increasingly popular in analyzing astronomy data in recent years. ML algorithms can automatically identify patterns and relationships within large data sets, allowing astronomers to make more accurate and efficient predictions and classifications. ML techniques apply in many areas of astronomy, including star and galaxy classification, data analysis, and image processing. One of the critical advantages of ML in astronomy is its ability to handle large amounts of data with many variables, which can be challenging to analyze using traditional statistical techniques. ML algorithms can also learn from the data, improving their accuracy and efficiency over time as more data becomes available. However, this paper focuses on NLP methods primarily, so we can address an excellent overview Baron, 2019 of general ML methods in astronomy and proceed to our topic.

Applying NLP methods to astronomical data is relatively unexplored for now. One of the first examples is the paper Murphy and Curra, 2006, where authors presented a corpus of approximately 200,000 words of text from astronomy articles, manually annotated with about 40 entity types of interest to astronomers. The authors report on the challenges in extracting the corpus, defining entity classes and annotating scientific text. They investigate which features of an existing state-of-the-art Maximum Entropy approach perform well on astronomy text and achieve an F-score of 87.8%.

The authors also discuss the advantages of the astronomy domain for NER, such as being representative of the physical sciences, having freely available papers, interesting entity types to annotate, and existing databases of astronomical objects. The paper also reviews comparable named entity corpora, discusses aspects of astronomy that make it challenging for NLP, and presents examples of interesting cases of ambiguity in the astronomical text. Finally, the authors describe experiments with retraining an existing Maximum Entropy tagger for astronomical named entities and use the tagger to detect errors and inconsistencies in the annotated corpus.

Another recent paper Grezes, 2021 discusses the development of a language model based on Google's BERT deep neural network transformer architecture, called astroBERT. The aim of the project is to automate the process of identifying and tagging named entities within the ADS (Astrophysics Data System) database, which is used by the astrophysics community. The authors used a large collection of recent astronomy ADS papers to train astroBERT, which outperformed BERT on the named entity recognition task on ADS data. The paper also discusses the related work on BERT and SciBERT, as well as the technical details of data preparation and training of the model.

Finally, authors Alkan, 2022 propose using Natural Language Processing (NLP) to extract and summarize information from astronomical reports. They introduce TDAC, the first publicly available corpus based on astrophysical observation reports for named entity recognition in time-domain astrophysics. The paper also highlights the differences between astrophysics corpora and characterizes the discourse used in astrophysics through corpus analysis. The Astronomy Bootstrapping Corpus, the Astro Corpus, and the DEAL Shared Task Corpus are discussed as examples of existing annotated corpora for astrophysical named entity recognition.

2.2 General Methods Overview

2.2.1 Text Information Extraction

Ideologically, there are 3 main Information Extraction (IE) approaches: rule-based, machine-learning-based (also statistical) and hybrid.

The classical approach to the IE task is a rule-based approach. Chiticariu and Reiss., 2013 make a comprehensive review of its pros and cons. The stated pros of a rule-based IE approach include that it is easy to comprehend, easy to incorporate domain knowledge, and easy to trace and fix the cause of errors. However, it also requires tedious manual work and may not be as scalable as other approaches.

We can also take a look at Wu, 2022 as an example of the latest successful incorporation of the rule-based IE. The paper proposes a rule-based approach to the IE task in the professional mechanical, electrical, and plumbing (MEP) domain. The proposed algorithms include matching algorithms for named entity recognition (NER) and relationship extraction The paper also introduces two novel ideas, "meta linking" and "path filtering," for discovering out-of-pattern entities/relationships. A comparison experiment shows that the proposed rule-based approach outperforms the selected deep learning NER models by 37% and 49% in extraction precision. This work coincides with ours, as it is also solving a very domain-specific IE problem.

The basic rule-based IE approach is incorporated in our project by introducing a set of pre-defined Regular expressions which are used to extract data on both the data preparation step and the feature extraction step.

Speaking of the machine-learning (statistical), the field classic is Peng and McCallum., 2006. At the time published, the CRF were able to outperform that time IE titans: hidden Markov models and SVM classifiers. The research took part in 2006 but the conditional random fields are still widely used in the IE hybrid techniques and NER problems (Settles, 2004, Klinger, 2011).

In this research, the CRF is used as a layer of a Neural Network NER model provided by Spacy.

Another more novel IE approach is presented by the Jiang and Chen., 2023. The authors propose automatic information extraction using entity recognition techniques to ease the burden of paper reading for AI researchers. The proposed approach involves creating a manually annotated dataset called the ACER dataset and utilizing the GIA-PME model, which uses a gated interaction attention mechanism and probability-matrix encoding to enhance entity recognition. The model achieves the best performance compared to existing models and significantly improves the F1 score on the ACER dataset (according to the Jiang and Chen., 2023 results).

A comprehensive review of the Deep Learning IE methods is provided by Nguyen, 2018. The author develops deep learning models for various information extraction problems, such as entity mention detection, relation extraction, and event detection as a counterpart to the traditional approaches which involve hand-designing large

feature sets and are limited and expensive. The experiments demonstrate the effectiveness of the proposed methods, particularly in domain adaptation and transfer learning settings. The proposed methods show promise in automating the representation learning process for efficient and effective information extraction. The author also achieved joining frameworks for solving multiple problems simultaneously.

The other promising method in the IE field is question-answering (QA) using large language models (LLM). The LLM's ability to understand context and reason over natural language text makes them an effective tool for extracting insights from complex and even domain-specific datasets. The recent researches in this direction Wei, 2023 and Pereira, 2023 show superior LLMs performance in IE, even with zero-shot prompting (Wei, 2023). Dunn, 2022 presented a simple sequence-to-sequence approach using GPT-3 to jointly extract named entities from complex scientific text. The approach is fine-tuned on approximately 500 pairs of prompts and completions and can extract information from single sentences or whole abstracts/passages. They were able to extract very structured information from the scientific text in both native English and JSON formats. Their targets and data type are very similar to what we are dealing with in this project and hence might be a very good example to follow up.

We also employ the power of pre-trained LLMs to solve the topic extraction problem in a fashionable manner.

2.2.2 Domain-Specific Text Classification

Although the citation prediction problem is being addressed using both classification and regression approaches, we prefer a classification one more due to its ease of understanding and interpretation of the results. We will consider the domain-specific text classification problem for the related work references in this subsection.

A problem somehow similar to ours was solved by Liu, 2005. This paper proposes a statistical method for extracting domain-specific terms from the scientific corpora. The method takes into account the distribution of a candidate word within domains using entropy impurity. The process includes a normalization step to deal with unbalanced corpora. The extracted domain-specific terms are applied in text classification as the feature space and outperform traditional methods in experiments.

Another domain-specific classification problem was solved by Wu, 2020. The paper proposes a method to incorporate domain-specific information into meta-embeddings, which have shown superior performances across different NLP tasks. Experiments on four text classification datasets demonstrate the effectiveness of the proposed method. Basically, their statement is "You have to train your own embeddings".

Benballa and Picot-Clemente., 2019 proposed a very interesting key to the text classification solution, combining the state-of-the-art LLMs with classical hand-crafted features. The paper outlines the three steps involved in the system: feature creation, dynamic meta-embedding, and combining information to classify tweets for hate speech. Despite the fact that some of the hand-crafted features such as bag-of-POS-tagging or bag-of-emoji were not of interest for the predictions and were reducing the F1 score (as described in Benballa and Picot-Clemente., 2019 experiments), the idea to enhance the word embeddings with some classical NLP features is worth taking note of. It is like "You have to train your own embeddings and not forget about the classical NLP features".

And while the previously mentioned papers were aimed to solve the domain-specific text classification problem using all the available arsenal of tools, including the products of Deep Learning Transformers, these researchers from London and Toronto Wahba and Steinbacher., 2023 have proven that using complex, attention-based embeddings is not always necessary. Their main statement is that the Support Vector Machine (SVM) classifier with Tf-Idf vectorization can perform comparably to state-of-the-art models such as pre-trained language models like BERT. Believe it or not, the paper argues that the monosemic nature of specialized words in the domain-specific text makes the use of contextualized embeddings less necessary. And their research is live evidence that there are cases, where a simpler and more explainable model can achieve results similar to a Wu, 2020.

Last but not least Xie, 2022 - straightforward research on domain text classification methods based on BERT. The researchers proposed a method based on word embeddings to solve the text classification problem in specific domains. The proposed BERT "VCA" model segments long domain texts into short sentence sequences, inputs them into the BERT to obtain word vectors, compresses the sequence vectors, and combines them with the Encoder layer to extract important domain features. The experiments showed an F1 increase of 1.2% and emphasized the need for a domain-specific pre-training. The BERT-based solution might not be the most explainable one, but it definitely can be one of the fastest considering the use of already pre-trained embeddings.

After conducting a review of the literature and related research, several key conclusions can be drawn. Firstly, astronomer telegrams have emerged as relatively new sources of information and have gained popularity over the past decade. Astronomers themselves are increasingly interested in tracking significant and rare observations, but the growing volume of telegrams creates challenges in identifying valuable insights. Secondly, the field of astronomical NLP is still in its early stages of development, although there have been notable projects (as described in the **Related Researches Overview**) that have utilized various NLP techniques for astronomical exploration.

The main objective of this research is to evaluate and apply different NLP techniques in order to understand their effectiveness in the field and apply them to solve the prediction problem at hand.

Chapter 3

Dataset

3.1 Dataset Creation

The dataset for this project was collected and composed from two primary sources of astronomical telegrams mentioned in the **Introduction**: ATel and GCN. The telegrams gathered on the corresponding resources will be referenced later in the text as ATels and GCNs.

The ATel website was parsed using the html-parsing libraries in Python and the underlying information was transformed into a more structural representation using a comprehensive set of Regular expressions. GCN website, however, already provides functionality to download all existing telegrams as one tar achieve. Each of the GCN telegrams is saved in a plain text format, with a few upper-case titles that help to identify the subject or the date of a telegram. Another set of Regular expressions was applied to extract the GCNs' data. The data from both sources was consolidated into a single dataset, incorporating all the necessary transformations to ensure a standardized format. This involved converting dates into a unified format, parsing author(s) information, subjects, etc.

Another challenge we had to face is that the same astronomical observation might be posted in multiple sources by the same reporter with a few differences in the content. To address this issue, we employed the Tf-Idf approach to vectorize the bodies of GCNs and ATels. These vector representations were then analyzed using pairwise cosine similarity metric, resulting in a matrix of size $D_{num\_atels} \times D_{num\_gcns}$, where each element represents the similarity between the corresponding ATel and GCN bodies. To identify probable duplicates, we set a similarity threshold of 0.9 based on empirical observations. In cases where a GCN and an ATel had a similarity score above this threshold, we prioritized the GCN and removed the corresponding ATel from the final dataset. By applying this step, we were able to identify and remove 258 duplicated telegrams. This process ensured that the final dataset was cleaned from duplicated instances, which could have potentially worsened the results of the prediction algorithm due to variations in citation rates across different sources.

As mentioned previously, the aim of this research is to analyze the informational value of a telegram. However, due to the broad nature of this concept, we have decided to measure it in terms of the citation or mention rate of a telegram. To achieve this, we conducted an analysis of the bodies of the telegrams and collected the mentions of other telegrams from them in order to define the target variable for our machine-learning problem. Good for us, there are agreed rules on how the telegrams should be referred to. The astronomers include the number and the source of the cited telegram in their report, e.g.: "The object identified in the previous epoch of imaging (GCN 9884)" or "This agrees reasonably well with previous photometric estimates of Zharikov et al. 1998 (ATel #34, #35)". Hence, by defining the multiple

Regex rules we were able to collect the vast majority of the direct possible mentions and citations. Worth to say, that GCNs are sometimes referring to the ATels and vice versa. To avoid human bias, it was important for us to exclude self-mentions from our analysis. Self-mentions refer to the situations where authors cite their own previous work in their posts.

3.2 Dataset Description

In total, around 15000 ATels and 33000 GCNs were collected and processed resulting in a dataset of the size of 48000 records. The collected ATels and GCNs were unified in their structure and the dataset end up having the following fields:

- *telegram_no* - the exact number of the telegram concatenated with the telegram source. E.g.: 134_*atel* or 135_*gcn*
- *date* - the telegram post date
- *subject* - the telegram subject
- *body* - the actual content of a telegram
- *from* - the telegram's author(s) credential information (an email). This field was chosen as a more appropriate and convenient replacement of the exact author(s) name(s) and surname(s).
- *atel_references* - defines which other ATels are cited in this telegram
- *gcn_references* - defines which other GCNs are cited in this telegram

A brief examination of the temporal distribution of telegrams, as illustrated in 3.1, indicates that the popularity of GCN and ATel was relatively low prior to 2005. This observation suggests that the earlier years may hold less significance in relation to the prediction problem. Additionally, an analysis of telegram volumes on a monthly basis helps to yield worthwhile insight, which will be further used in the feature engineering process.

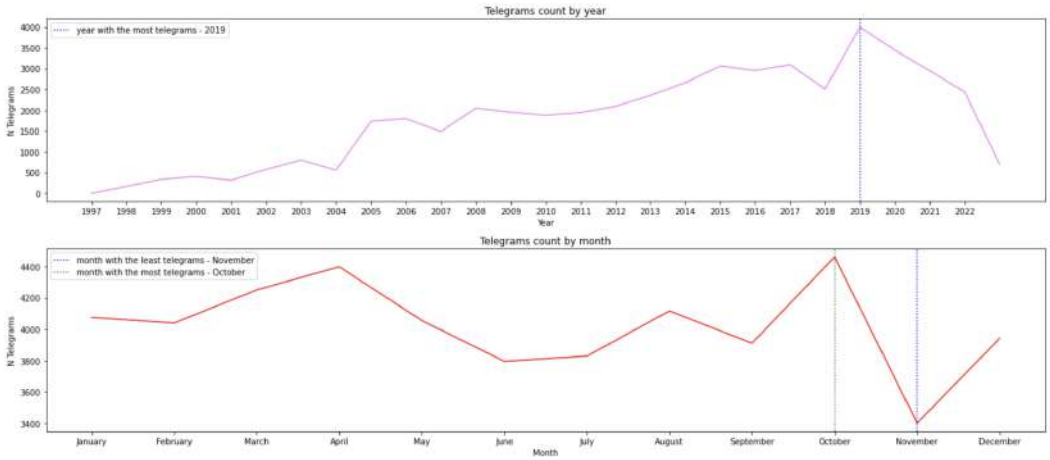

FIGURE 3.1: Temporal Distribution of Telegrams

3.3 Target Value and Labels Assignment

The citation rate of each telegram was calculated as the number of telegrams that make reference to it. For instance, if ATel #101 is citing ATel #100, ATel #100 will get a +1 to its citation rate.

If we take a look at the distribution of a target value (Figure 3.2), we can notice that 75% of the telegrams are cited no more than once. The purpose of defining the target value is to align with the primary hypothesis, which says that telegrams with higher citation rates are likely to contain more interesting observations. To set a clear threshold, we decided to consider a telegram potentially interesting only if it is cited at least three times which corresponds to the top 20% percentile of the citation rate.

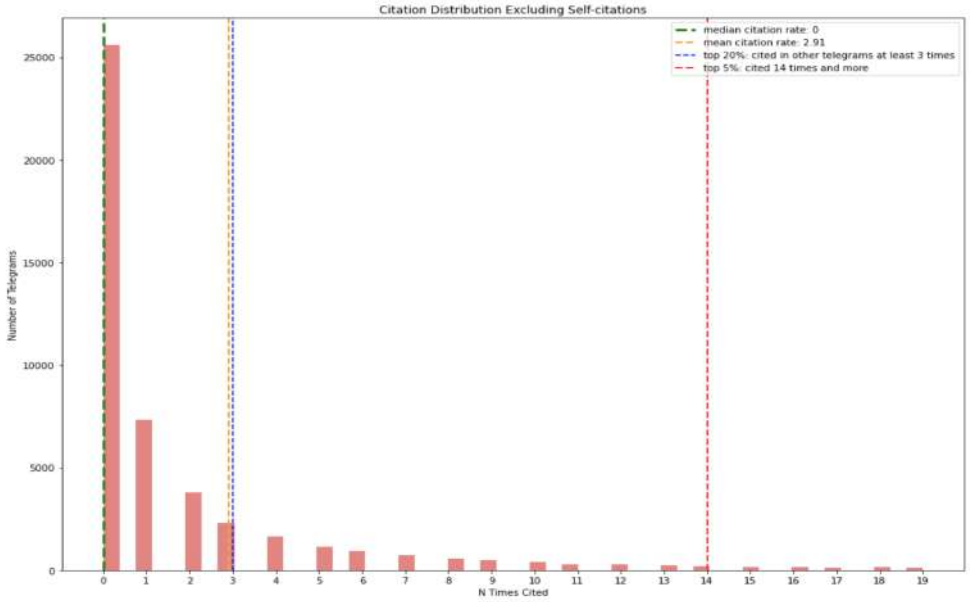

FIGURE 3.2: Citation Rate Distribution

The target variable for the classification problem was divided into three classes: very interesting (class 2), interesting (class 1) and not interesting (class 0). The distribution of the target labels shown in Figure 3.3 indicates that only 20% of the telegrams were labelled as interesting while the remaining 80% were labelled as not interesting. This suggests that the classification problem is imbalanced, with significantly more instances belonging to the majority class than the minority class. We will discuss our approach to handling class imbalance in the **Experiments** chapter.

Note: Later in the research, we will consider merging classes 1 and 2 into one, turning the classification problem into a binary one. This will allow us to reduce the impact of the over-cited telegrams and concentrate only on separating interesting telegrams from not interesting, as stated in the research goals.

Based on the distribution depicted in Figure 3.4, certain years exhibit significantly higher citation counts, indicating a greater overall interest in the corresponding topics. Notably, the year 2015 stands out with a notable increase in citations, which can be linked to the discovery of gravitational waves. Also, the year 2010 has the highest average citation count, likely due to the discovery of the first known Earth Trojan asteroid, indicating its significance in the field.

Examining the insights by month reveals specific periods when interesting cosmic objects are more frequently observable. For instance, in June, the Galactic Center,

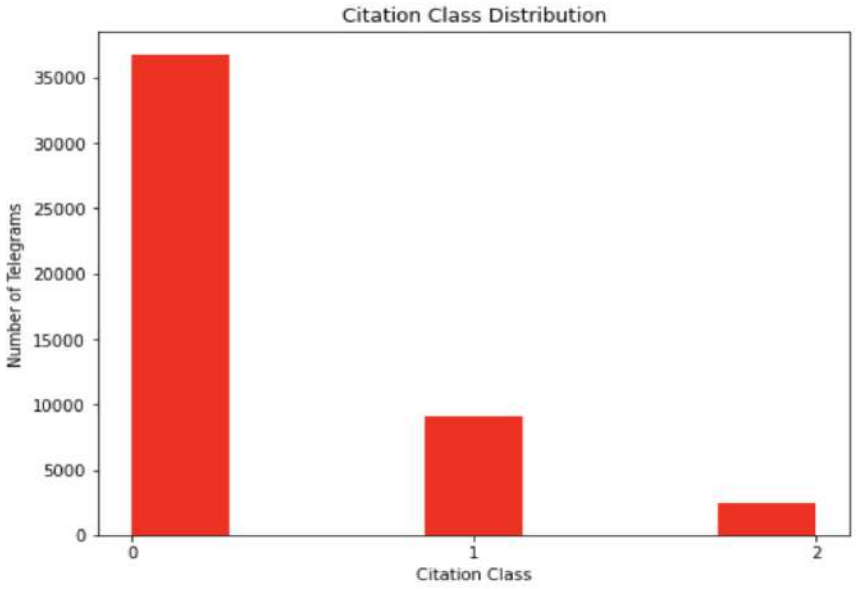

FIGURE 3.3: Citation Class Distribution

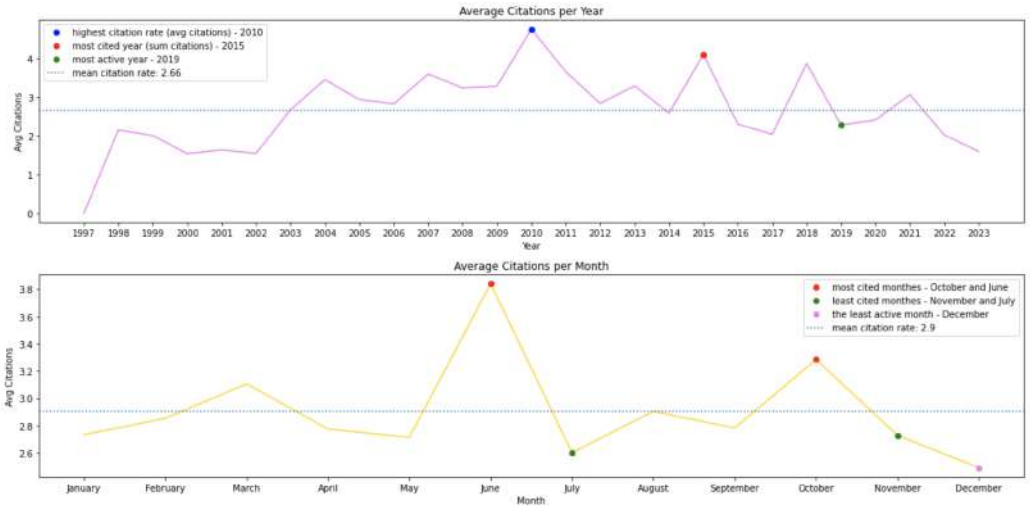

FIGURE 3.4: Temporal Distribution of the Citation Rate

located at the core of our Milky Way galaxy, becomes more prominently visible in the night sky. Oppositely, December poses challenges for astronomical observation as the Earth's position in its orbit obstructs the view of certain celestial objects. This, coupled with the reduced visibility of the Milky Way during that period, makes December the least "interesting" month in terms of astronomical observations. In relation to November, it can be observed that this particular month does not feature any significant astronomical events that are exclusive to it. Additionally, in many regions, November corresponds to the transition into colder weather, which often leads to cloudier skies and reduced visibility for observations.

Chapter 4

Astronomical Data Features Extraction

With the successful completion of the initial objective outlined in the **Introduction** goals, which involved the creation of a clean and tagged dataset, we now shift our focus to the next stages of our research. This includes text information extraction (IE) and feature engineering. These stages will enable us to extract relevant information from the text data and engineer informative features that will enhance our models' predictive capabilities.

4.1 Word2Vec

As some of the subsequent features that we extract from the text, named entities and topics in particular, require a vector form representation in order to be used as an ML model input, we begin this step by introducing a word2vec solution that will be further used for the stated above purposes. By leveraging word2vec, we can capture the semantic relationships between words and generate meaningful representations that capture the context and meaning of the text.

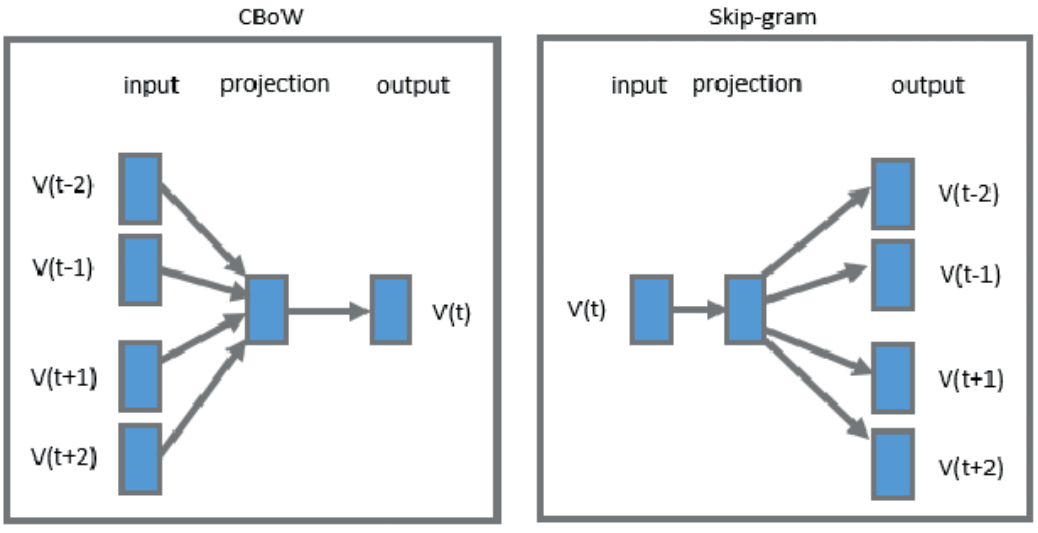

FIGURE 4.1: Word2Vec Representation Models: CBOW and Skip-gram. Image from the internet.

Due to the unique nature of astronomical data, we need a fast and scalable solution to train custom word2vec embeddings of different dimensions. Traditional

word2vec models might not capture the specific nuances and domain-specific terminology present in astronomical texts. For instance, the existing BERT word-level embeddings or Spacy embeddings taken from the *en_core_web_lg* model were not able to capture the similarity between the "FAST" and "MOST" (both are telescopes) or between "blazar" and "agn" (blazars belong to the group of the **a**ctive **g**alactic **n**ucleus - agn).

To create our custom word embeddings, we employ BLOOM-based embeddings and utilize the fine-tuning functionality offered by **Floret**. Floret stands out as a lighter and faster compressor compared to one of the widely used word2vec methods, FastText. We preprocessed the telegram bodies using the NLP best practices (special character removal, stop-words removal, etc.) and trained Floret embeddings of sizes 128 and 256 for further inference. CBOW and Skip-gram are two popular algorithms used for training word embeddings (Figure 4.1) in the context of word2vec. While both algorithms have their strengths and weaknesses, CBOW is usually preferred because of its computational efficiency and the tendency to work better when there's enough training data. Floret provides an option to choose between both of them.

In our training process, we utilized the CBOW model option with the default training settings, except for the adjustments made to the dimensionalities.

4.2 NER

To develop our Named Entity Recognition (NER) model, we focused on identifying and extracting specific entities of interest in the astronomical domain, namely telescopes and sources. Through communication with our astronomer colleagues, we were able to determine that these entities play a crucial role in identifying interesting observations. Worth noting, that by telescopes we are sometimes referring to the observatories, and sources may identify any observable cosmic object (for instance, we are not separating stars from planets). The theoretical formula of an interesting observation is pretty straightforward:

$$\text{powerful telescope} + \text{rare source} = \text{interesting observation}$$

In order to approach the solution, we selected the **Spacy** Python library as our tool of choice for its comprehensive documentation, computational efficiency, and overall convenience for the task at hand.

4.2.1 Training Attempt One

To have something to start with, our colleagues provided us with a defined list of telescopes and sources that they actively track in the telegrams as interesting entities. We transformed this list into a set of pre-defined regex rules, encompassing over 2600 entities. This set included 107 telescopes and approximately 2500 sources. These regex rules were then used to label the bodies of the telegrams in a format that corresponds to the required training format for Spacy NER models. Each labelled entity is represented as a tuple $(span\_start, span\_end, label)$, indicating the starting and ending positions of the entity within the text, along with its corresponding label - either TELESCOPE or SOURCE. Due to the model's generalizing and in-context-seeking ability, we expected the model to identify a larger set of entities compared to the rule-based (regex) approach that we were provided with. Moving forward in

the text, we will refer to the model inference phase when applying the Named Entity Recognition (NER) model to the entire dataset, rather than a separate test subset.

We used the *en_core_web_sm* built-in Spacy model as a base model for training. The model was trained in 2 epochs, achieving an F-score of 99.79 on the validation set (taken as 15% of the total data). During the inference, we discovered that the model was able to find 40 new unique sources and 3 new telescopes that were not previously included in the labels, while it was almost never missing any of the provided rule-based labels. Despite the promising training metrics and newly found entities, we have also noticed some issues after checking the inference results:

- Since we had a relatively small number of interesting entities used for labelling, approximately 9.73% of the annotated telegrams did not have any entity labels specified. Although the model was able to capture slightly more of the relevant information, there were still 9.68% of the telegrams without any entities identified by the model. There also were only 1.02 entities per telegram on average, which indicates that a significant number of telegrams were missing either the source or the telescope information, as identified by our NER model.

- The only thing that could neglect the previous problem could be a strong correlation between the amount of found entities and the citation rate of the telegram. For instance, if most of the telegrams with the missing entities would correspond to the not-interesting citation class, it could already be a powerful signal alone. However, despite we only trained the model on a subset of telescopes and sources stated as interesting ones, there was no correlation between the number of entities identified and the citation rate of the telegrams (Figure 4.2). This implies that telegrams without any identified entities could be equally relevant to all citation classes.

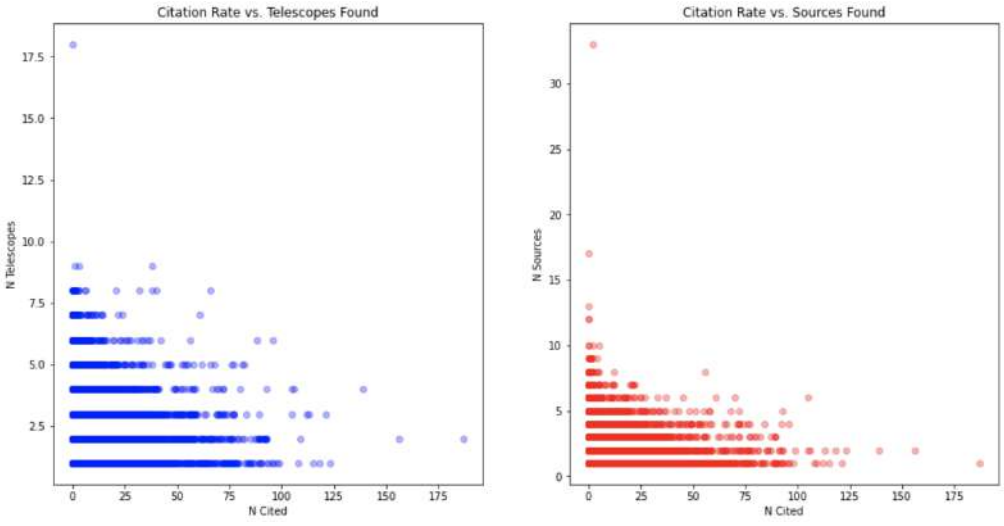

FIGURE 4.2: Correlation between the amount of Extracted Entities and Citation Rate

We decided to iterate on this step and enhance the training data in order to increase the extraction coverage and achieve better entity representation for each telegram.

4.2.2 Training Attempt Two

An additional source of labels was required to create more annotations and increase the NER coverage on inference. "Those who seek shall find" - Matthew 7:7. We found out that Wikipedia provides extensive lists of telescopes categorized by their types, including optical, gamma-ray, radio telescopes, space telescopes, and more. It also has lists of astronomical sources such as stars, planets, and galaxies. Hence, the Python parsing script was written and by leveraging that Wikipedia information we achieved to augment our training data with a wide range of entities, adding 717 new telescopes and 3050 new sources to our training labels. With the doubled annotations the model was training noticeably longer achieving the maximum F1-score of 98.9.

The results of the model inference were significantly more promising this time. Only 3% of the telegrams did not have any extracted entities, resulting in an average of 3.08 unique entities found per telegram. This indicates a three-fold improvement in model coverage. Consequently, the correlation depicted in Figure 4.2 has become insignificant. An example of the NER results can be observed in Figure 4.3. More NER examples in a text format can be seen in Appendix A.

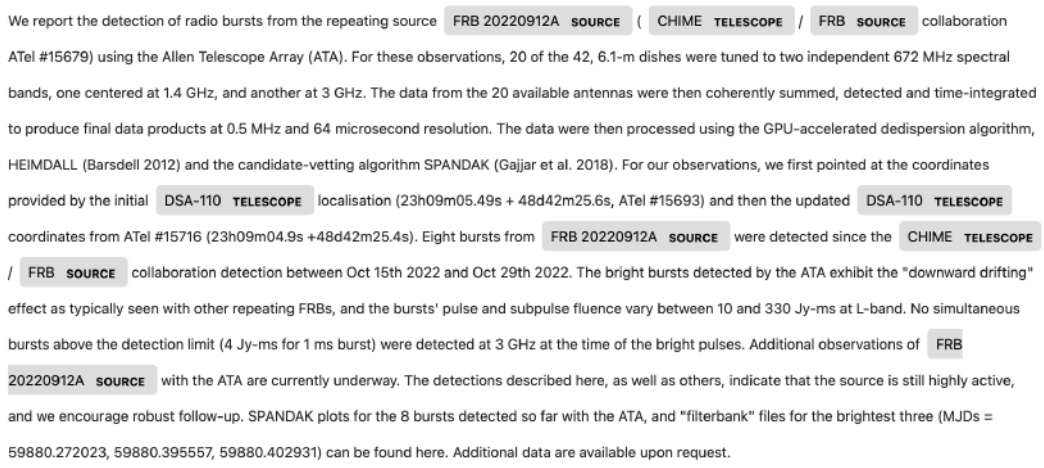

FIGURE 4.3: NER Predictions with Spacy

Notably, as shown in Table 4.1 the new NER model was able to identify more than a hundred previously undefined telescopes while missing even slightly more telescopes that were labeled with the regex patterns. A similar pattern emerged for the sources but with even more unique entities undefined by NER this time. Considering these facts, we made a decision to combine the NER predictions and the regex patterns in order to achieve the maximum entity coverage and extract as many unique telescopes and sources as possible. Later in the text, when discussing the extracted entities from the telegrams and the corresponding NER results, we will be referring to the combined outcome of both the NER model and the rule-based approach.

Entity Type	Entities not found by NER	Entities not found by RB
TELESCOPE	101	131
SOURCE	947	307

TABLE 4.1: NER and RB comparison

Proceeding with the analysis of the extracted entities. One of the interesting insights can be drawn from Table 4.2. We have observed that we possess more knowledge about interesting telescopes compared to interesting sources. This indicates that we can extract information about telescopes that are capable of discovering rare phenomena in the cosmos, which in turn increases the likelihood of their observations being cited in the future. The number of such powerful telescopes is limited, and it is expected that we can identify most of them since they are frequently mentioned in the training data, particularly those operated by the largest observatories and institutes. On the other hand, there is a larger number of amateur telescopes (class 0) that are only capable of detecting generic objects, and we tend to miss them. However, the number of rare and interesting sources is much larger, making it challenging for us to capture all of them. The training data primarily consists of well-known source labels, such as popular stars and black holes, thereby limiting our coverage of the diverse range of rare sources.

Citation Class	**Telegrams with missing Telescopes**	**Telegrams with missing Sources**
0	19%	11.7%
1	13.9%	11.9%
2	11.2%	19.7%

TABLE 4.2: Missing Entities per Citation Class

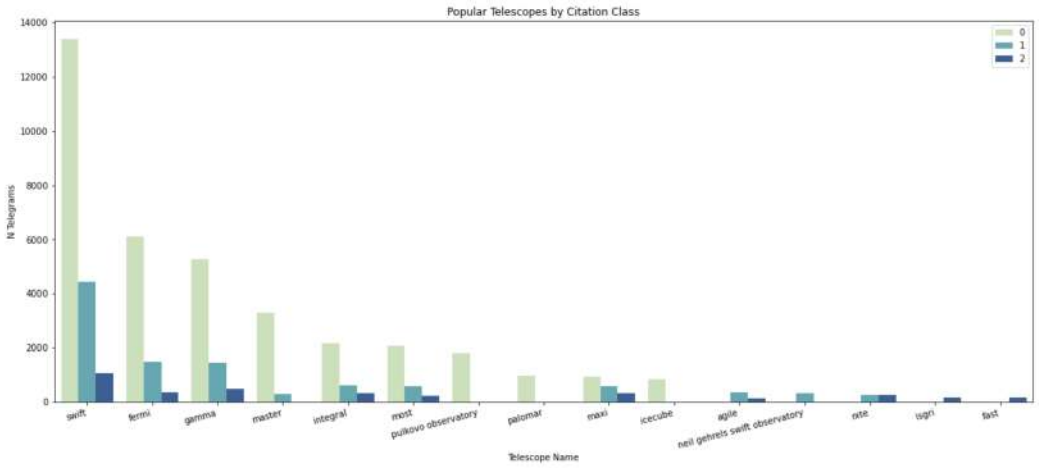

FIGURE 4.4: Telescopes by the Citation Class

A step towards the upcoming prediction problem would be to see the extracted entities' distribution with respect to the citation classes. Figure 4.4 displays the distribution of the 10 most frequent telescopes among different citation classes. The analysis of this figure reveals that the presence of certain telescopes in astronomical telegrams may be related to their citation rates. For instance, telegrams mentioning the MASTER telescopes or Palomar (observatory) are generally not cited, with most of them belonging to class 0. In contrast, reports mentioning the observations obtained by AGILE (satellite) or FAST (**F**ive-hundred-meter **A**perture **S**pherical radio **T**elescope) are mainly assigned to the highly-cited telegrams class. Figure 4.5 illustrates a similar pattern for the top 10 most frequent sources. Sources like the sun (obviously), stars, or radio sources are likely to be generic and commonly observed, which explains their high frequency. In contrast, black holes or neutron stars, being

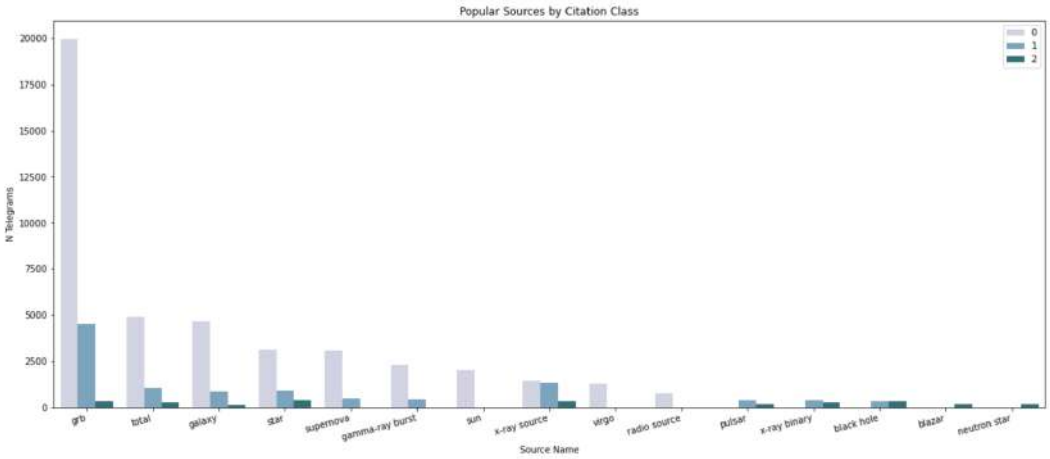

FIGURE 4.5: Sources by the Citation Class

more unique and rare in nature, are typically mentioned in the most cited telegrams, highlighting their significance in astronomical observations. Indeed, in the case of both telescopes and sources, there are entities that are mentioned equally across all citation classes, reflecting a similar distribution as the citation classes themselves.

These findings suggest that a traditional machine learning algorithm, such as a Random Forest Classifier (RFC) or LGMB, has the potential to differentiate the relevance of telegrams by solely considering the extracted entities. The methodology for predicting citations based on the extracted entities and their corresponding vector representations will be detailed in the **Citation Prediction** section.

4.3 Doc2Vec

The next approach we explored involved capturing the context of the entire telegram body to predict its future informational value and analyze the patterns or their absence in the context-citation relationship. As an appropriate technique, Doc2Vec was chosen. This technique for generating vector representations of documents is an extension of Word2Vec and is designed to capture the semantic meaning of an entire document, as opposed to just individual words. This vector can then be used as an input to machine learning algorithms of a choice.

Doc2Vec models learn from the context and patterns of the input text, and these patterns can be easily corrupted by the presence of noise or irrelevant information. In order to avoid this kind of situation, the telegram bodies were preprocessed and normalised following standard practices such as stop-words removal, punctuation and special characters removal, and lowercase casting.

The Doc2Vec algorithm of our choice is called **Doc2VecC**, described in “Efficient vector representation for documents through corruption" Chen, 2017. It represents each document as a simple average of word embeddings, capturing the document’s semantic meaning during the learning process. Doc2VecC includes a corruption model that favours informative or rare words while forcing common and non-discriminative words to be close to zero. Due to its proven efficiency and reliability, we employed this algorithm in our approach.

The implementation of Doc2VecC used in this project was obtained from Stanford University’s GitHub repository (https://github.com/mchen24/iclr2017). This implementation provides a range of settings and options for representation learning

architectures, including the CBOW model and Skip-gram (Figure 4.1). We opted for the CBOW model with most of the default settings but with a slight modification to the CBOW window parameter. Instead of the default window size of 10, we increased it to 15 to capture a wider range of surrounding context for each training iteration. The document embeddings' dimensionality was set to 256.

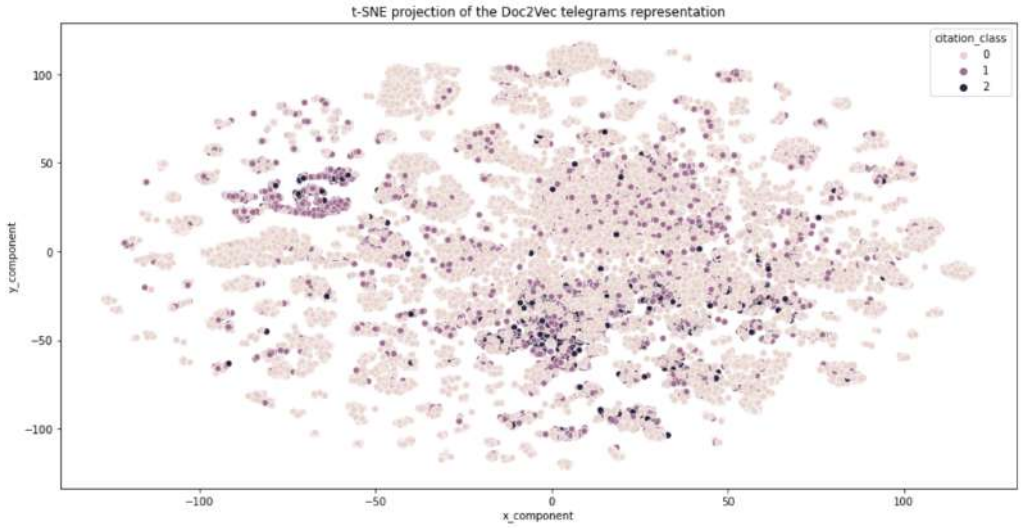

FIGURE 4.6: t-SNE projection of Document Vectors

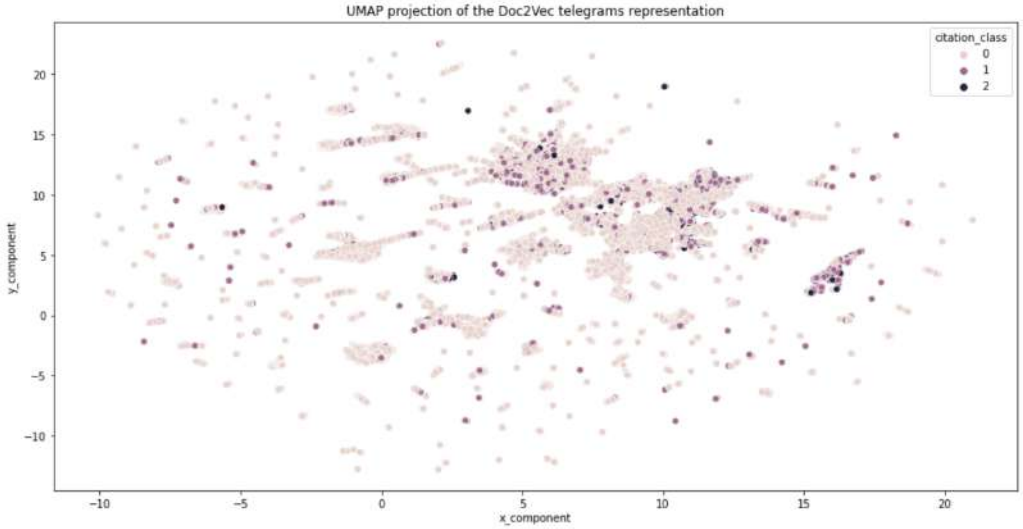

FIGURE 4.7: UMAP projection of Document Vectors

Although our initial intention was to use doc2vec embeddings for the subsequent citation prediction task, we discovered some interesting insights by examining the 2D projections of the resulting vectors in relation to the target value. Of course, in the ideal case scenario, we expected that the telegrams with high citation rates tend to cluster together and be visually separatable from the low citation rate clusters. The t-SNE projection in Figure 4.6 and the UMAP projection in Figure 4.7 reveal that there is no clear separability between the citation class clusters. Both visualizations indicate a lack of distinct and well-separated clusters corresponding to different citation classes. However, it is worth noting that despite the absence of clear separability, there is one noticeable cluster in both visualizations that mostly

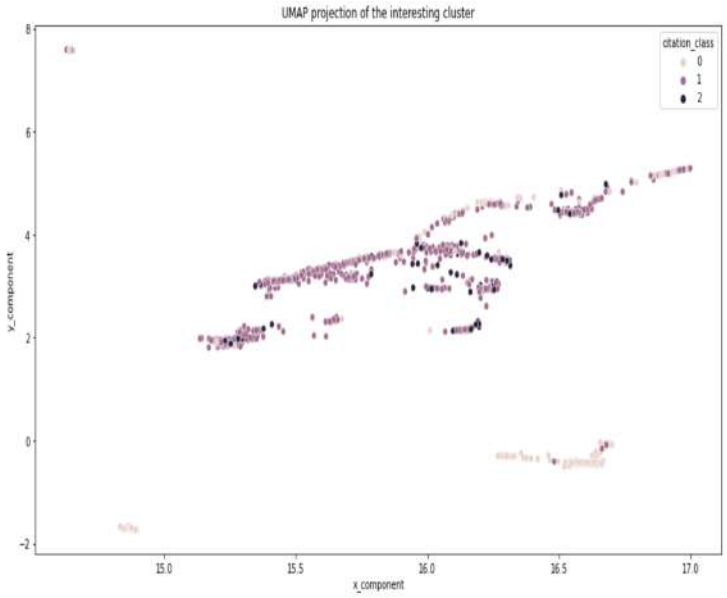

FIGURE 4.8: UMAP projection of the interesting DocVec cluster

contains telegrams labeled as "interesting" and a few "very interesting". This cluster is located on the right side of Figure 4.7 and the left side of Figure 4.6, and can be seen more closely in Figure 4.8.

Further analysis of the intriguing cluster reveals, that all of the telegrams in that cluster belong to GCN info source, 94% of them are related to the GRB (**G**amma-**r**ay **b**urst) observations mainly made with the Swift telescopes (96%). More high-level details can be seen in the WordCloud Figure 4.9. It is worth noting that the average citation rate for the telegrams in this cluster is 6.4, which is more than two times higher than the global average of 2.9. This indicates that telegrams within this cluster, which are predominantly focused on GRB observations with the Swift telescopes, tend to receive a higher level of attention and citations within the astronomical community. It is noteworthy that 71% of these posts belong to just four different authors, indicating that certain authors may consistently receive more citations and have a greater impact compared to others.

These insights emphasize the importance of considering the topic information when predicting the citation rates of telegrams, as well as the importance of finding a way to represent the authors as the feature. Subsequent experiments involving the use of Doc2Vec embeddings, as well as their combination with other signals, will be detailed in the **Citation Prediction** chapter.

4.4 Topics Extraction

In this research, the topic extraction model is needed to identify and extract the underlying themes or subjects discussed in the astronomical telegrams. While we have already obtained the complete document vector representations in the previous section, we believe that extracting topics can be a great complement to the Feature Extraction pipeline and serve two main purposes:

1. Feature Engineering: Topics can serve as informative features in machine learning models for predicting citation rates or other relevant outcomes. Additionally, we can generate novel topic-based features, such as the citation rate of

FIGURE 4.9: Intriguing Cluster's WordCloud

related topics and the relative frequency of topics.

2. Interpretability: Topics provide a higher-level representation of the telegrams, allowing researchers to interpret and analyze the content in a more meaningful way. While their impact on the citation prediction problem may not surpass that of the entire document embeddings, topics offer a concise summary of the key themes and subjects discussed in the telegrams.

For this research, we aimed to explore various approaches for each step, and as a result, we decided to integrate a Large Language Model into our methodology. In line with the current trend, we selected the model offered by **OpenAI** for this purpose. Given the uniqueness of astronomical data and the specific requirements of our research, it was necessary to perform fine-tuning on the chosen OpenAI model. OpenAI provides a wide range of options for both direct inference and fine-tuning, allowing us to tailor the model to our specific needs and improve its performance in the context of astronomical data analysis. Following OpenAI's general recommendations (and the budget limited by the author's purse), we selected the **ADA** (**A**dversarial **D**ebiasing **A**ugmentation) model as the most suitable option for our research. ADA is known for its efficiency in fine-tuning and its cost-effectiveness during inference. For instance, ADA's fine-tuning cost per thousand tokens is $0.0004, with a usage price of $0.0016 per thousand tokens while a more powerful Davinci (GPT-3) model would cost $0.03 and $0.12 respectively, making it 18 times more expensive. These factors made the ADA model a practical choice for addressing topic extraction tasks.

4.4.1 Annotations Preparation

Like any other model, ADA also requires labeled annotations for fine-tuning. In order to fine-tune the model, we needed to provide it with prompts and their perfect completion. The question arises: "Where do we get enough astronomical data with tagged themes?". And once again we are lucky enough, as the ATel website already has the built-in functionality that tags each of the newly added telegrams with a

set of related topics. With a couple of additional parsing steps, we were able to effortlessly obtain approximately 15,000 telegrams tagged with 54 different topics.

When it comes to the annotation prompt format, there are several strict requirements that need to be followed to ensure compatibility with the fine-tuning process. The requirements for the annotation prompt format are generally independent of the specific model and framework used for fine-tuning. One common requirement is that each prompt should end with a specific set of characters, such as `\n\n###\n\n`. This ending pattern helps in identifying and separating the prompts from the rest of the text during the fine-tuning process. Another requirement is that all completions should have the same set of characters at the end, such as ###. This consistent ending allows for easy extraction and processing of the completed annotations. By following these requirements, there is no need for additional prompt engineering, such as explicitly instructing the model to extract the related topics. The annotations can be straightforwardly generated by providing the text without any explicit instructions, as long as the prompts and completions follow the specified format.

Following the recommended practices, we prepared 15000 annotations with telegram bodies in prompt and ATel-generated tags in completion.

4.4.2 Fine-tuning

The documentation provided by OpenAI is comprehensive and helpful in understanding the fine-tuning process. Once we created the fine-tuned model, we submitted it to the fine-tuning queue for processing. The cost for fine-tuning the model was set at the affordable price of $9, considering the significant amount of data involved. After approximately 8 hours, we were notified about the fine-tuning completion.

4.4.3 Inference and Analysis

After completing the fine-tuning process, we were able to utilize the model by sending completion requests. The total cost of performing inference on approximately 30,000 GCNs, following the iterative approach described below, amounted to approximately $40. To assess the model's performance, we first tested it on a subset of the ATel training data, which consisted of approximately 5000 telegrams. The completion process for each telegram took some time to generate the predicted topics. We then evaluated the model's performance by comparing the predicted topics with the true labels (ATel tags) and measuring precision, recall, and F1 score. The obtained metrics were as follows: precision of 63.1%, recall of 67.8%, and F1 score of 65.3%. Considering the affordable cost of the model and the domain-specific nature of the data, we considered these metrics to be acceptable.

To run the process of predicting topics for all of the GCNs, we implemented an asynchronous request-sending mechanism. Initially, sending synchronous requests for each telegram took around 5 seconds, which would have been too time-consuming for processing 33,000 telegrams. With the asynchronous implementation, we were able to predict the topics for all telegrams in less than an hour. Fortunately, OpenAI does not set any request limits for their ADA models, unlike more advanced models such as Davinci or GPT3.5-turbo. Moving on to the investigation of the results, we found that the model was successful in identifying numerous previously unlabeled topics. However, there was a downside to this as a significant portion of these newly identified topics were unrelated to astronomy. Some of the topics were also questionable, and it was necessary to consult a domain professional to validate their relevance. Out of the numerous topics extracted by the model, around 10000 in

total, a large portion of them appeared only once or twice and were not relevant to astronomy. Only 766 topics were predicted more than twice. While there were some promising new topics, we wanted to minimize the noise in the predictions. Therefore, we decided to focus only on the 54 topics that were present in the ATel tags used for labelling, calling them **benchmark topics**. This can be seen as a form of multi-label topic classification, where we narrowed down the topics extraction to a specific set of 54 relevant topics. To ensure that each telegram received at least two benchmark topics, we implemented an iterative process during the prediction phase. As generative models have a random nature, approximately 10% of the GCNs initially did not have any predicted topics that matched our new benchmark criteria.

After multiple iterations and data validation, we successfully predicted at least two benchmark topics for each telegram in the dataset. Note that we were not re-predicting the topics for ATels, as they were already there. With the extracted topics in hand, we are now ready to perform data analysis to gain insights and formulate theories based on the information seen.

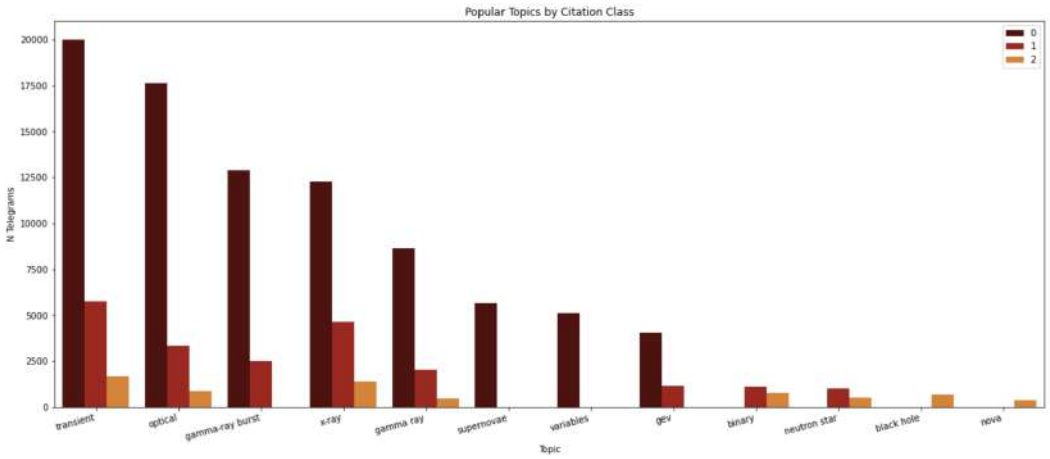

FIGURE 4.10: Popular Topics by Citation Class

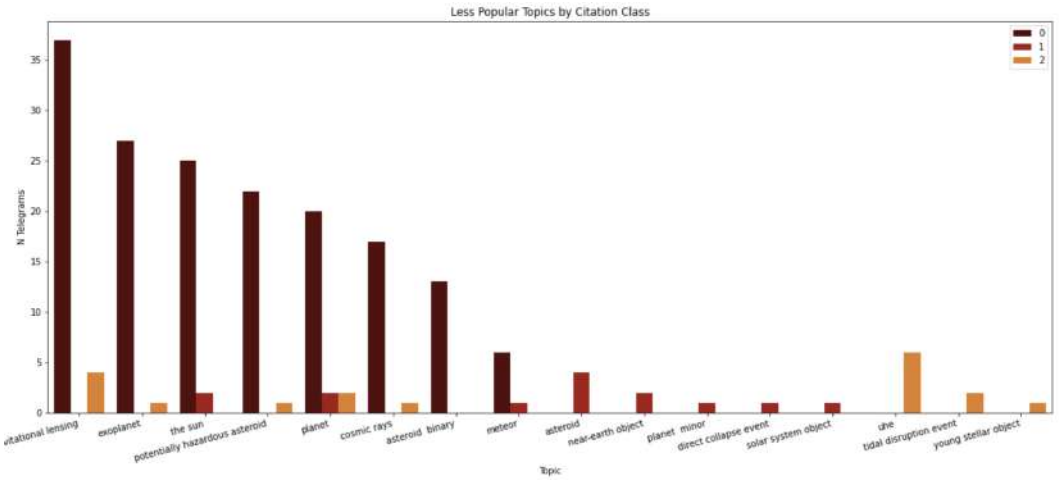

FIGURE 4.11: Less Popular Topics by Citation Class

In Figure 4.10, we observe the distribution of the most frequent topics in relation to the citation classes. The pattern of this distribution is similar to the distribution we observed for NER sources with respect to the citation classes, although with some

minor differences. Upon closer examination, we can identify that certain topics such as supernovae, gamma-rays, and variables (variable stars) are usually associated with telegrams that receive fewer citations. These topics tend to be more common in telegrams that are less cited overall. On the other hand, topics like binaries (binary stars), neutron stars, and black holes are often associated with telegrams that receive a higher number of citations. These topics appear to be more prevalent in telegrams that are cited more frequently. Figure 4.11 presents the distribution of less frequent topics in relation to the citation classes. Ironically, we observe that certain topics, such as "potentially hazardous asteroid" and "meteor," tend to have a lower citation rate despite their possible importance. This observation challenges the human intuition that topics related to potentially hazardous events or celestial phenomena like meteors would naturally receive more citations.

In Figure 4.12, the plot demonstrates the temporal aspect of topics citation, revealing shifts and trends in astronomers' interests over time. Although we don't possess enough knowledge to make definitive conclusions about the topics shown, this observation itself suggests that incorporating temporal information into future topics-based features such as the recent citation rate of topics relative to the global rate, could help to capture the evolving nature of community interests.

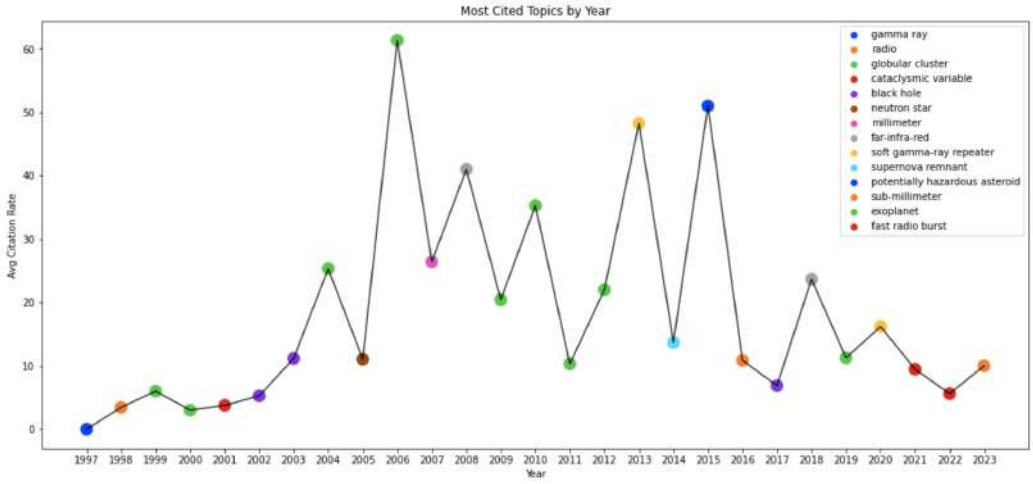

FIGURE 4.12: Most Cited Topics by Year

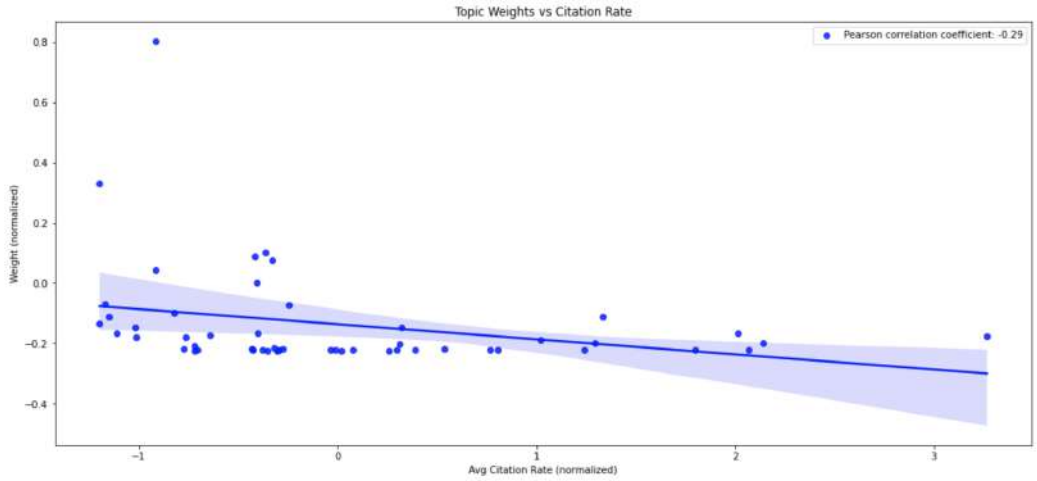

FIGURE 4.13: Topics Relative Frequency vs Citation Rate

Finally, in order to gain a better understanding of the relationship between topic frequency and telegram citation rate, we introduced topic weights calculated as follows:

$$\text{Topic Weight} = \frac{1}{\text{Num of Obseravtions with this Topic}} * \frac{\text{Total Num of Observations}}{\text{Num of Unique Topics}}$$

By incorporating these topic weights, we can visualize a scatterplot (Figure 4.13) to examine a correlation (or better say, the absence of it) between the rarity of a topic and the average citation rate of telegrams related to that topic. As no correlation can be observed between the rarity of a topic and the average citation rate of telegrams, we can conclude that the relative rarity of a topic alone does not guarantee a high citation rate. This suggests that other factors and signals are likely at play in determining the citation rates of telegrams.

With all the information discussed so far, we will move to the final section of this chapter, and leave the exploration of using topics for predicting citations for the future paper chapter.

4.5 Feature Engineering

In the previous sections, we discussed our approaches for extracting and representing text information from telegrams. We encountered various challenges along the way and gained valuable insights from our analysis. We employed several machine-learning techniques to represent the telegrams' content and underlying information. Moving forward, we will now focus on engineering the features that capture additional information beyond the direct context of the telegrams. These features, commonly referred to as meta-features, provide complementary insights that can enhance our prediction task.

4.5.1 General Features

We will begin with a couple of generic high-level features, that still can have their use cases:

- $refs\_count$ - represents the count of references and citations to other posts included in the telegram body. A higher value suggests that the post is more comprehensive and supported by a significant amount of evidence from external sources. It is also possible that some authors may not ignore or overlook someone who has referenced them previously, hence the telegram with more citations more likely will be highly cited.

- *telegram_len* - the char length of the telegram body. The longer length might indicate more substantial observations. On the other hand, longer telegrams can be more challenging to read and comprehend.

- *month* - the number of the month when observations took place. Regarding the data analysis we made before, some months (June, October) get more citations than others (December, July) due to the changing weather conditions or the Earth's positioning in space.

Continuing with the features that emphasize the time-series nature of the telegrams and can help to track the short- and long-term trends in astronomers' activity.

These features are computed based on the information available at the time of new telegram publication, without taking into account any future information. In other words, the feature values are determined solely based on the data that was already available up to that point in time.

Note: In the future context, when we refer to "month," we mean the period of the past 30 days leading up to the current moment. Similarly, when we mention "year," we are referring to the period of the past 365 days leading up to the current moment.

4.5.2 General Trend Features

- *month_citation_rate_to_global_ratio* - the average number of citations this month divided by the global average number of citations. Reflects the relative interest in recent posts. A higher value of this metric (above 1) suggests an increased level of interest in the posts made during the current month compared to the overall average citation rate. It might indicate the general increase in interest in the ATEL resource or a month that is more suitable for all sorts of observations.
- *year_citation_rate_to_global_ratio* - calculated as the average number of citations during the current year period divided by the global average number of citations. A higher value of this metric (above 1) suggests a growing level of interest in posts made thru the current year.

4.5.3 Author-based features

- *author_all_time_citation_rate_to_global_ratio* - calculated as the average number of citations received by this author(s)' all previous publications divided by the global average number of citations. A value greater than 1 indicates that the author may be relatively more popular or well-known compared to other authors. For instance, we observed this metric to be noticeably higher for the authors with NASA credentials.
- *author_year_citation_rate_to_global_ratio* - calculated as the average number of citations received by the author's posts during the current year and dividing it by the global average number of citations during the same period. This might indicate the growing interest in the author's posts during the previous year. For example, it could reflect a situation where a well-known observatory has acquired a new, more powerful telescope, leading to greater potential for significant discoveries.
- *author_month_citation_rate_to_global_ratio* - calculated as the average number of citations received by the author's posts during the current month and dividing it by the global average number of citations during the same period. Probably will help to catch the short-term trends in the author's popularity (if any).
- *author_activity_frac_year*- the ratio of the number of posts by the author during the current year to the total number of posts by this author. A lower value (minimum of 0) indicates that this is the author's first post in the current year, suggesting limited activity. A higher value (maximum of 1) indicates that all of the author's activity occurred in the current year, suggesting that the author may be relatively new to the field or have a focused presence within a specific time frame.

- *author_activity_frac_month* - the ratio of the number of posts by the author during the current month to the total number of posts by this author. Same as the feature mentioned above but for the month time period.

4.5.4 Topics-based features

An important detail to note about the topics-based features is that they were calculated taking into account the varying number of topics extracted from the telegrams. To ensure comparability, the average citation rate for each topic was calculated relative to the number of topics present in the telegram. This means that the topic citation rates of telegrams with two identified topics and those with ten identified topics were placed on the same scale, allowing for fair comparisons across different telegram-topics compositions.

- *topics_all_time_citation_rate_to_global_ratio* - computed as the average number of citations received by all previous publications that share the same topics as the current one, divided by the global average number of citations. A value greater than 1 suggests that these specific topics, in combination, have a higher level of interest and citation compared to other topics. It indicates that publications within these topics tend to receive more attention and citations in general.

- *topics_year_citation_rate_to_global_ratio* - computed as the average number of citations received by previous publications posted this year that share the same topics as the current one, divided by this year's global average number of citations. A value greater than 1 suggests that these specific topics have a higher level of interest during the current year than the other topics. This feature can be particularly useful in cases such as the discovery of gravitational waves, where specific topics gain significant attention and engagement from the scientific community over a period of years.

- *topics_month_citation_rate_to_global_ratio* - computed as the average number of citations received by previous publications posted this month that share the same topics as the current one, divided by this month's global average number of citations. A value greater than 1 suggests that these specific topics have a higher level of interest during the current month than the other topics. May be related to the better visibility of the specific cosmos objects during this month.

- *topics_activity_frac_year* - the ratio of the number of posts with the same topics during the current year to the total number of posts with these topics. A lower value (minimum of 0) indicates that these topics are not being actively discussed this year. A higher value (maximum of 1) indicates that most of these topics' related posts occurred in the current year, suggesting that these topics might either be more common for a specific time frame or they are getting more popular recently.

- *topics_activity_frac_month* - same as the feature mentioned above but for the month time period.

- *topics_weight_coef_all_time* - the weight coefficient of the topics which takes into consideration their frequency during the whole time. For each topic, we calculate its relative importance (weight) as described in 4.4.3. By applying weights to the citation rates of the topics, we can highlight the less frequent

topics by normalizing their citation rates based on their respective weights. For example, topics such as "transient" or "optical" are commonly observed in many posts and therefore have weights of 0.1 and 0.2, while the topic "meteor" is rare and has a weight of 400. The overall weight coefficient is calculated by dividing the raw citations by the weighted citations. A higher value indicates the presence of rare topics. While the weights calculated for the features did not exhibit a direct correlation with the target values (as shown in Figure 4.13), it is still possible that their combination with other features could have an impact on the prediction task.

- *topics_weight_coef_year* - same as above, but computed in a year timeframe.
- *topics_weight_coef_month* - same as above, but computed in a month timeframe.

4.5.5 Features Analysis

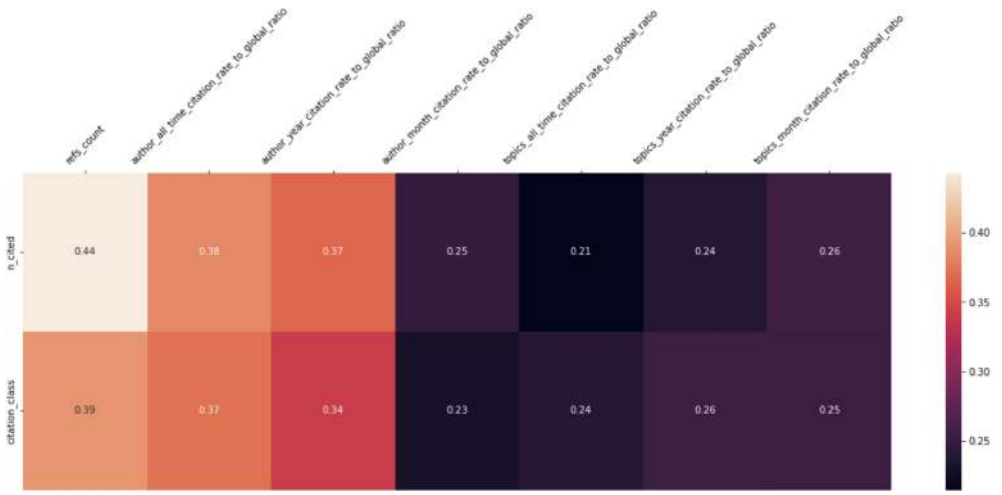

FIGURE 4.14: Features Correlation Heatmap

We investigated the correlation between features and the target variable by creating a heatmap, as depicted in Figure 4.14. To ensure that only features with at least a potential weak correlation were considered, we set a minimum threshold of 0.2. Analysis revealed that the *refs_count* feature has the strongest correlation of 0.44. This finding provides evidence to support the idea that when authors reference a greater number of telegrams in their own publications, it increases the likelihood of their telegrams being cited by others. In addition to that, we can also observe a moderate correlation between the citation rate and some of the author-based features. This finding provides further evidence that certain authors may have better reputations or access to more resources, such as powerful telescopes, which could contribute to their ability to make significant observations. And finally, the weakest out of the strongest, a few of the topics-based features also show an interesting behaviour identifying the predictable trend in a topics-citation relationship.

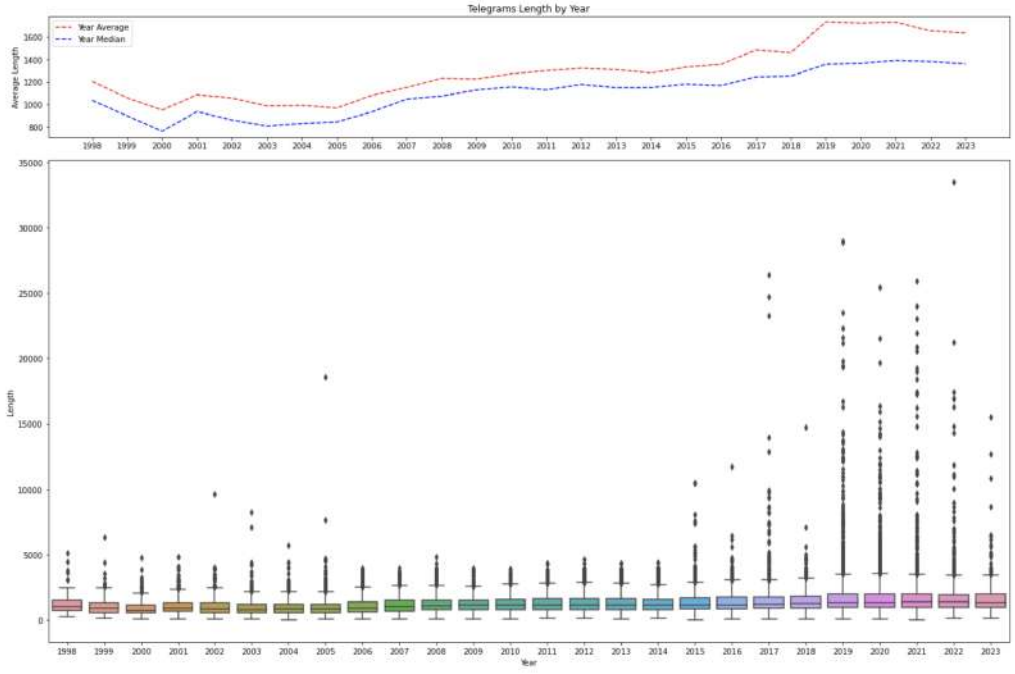

FIGURE 4.15: Telegram Length Boxplot

When examining the temporal boxplots of telegram length as shown in Figure 4.15, several untrivial observations come to light. One notable trend is the substantial increase in the average and median length of telegrams in recent years. Additionally, there is a notable increase in the number of telegrams that exceed 10000 characters, indicating a higher frequency of exceptionally long telegrams. These findings suggest a potential shift towards more extensive and detailed communication among astronomers over time. Or maybe the astronomers became using language models, such as Chat-GPT, to generate more refined and competitive observations. However, as any correlation between the telegram citation rate and length was not spotted, the notable influence of this feature on the citation prediction problem is doubtful.

Other observations that contribute to our understanding of the data can be drawn from the analysis of the authors' data. On average, authors generated 16.15 telegrams, with some authors producing hundreds of messages. The University of Leicester holds the record for the highest number of telegrams generated, with a maximum of 1757. Interestingly, the majority (76%) of productive authors, in terms of the number of telegrams, do not have the highest citation rates. Among the top 10 most productive authors, only two of them have a citation rate to the global average ratio (defined above as *author_all_time_citation_rate_to_global_ratio*) higher than the average ratio of 0.92. These authors are affiliated with the NASA credentials (milkyway.gsfc.nasa.gov) and David Palmer's group from the Los Alamos National Laboratory (palmer@lanl.gov), with citation rate to global average ratios of 1.96 and 2.40, respectively. Among the 100 most generative authors, 24 of them have an all-time citation rate above the average, while the entire group of the top 100 authors has an average citation value of 0.82, which is lower than the global average. This suggests that many large scientific astronomy groups, such as universities and observatories, often report routine observations rather than focusing solely on significant discoveries.

Surprisingly, a significant number of the most cited authors with a citation rate

to the global ratio of 10 and above (as investigated in the group of the top 100 most cited ones) have only a small number of posts, usually 2 or 3. These highly cited authors do not necessarily belong to specific scientific groups, and their credentials vary from universities to personal affiliations. This finding suggests that relying on author-based trend features may not be effective in identifying significant discoveries made by relatively new authors. And once again, it highlights the importance of focusing on the content of the telegram itself when determining the significance of the observations.

With all the information we have extracted and the features we have created in this chapter, we will now proceed to predict the citation rates of telegrams.

Chapter 5

Citation Prediction

5.1 Data Preparation

In order to utilize some of the extracted features, named entities and topics, in machine learning algorithms, we needed to convert them into a suitable vector form. To accomplish this, we employed multiple algorithms that are specifically designed for feature representation and transformation. In a previous chapter, we introduced the first approach called word2vec for generating word embeddings. We employed Floret and trained two versions of word embeddings, one with a dimensionality of 128 and another with a dimensionality of 256. These embeddings are now ready to be tested and incorporated into our models. Additionally, we utilized an algorithm known as Bag-of-Words as an alternative approach. This algorithm allows us to compare its performance with word embeddings in terms of computational efficiency and the range of achievable metrics.

When encoding named entities using the Floret embeddings, we followed a specific approach. First, we encoded the telescopes and sources separately, taking into account any duplicates and averaging the vectors if there were multiple instances of the entity (e.g. multiple telescopes found in the telegram). Next, we created a final feature vector by stacking the encoded telescope and source vectors together. The size of this feature vector was determined by the dimensionality of the Floret embeddings, which was either 256 or 512 depending on the chosen embedding size (128 or 256). In cases where there were missing values for telescopes or sources, we handled them by replacing the absent values with specific placeholders. For missing telescopes, the placeholder used was "no telescope found," and for missing sources, the placeholder used was "no source found." This ensured that all missing telescopes had the same vector representation and all missing sources had the same vector representation, allowing for consistent handling of missing data in our feature encoding process.

For encoding the topics vector using the Floret embeddings, we followed a similar approach as with the named entities. Each topic was encoded individually using the Floret embeddings and their encoded vectors were averaged to obtain a representative vector.

To incorporate the Bag-of-Words approach, we transformed our entities (both named entities and topics) into a single-string representation. For example, telescope entities such as ["MOST", "Adolphson Observatory", "Atlas"] were transformed into the string "MOST Adolphson Observatory Atlas". With the Bag-of-Words approach, the transformed entity strings were further processed to create a vector representation. Each column of the resulting vector corresponded to a unique word in the dataset. Indeed, in the case of the named entities, the Bag-of-Word approach was applied to telescopes and sources separately, so that these entities and sources had their

own dictionaries of unique words. The corresponding BOW vectors were stacked after that.

Entity type	Method	Vector Size
Named Entities	Floret 256	512
Named Entities	Floret 128	256
Named Entities	BOW	2112
Topics	Floret 128	128
Topics	BOW	74

TABLE 5.1: Created Vector Representations of the Extracted Entities

The table 5.1 shows all of the created representations of both extracted named entities and extracted topics.

After brief experiments, we evaluated the performance and computational efficiency of the created representations using two main criteria: the average fold fit time using cross-validation and the learning capacity measured by the test score on an overfitted model with the same parameters. Based on these evaluations, we determined that the Floret 128 representation provided the best trade-off between computational efficiency and performance for both named entity recognition (NER) predictions and topics extraction. Therefore, the final vector size for NER predictions is 256, while for topics extraction it is 128.

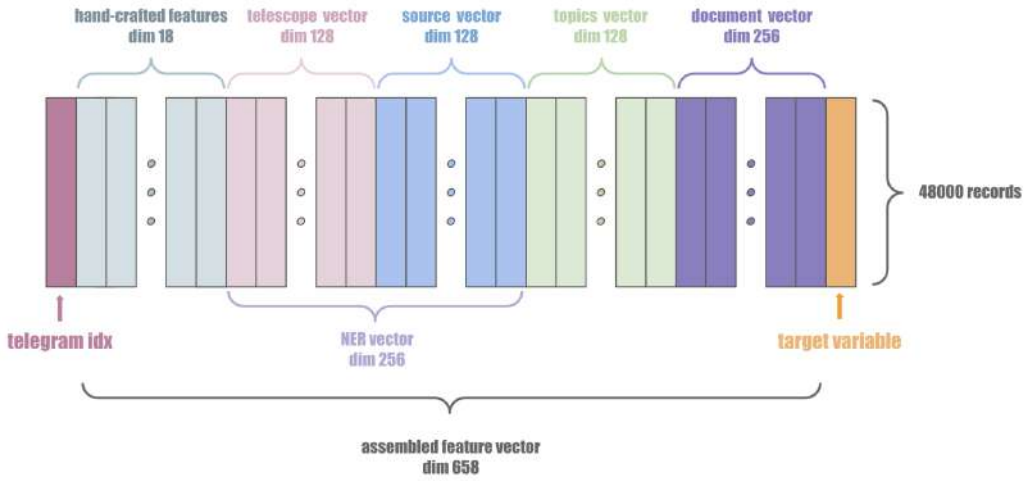

FIGURE 5.1: Assembled Data Schema

As the other features, namely document vectors and hand-crafted features are already in a form suitable for the prediction task, you can see a scheme of the final (later in text referenced as **assembled**) version of the training data in Figure 5.1. This scheme represents the dataset that is a culmination of the feature extraction process and is now ready to be utilized for citation prediction.

5.2 Experiments Setup

Before we dive into the experiments and their results, it is important to establish a set of clear rules that were followed throughout these experiments and will help to understand the results obtained.

1. In addition to our primary goal of achieving separability between interesting and not interesting telegrams for the classification problem, we will also explore the regression problem of predicting the exact number of citations. This approach will provide insights into our capabilities in this direction and can potentially be integrated into a stacking or blending solution to enhance the overall predictive power.

2. The prediction results obtained on a subset of the data will be referred to correspondingly. For example, we will have "docvec-based" results, "ner-based" results, "features-based" results, etc. On the other hand, the results obtained on the entire assembled dataset (as depicted in Figure 5.1) will be referred to as "assembled" results.

3. The results of the classification models will be referred to as "class" results, while the results of the regression models will be referred to as "reg" results. Indeed, the regression target variable is the future exact amount of citations.

4. The experiments conducted for binary classification and three-class classification tasks will be referred to as "binary" and "multiclass" experiments, respectively.

5. The random seed was fixed and remained the same on each preprocessing and training step involved.

6. Regarding the classification experiments:

   (a) To address the issue of imbalanced class label distribution, stratification was applied during the Train/Test/Val split and K-Fold split to ensure a representative distribution of classes in each subset.

   (b) Multiple imbalanced classification techniques were employed and compared on baseline models. These techniques included resampling, class weighting, and specific parameter adjustments. The best-performing technique was selected and applied during the final tuning step in each subsequent model iteration. This additional step will be indicated in the experiments table in the corresponding column.

   (c) To account for the imbalanced nature of the data, the chosen classification metric is the **balanced accuracy** (also known as weighted accuracy), which considers the class distribution and gives more weight to the minority classes. However, for the sake of brevity in the paper, it will be referred to as simple accuracy.

7. Regarding the regression experiments:

   (a) In order to handle extremely cited telegrams and mitigate the potential noise they introduce, we occasionally employ a technique where we limit the upper 2.5% percentile of target values. This means that any citations above a threshold of 20 citations are capped at that threshold. If this approach leads to an improvement in the test metrics, it will be indicated in the experiments table.

   (b) The metric of our choice for this regression task is Mean Absolute Error (**MAE**). This metric is the most intuitive to interpret, as it directly shows in how many future citations we are mistaking on average.

8. We consistently removed the first 1000 observations, which corresponded to the years 1997 and 1998. During this period, the resources were not widely used, and meaningful trends could not be identified. Similarly, we excluded the most recent 100 observations, as these telegrams might not have received their citations yet, which could potentially mislead the model.

5.3 Experiments

For each of the subsets of data and the prediction problems mentioned (binary and multi-class classification, and regression), we employed several machine learning algorithms, including Random Forest, LGBM, and neural network (NN) models based on Keras. In the experiment results, we only report the performance of the best-performing tuned models for each of these experiments. The best results will be highlighted with bold text, the different prediction tasks will are grouped and separated by the double line.

Based on the results presented in Table 5.2, the assembled dataset performed the best among the different subsets of data. Furthermore, the examination of feature importances using the best performing multi-class setup (Figure 5.2) revealed that the document vector representation was the most important feature. This suggests that the information captured in the document vectors played a significant role in the prediction task. On the other hand, the topics vector was found to be the least important among the extracted feature vectors, indicating that the topics information was already partially captured in the document vectors.

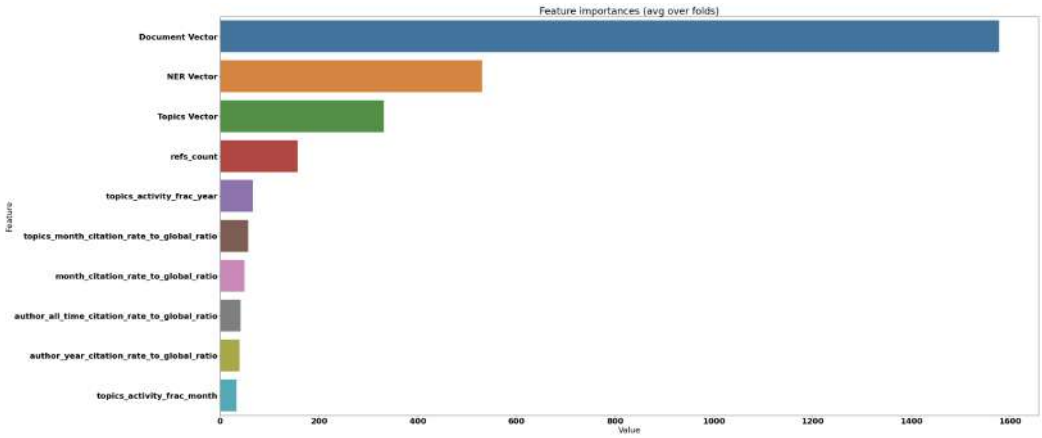

FIGURE 5.2: Feature Importances

Regarding the hand-crafted features, the reference count was identified as the most impactful feature. This finding supports the idea that the more telegrams an author references, the more likely they are to be referenced by others. Additionally, some of the author-based and topics-based features showed importance, which aligns with the observation of their weak correlation with the target variable and the idea of the presence of specific trends in the astronomers community.

5.4 Final Shot

Our goal was to maximize the separability between the interesting and uninteresting telegrams, ultimately aiming to improve the existing binary classification metric.

Specifically, we aimed to surpass an accuracy threshold of 80%, which is regarded as an acceptable metric in this context. We decided to further involve the **Model Stacking** technique (Figure 5.3). Model stacking is a technique used to enhance the predictions of machine learning models by combining their outputs and feeding them as input to another machine learning model known as a meta-learner. This approach leverages the diversity and strengths of multiple base models to improve overall prediction performance.

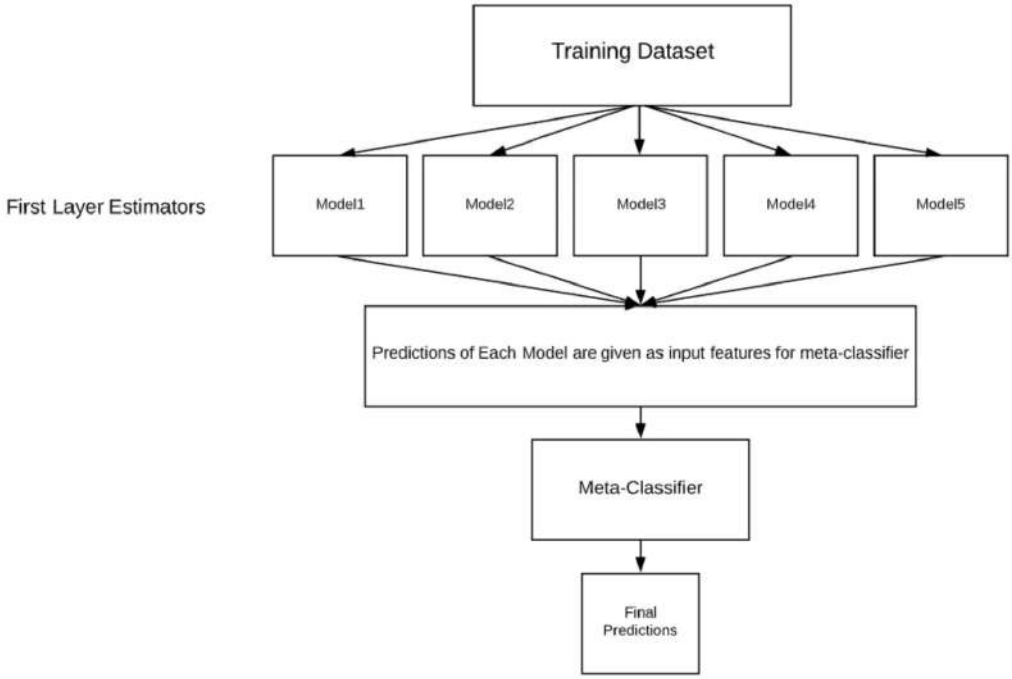

FIGURE 5.3: Model Stacking. Image from the internet.

To implement the model stacking approach, we have selected the following models:

- LGBM regression model
- Keras NN regression model
- LGBM multiclass model
- Keras NN multiclass model
- KKN multiclass model. As you could note, the KKN multiclass model was not listed in the previous experiments because it had much worse metrics compared to the other models. However, since model stacking considers different prediction principles and aims to find interesting patterns in the predictions, the KKN model could still have an impact on the stacking process due to its distance-based algorithm. Further analysis of the meta-model's feature importances supports this assumption.

The training data used for model stacking was the assembled dataset. This dataset was split into a training set and a holdout set. The holdout set predictions from the base models are then used to further train the meta-model, following the

principles of model stacking (Figure 5.3). The test subset used for evaluation is the same as the one used in the previous experiments section.

For the model stacking approach, we have chosen the LGBM Classifier as the meta-model. The task is binary classification, and we trained and fine-tuned the classifier using the holdout set predictions from the base models. The final evaluation was performed on the test set. The stacking approach proved to be beneficial, as it helped us achieve an accuracy of **83.3%**, which is an improvement of more than 5% compared to the previous best binary classification result.

The final classification report, shown in Table 5.3, provides an overview of the model's performance. The achieved recall for class 1 suggests that we are able to identify 84% of the interesting telegrams, although this comes at the expense of precision for this class. On the other hand, the high precision for class 0 indicates that we are very accurate at identifying telegrams that are not interesting.

Overall, the performance of the model indicates that we have successfully extracted enough information from the telegrams to achieve a satisfactory level of separation between highly cited (interesting) and not interesting telegrams.

Task	Algorithm	Add. Techniques	Test Metric
DocVec-based binary class	LBGM Classifier	oversampling	75.7 accuracy
NER-based binary	LBGM Classifier	"balanced" class_weight	68.4 accuracy
Topics-based binary	LBGM Classifier	"balanced" class_weight	67.2 accuracy
Assembled binary	**LBGM Classifier**	**"balanced" class_weight**	**79.4 accuracy**
DocVec-based multi-class	LBGM Classifier	oversampling	65.3 accuracy
DocVec-based multi-class	Keras NN	adding custom class weights	63.5 accuracy
NER-based multiclass	LBGM Classifier	"balanced" class_weight	57.4 accuracy
Topics-based multi-class	LBGM Classifier	"balanced" class_weight	58.4 accuracy
Features-based multi-class	LBGM Classifier	oversampling	64.3 accuracy
Assembled multiclass	**LBGM Classifier**	**"balanced" class_weight**	**68.7 accuracy**
Assembled multiclass	Keras NN	adding custom class weights	66.7 accuracy
DocVec-based reg	LBGM Regressor	capping at upper 2.5%ile	2.87 MAE
DocVec-based reg	Keras	none	2.24 MAE
NER-based reg	LGBM Regressor	capping at upper 2.5%ile	2.82 MAE
Topics-based reg	LGBM Regressor	capping at upper 2.5%ile	2.88 MAE
Features-based reg	LBGM Regressor	capping at upper 2.5%ile	2.51 MAE
Assembled reg	LBGM Regressor	capping at upper 2.5%ile	2.13 MAE
Assembled reg	**NN Keras**	**none**	**2.0 MAE**

TABLE 5.2: Experiments

Class	**Precision**	**Recall**
0	94	82
1	60	84

TABLE 5.3: Final Binary Classification Report. Balanced Accuracy achieved: 83.3

Chapter 6

Conclusions and Future Work

6.1 Conclusions on the Results

Throughout this research, we significantly replenished the vocabulary of astronomical terms and achieved the following results as stated in the Introduction Goals:

1. During our research, we collected and prepared a custom dataset consisting of observations from two of the most popular astronomer message sources - ATel and GCN. This dataset comprises a total of 48,000 observations. The data collection process involved parsing the relevant information from these sources and ensuring the dataset's quality and data format consistency for further analysis and modelling. Additionally, we developed functionality that allows for easy maintenance and updating of the dataset. This functionality enables the population of the dataset with new observations from the chosen source by taking the last populated telegram number, the desired number of telegrams to fetch, and produces an updated dataset as output.

2. We incorporated multiple advanced machine learning algorithms to extract and represent the information from the data. This included the development of the following:

   (a) Combined Named Entity Recognition (NER) model: This model was built using a combination of a Spacy-based Neural Network and a set of regular expressions. It is capable of identifying and extracting information about more than a thousand unique telescopes and over five thousand different cosmic objects referenced as sources in the telegrams.

   (b) Topic Extraction model: This model based on the OpenAI cheapest ADA model was designed to extract topics from the telegrams and was capable of determining numerous topics. However, it is recommended to limit the number of topics to a benchmark of 54 for better performance. This model served as a valuable complement to GCNs analysis, as the GCN resource did not have built-in topic detection functionality or corresponding text fields.

   (c) Astronomy Text Representation Models: a Word2Vec Floret-based model that utilizes BLOOM embeddings and the CBOW Word2Vec algorithm, and a Doc2Vec model based on the Doc2VecC algorithm. Both of these text representation models were crucial in the citation prediction tasks, as they provided meaningful and context-aware representations of this domain-specific text data.

   (d) We have created a comprehensive set of engineered features to capture the temporal nature of the astronomy telegrams and their meta-context.

These features include various temporal features and contextual features that provide valuable information for analysis and prediction.

3. We conducted an extensive analysis of the insights gained from each of the abovementioned parts. This analysis helped us gain a better understanding of the data and identify areas for improvement in future.

4. Finally, we developed a machine learning stacking solution that achieved an acceptable balanced accuracy of 83.3%, allowing us to effectively separate interesting and not interesting telegrams. This indicates that the features and the extracted information we utilized were valuable and contributed to the success of the citation prediction task. The final precision and recall metrics of our model align with the needs of potential end-users. With a precision of 94%, our model can accurately identify telegrams that are not interesting, saving time and effort in manually reviewing them. Additionally, with a recall of 84% on interesting telegrams, our model is effective at highlighting potentially significant observations. These metrics demonstrate the practical utility of our model in assisting astronomers in quickly filtering and identifying relevant information from a large volume of telegrams.

All the notebooks and related code with the described experiments can be found in the following GitHub repository: https://github.com/tamara-is-home/astro-diploma-nlp

6.2 Future Work

Future work in this direction could involve the following aspects:

- Continuous data updating and model retraining: As new telegrams are generated over time, it is essential to continuously update the dataset and retrain the model to ensure its relevance and accuracy. Developing automated pipelines using different scheduling solutions (for example an Airflow) to fetch and integrate new observations into the dataset can help keep the model up to date. Also providing astronomers with automated batch predictions via email with the newly appeared telegrams that are likely to be worth their attention can be a valuable feature. It can even have a paid subscription.

- Incorporating more advanced NLP techniques: While we have employed techniques such as named entity recognition and topic extraction, there are other advanced NLP methods that could be explored. This includes using deep learning models like Transformer-based architectures (e.g., BERT) and more advanced LLMs (e.g., GPT-4). Although performing a thorough comparison of different methods for each part of the project could be a standalone paper, it would be valuable to further evaluate and compare the existing approaches in the field in order to leverage the quality of the extracted information.

- Enhancing model interpretability: Even though we have achieved acceptable accuracy and performance, it is crucial to understand the reasons behind the model's predictions. While we already have techniques in place to extract interesting entities such as the telescope and source names, as well as topics, it would be beneficial to further enhance the interpretability of the models by transforming existing features. For example, we can convert features such as the relative popularity of authors or the rarity of topics into a score representation (1 - 10) and utilize those scores during the predictions' interpretations.

- Telegram Interest Prediction: Our research focused on using the citation rate of telegrams as an indicator of their informational value. However, it is important to acknowledge that there probably are alternative approaches to predicting the significance of telegram observations. By leveraging multiple advanced ML techniques, it is possible to measure the relevance and impact of the observed phenomena or events mentioned in the telegrams in completely another way, and who knows, maybe those ways produce many potent results...

Overall, further research and development in automated telegram analysis and knowledge extraction can have a significant impact on the field of astronomy. By utilizing advanced techniques in natural language processing, machine learning, and data analysis, it is possible to change the way astronomers approach their research and accelerate scientific discoveries, bringing humanity closer to the great unknown.

Appendix A

NER Examples

Example 1: The 2-m **Liverpool Telescope** (*TELESCOPE*) robotically followed up GRB110402 (**SWIFT** (*TELESCOPE*) trigger 450545; Ukwatta et al. GCN 11857) 14.85 min after the **GRB** (*SOURCE*) trigger time. We identify a faint source detected within the XRT error circle in R, i′ and z′ band images. Fading is yet to be confirmed. This message may be cited.

Example 2: Comparison of R band images of the error box of **GRB 980425** (*SOURCE*) (Soffitta et al. 1998; IAUC 6884) taken at the ESO NTT telescope on April 28.37 UT (900s) and May 1.33 UT (900s) shows no variation 0.3 mag down to 22.8 mag at the location of the transient **BeppoSAX** (*TELESCOPE*) **NFI X-ray source** (*SOURCE*) 1SAXJ1935.3-5252 (Pian et al. 1998; GCN #61). This message is citeable.

Example 3: Monitoring of the time-variable **radio source** (*SOURCE*) (GCN 63, IAUC 6896, GCN 70) coincident with a **supernova** (*SOURCE*) candidate proposed by Galama et al. (GCN 60, IAUC 6895) has continued with the **Australia Telescope Compact Array** (*TELESCOPE*) at 20, 13, 6 and 3 cm. The **radio source** (*SOURCE*) may have reached a peak on May 7 1998 at 6 and 3 cm of 45 and 49 mJy, respectively.

Example 4: M. Williams (PSU) reports on behalf of the **Swift** (*TELESCOPE*) Team: **Swift** (*TELESCOPE*) resumed observations of **GRB 221009A** (*SOURCE*) on February 7 at 00:51 UTC after the end of **Sun** (*SOURCE*) constraint, 10 Ms after the **Fermi** (*TELESCOPE*) /GBM trigger (Veres et al., GCN Circ. 32636). Further observations are planned for this weekend.

Example 5: We observed the **BeppoSAX** (*TELESCOPE*) MECS error circle of **GRB** (*SOURCE*) /XRF 020427 (GCN 1386) at 8.7 GHz with the **Australia Telescope Compact Array** (*TELESCOPE*) (ATCA) for 6.2 hours centered on May 11.8 UT. We detect one **radio source** (*SOURCE*) within the MECS error circle with a flux density of 190+/-35 microJy. We find no counterparts to the three Chandra sources reported by D. Fox (GCN 1387), down to a 2-sigma limit of 70 microJy.

Example 6: Following the detection of **SGR** (*SOURCE*) -like behaviour from the anomalous x-ray **pulsar** (*SOURCE*) 1E 2259+586 by **RXTE** (*TELESCOPE*) (GCN#1432), we have observed the field with the auxiliary port camera at the 4.2-m **William Herschel Telescope** (*TELESCOPE*) at La Palma on June 20, 02:53-03:35 UT. We do not find optical emission within the Chandra 0.6" radius error circle (Hulleman et al. 2001, ApJ 563, L49) down to limiting magnitudes of R = 24.8 and I = 20.0.

Bibliography

Alkan Atilla Kaan, et al. (2022). *TDAC, the First Time-Domain Astrophysics Corpus: Analysis and First Experiments on Named Entity Recognition.*

Baron, Dalya. (2019). *Machine learning in astronomy: A practical overview.*

Benballa Miriam, Sebastien Collet and Romain Picot-Clemente. (2019). *Saagie at semeval-2019 task 5: From universal text embeddings and classical features to domain-specific text classification.*

Chen, Minmin. (2017). *Efficient vector representation for documents through corruption.*

Chiticariu Laura, Yunyao Li and Frederick Reiss. (2013). *Rule-based information extraction is dead! long live rule-based information extraction systems!.*

Dunn Alexander, et al (2022). *Structured information extraction from complex scientific text with fine-tuned large language models.*

Grezes Felix, et al. (2021). *Building astroBERT, a language model for astronomy & astrophysics.*

Jiang Xiaobo, Kun He and Yongru Chen. (2023). *Automatic information extraction in the AI chip domain using gated interactive attention and probability matrix encoding method." Expert Systems with Applications*.

Klinger, Roman. (2011). *Conditional Random Fields for Named Entity Recognition.*

Liu Tao, et al. (2005). *Domain-specific term extraction and its application in text classification.*

Murphy Tara, Tara McIntosh and James R. Curra (2006). *Named entity recognition for astronomy literature.*

Nguyen, Thien Huu. (2018). *Deep learning for information extraction.*

Peng, Fuchun and Andrew McCallum. (2006). *Information extraction from research papers using conditional random fields.*

Pereira Jayr, et al. (2023). *Visconde: Multi-document QA with GPT-3 and Neural Reranking.*

Settles, Burr (2004). *Biomedical named entity recognition using conditional random fields and rich feature sets*.

Wahba Yasmen, Nazim Madhavji and John Steinbacher. (2023). *Attention is Not Always What You Need: Towards Efficient Classification of Domain-Specific Text.*

Wei Xiang, et al. (2023). *Zero-shot information extraction via chatting with chatgpt.*

Wu Lang-Tao, et al. (2022). *Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web.*

Wu Xin, et al. (2020). *Task-oriented domain-specific meta-embedding for text classification.*

Xie Shaohui, et al. (2022). *Research on domain text classification method based on BERT.*