Project Draft

Scanning Reddit for identification of Science Topics



|  |  |
| --- | --- |
| Name: | Supervisor: |
| University ID: |  |
| Word Count: 3243 (Without References) |  |
| Date: |  |

Table of Contents

[Table of Figures 2](#_Toc68880288)

[Abstract 3](#_Toc68880289)

[Introduction 3](#_Toc68880290)

[Background 4](#_Toc68880291)

[Aim 4](#_Toc68880292)

[Research Questions 4](#_Toc68880293)

[Ethical Considerations 4](#_Toc68880294)

[Literature Review 5](#_Toc68880295)

[Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets 5](#_Toc68880296)

[GraphIE: A Graph-Based Framework for Information Extraction 6](#_Toc68880297)

[Information extraction from text messages using data mining techniques 9](#_Toc68880298)

[Project Timeline 10](#_Toc68880299)

[Methodology 10](#_Toc68880300)

[Project Design 10](#_Toc68880301)

[Experimentation 10](#_Toc68880302)

[Dataset 11](#_Toc68880303)

[Machine Learning 11](#_Toc68880304)

[Conclusion 11](#_Toc68880305)

[Project Evaluation 12](#_Toc68880306)

[References 12](#_Toc68880307)

# Table of Figures

[Figure 1 - Methodology for LR 5](#_Toc68880328)

[Figure 2 - Accuracy for Speech tagging Datasets 6](#_Toc68880329)

[Figure 3 - F1 Scores for NER 6](#_Toc68880330)

[Figure 4 - Architecture and Framework for GraphIE 7](#_Toc68880331)

[Figure 5 – Results 8](#_Toc68880332)

[Figure 6 - Statistics for EDUCATION and JOB datasets 8](#_Toc68880333)

[Figure 7 – Results 9](#_Toc68880334)

[Figure 8 - Project timeline with Gantt chart 10](#_Toc68880335)

# Abstract

Social media has become one of the most influential technologies of the 20th century. And with the continued exponential growth of social media, it will also remain the most influential technology in the 21st century. Social media is now gaining attention not only as an entertainment platform, but also a sharing platform. The sharing of anything, be it solutions to technical problems, fashion trends or the breaking news, social media is there with all.

This has created a motherlode of data just from social media itself. This amount of data makes it almost impossible to find something easily. This research aims to deal with this problem by replacing the human in the equation with an algorithm/ensemble.

This research project culminates Machine Learning with Natural Language Processing (NLP) to scan through posts on Reddit and tell apart the posts belonging to the computer science domain. The output of the research thus provides an experimental analysis of how Machine Learning can come handy in a scenario like this.

**Keywords:** Reddit, Machine Learning, Natural Language Processing, Ensemble

# Introduction

The extraction of information from web-based media content has as of late become a functioning exploration theme, following early tests that demonstrated this classification to be very trying for cutting edge calculations (Bontcheva and Derczynski, 2016). For finding arising information, there is today another and very amazing source: socially created content. Interpersonal organizations are gigantic (1.8 billion of clients) and new (their substance is created while occasions happen) (Brambilla et al., 2017).

The reason behind the beginning of information extraction from social media is precisely because it has started becoming a hub for diverse information and data that we would never have expected to find on it 5 years ago. This project will strive to identify posts on Reddit that belong in the sciences domain.

This document will talk about the accompanying subheadings ahead; the introduction heading will continue to describe the establishment of the research with the objectives this endeavor will achieve. The requests that this investigation will endeavor to answer to will in like manner be inspected close by the ethical concerns of the endeavor and its course of occasions. The methodology i.e., the approach will look at how the Literature Review (LR) will be performed and what the assignment design is. It will moreover discuss in detail the instruments and theories used for achieving the objectives this project has. The methodology will end with a project appraisal that will pass on the modules that the assignment includes and the solicitation wherein they ought to be executed.

Following the former segment will be the (LR) which will discuss in detail the revelations of researchers exploring in a similar territory. The LR will summarize the methodology, investigations and disclosures of every assessment paper completed for pack in this errand. The end will be the last fragment summarizing the examination followed by references, list and glossary.

## Background

Reddit has been a popular social media website from the early times of the social media craze and is still more popular than ever. A lot of great and informative content makes its way to Reddit first and then to other places like informative blogs and solution forums. Reddit’s layout of the website is a little unique and people not used to Reddit but used to other platforms like Facebook get a little lost during navigation.

Also, since Reddit has been here for a long time, there’s a lot of relevant data available on the platform for anything you search but it’s almost impossible to find what you exactly want even though it’s available due to the sheer amount of data. Many people treat Reddit as a forum and many pages on Reddit are truly forums where people from all over the world share solutions to problems for every domain. This project can help people make use of these forums much more efficiently.

## Aim

The project’s aim is to provide an experimental proof for the usefulness of automating the act of identifying posts related to computer science on Reddit using Machine Learning. For this, the following is performed:

* Literature Review of the work done by researchers in the same domain.
  + Summarizing the research paper by summarizing its methodology and results.
* Dataset
  + Dataset selection and finalization
  + Data pre-processing and data cleaning
  + Data visualization and statistical analysis
* Machine Learning
  + Algorithm Workflow/Pipeline Design
  + Algorithm Implementation
  + Algorithm Training
  + Algorithm Prediction
* Performance Metrics
  + Algorithm Performance Testing using Cross-Validation
  + Confusion Matrices

## Research Questions

The following questions are what behind the motivation for this project:

1. Can Machine learning be of any support in identifying computer science related posts from Reddit?

## Ethical Considerations

Morals is, unquestionably, perhaps the main subjects to arise in AI and man-made brainpower in the course of the most recent year. While the explanations behind this are unpredictable, it all things considered underlines that the region has arrived at mechanical development. Morals should be viewed as a significant reasonable thought for anybody utilizing and assembling AI frameworks.

[Amazon rejected its AI instrument for employing in October 2018](https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G) since it showed huge predisposition against female occupation candidates. Significantly more as of late, a month ago it arose that calculations worked to recognize disdain discourse online have [in-constructed predispositions against individuals of color](https://futurism.com/the-byte/google-hate-speech-ai-biased).

Albeit these might seem like edge cases, it's imperative that everybody in the business assumes liability. This isn't something we can surrender to guideline or different organizations individuals who can truly influence change are the designers and specialists on the ground.

## Literature Review

The value of social media was realized a few years ago. The research on this domain started then. From that time, there has been a lot of work done in this field, and to read and understand that work is called Literature Review (LR). A strong and properly performed LR can become the validation point for the integrity of the project. Due to its importance, a specific methodology has been assigned to perform this task efficiently, shown in *Figure 2*.

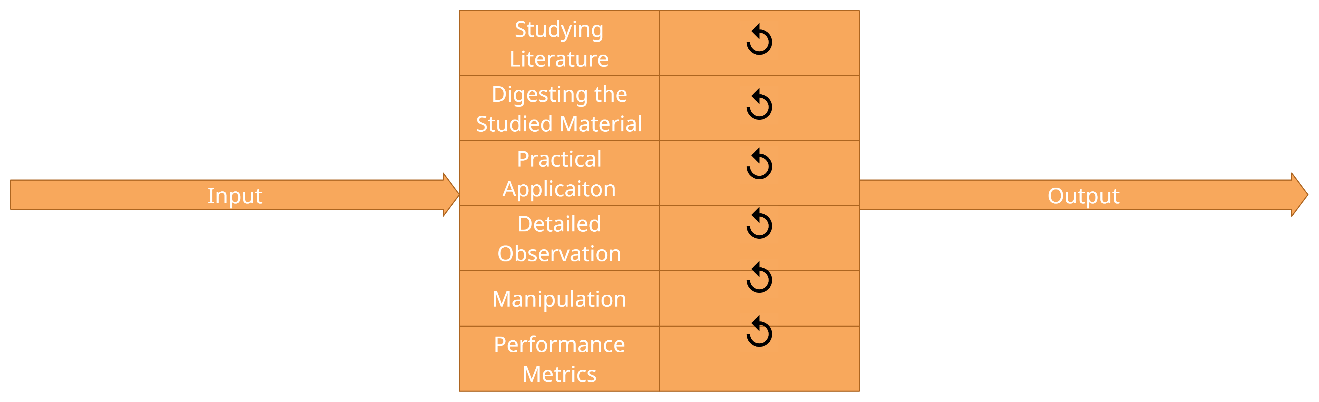


Figure 1 - Methodology for LR

As mentioned above, work is being done on this topic since a few years back. In this time, a lot of review material, namely research papers have accumulated for anyone that needs to do a Literature Review. But reviewing all the papers until now is not feasible. Therefore, a plan for selecting quality papers needs to in be place. The plan is as follows:

* English must be the only language in the papers.
* Must be distributed in journals with a high effect factor.
* Complete and free access should be accessible for the Paper and the Journal.
* All the examination datasets and code for the papers should be accessible free of charge.
* Papers sooner than 10 years are not satisfactory.

### Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets

Perform multiple tasks learning is compelling in decreasing the necessary information for learning an assignment, while guaranteeing serious exactness regarding single undertaking learning. We study adequacy of multi-dataset, perform multiple tasks learning in preparing neural models for four arrangement labeling undertakings for Twitter information, in particular, grammatical feature (POS) labeling, lumping, super sense labeling, and named substance acknowledgment (NER). We use – 7 POS, 10 NER, 1 Chunking, and 2 super sense – labeled openly accessible datasets.

#### Methodology

The creators previously consolidated various freely accessible information hotspots for POS, NER, CHUNK, and CCG undertakings. Every one of these datasets is involved English tweets clarified for arrangement labeling undertakings. For preparing our model each word in the arrangement is addressed by its ELMO portrayal. ELMO portrayal is extricated from a huge pre-prepared model dependent on English book. Three settings for looking at the models are thought of. Single information (S), here the model is prepared on just a solitary dataset of a solitary errand.

The other model is a multi-dataset single-task model (MD), which uses all the datasets from a solitary errand, with all datasets sharing a portrayal. At last, two kinds of multi-dataset perform various tasks models are used one in which all dataset and errands share a typical portrayal (MTS), and another where each undertaking has its own portrayal layer (MTL). The common portrayal layer in all models is a bidirectional long transient memory (biLSTM) layer. At long last, every one of the models have an exceptional restrictive irregular field (CRF) layer for each dataset or undertaking. The models are prepared utilizing the joint CRF misfortune for all errands and datasets. Every one of the secret units in the model are of size 100. The models are prepared utilizing the Adam enhancer.

#### Results

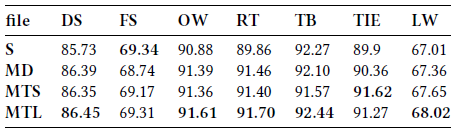


Figure 2 - Accuracy for Speech tagging Datasets

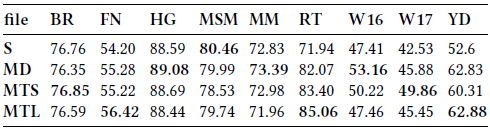


Figure 3 - F1 Scores for NER

### GraphIE: A Graph-Based Framework for Information Extraction

In this paper, GraphIE is presented. It is a system that works over a diagram addressing an expansive arrangement of conditions between text-based units (for example words or sentences). The calculation proliferates data between associated hubs through chart convolutions, producing a more extravagant portrayal that can be abused to improve word-level forecasts.

GraphIE improves predictions by automatically learning the interactions between local and non-local dependencies in the input space. This approach integrates a graph module with the encoder-decoder architecture for sequence tagging. At the core of this model, a recurrent neural network sequentially encodes local contextual representations and then the graph module iteratively propagates information between neighboring nodes using graph convolutions. The learned representations are finally projected back to a recurrent decoder to support tagging at the word level.

#### Methodology

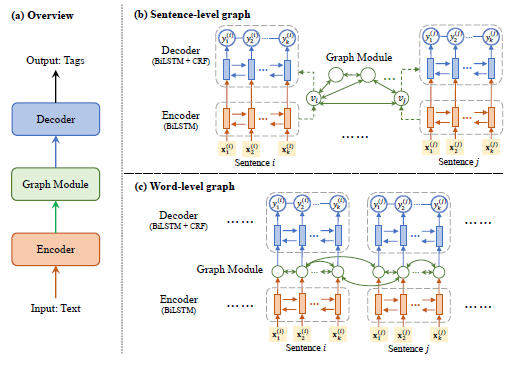


Figure 4 - Architecture and Framework for GraphIE

Three components in the GraphIE model:

1. The **encoder** generates local context-aware hidden representations for the textual unit with a Recurrent NN. The encoder RNN is defined by the following: where x denotes the input sequence, h denotes the hidden states and denotes the encoding parameters.
2. The **graph module** captures the graphical structure between textual units. The graph module is a Graph ConvNet (GCN). It consists of two parts:
   1. The first part gets the information from all nodes in the previous layer: where W is the weight for each input.
   2. The second part aggregates information from the neighbors of each node: where d is the degree of node (i.e. the number of edges connected to) and is used to normalize, ensuring that nodes with different degrees have representations of the same scale.
3. The **decoder** uses the data generated by the graph module to perform labelling at the word level. It consists of a BiLSTM + CRF tagger.
   1. BiLSTM:
   2. CRF:

#### Experiment

##### Textual IE

The creators led probes two NER datasets: the CoNLL-2003 dataset (CONLL03), and the CHEMDNER dataset for compound element extraction. We keep the standard split of every corpora. A word level diagram was made with two kinds of edges:

1. Nearby edges: forward and in reverse edges are made between adjoining words in each sentence, permitting neighborhood logical data to be used.
2. Non-nearby edges: re-events of a similar token other than stop words are associated, so data can be spread through, empowering worldwide consistency of labeling.

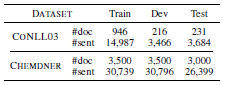


Figure 5 – Results

##### Social Media IE

Two datasets were built, EDUCATION and JOB, from the Twitter corpus. The first corpus contains a large number of tweets produced by ≈ 10 thousand clients, where the training and occupation specifies are clarified utilizing far off oversight. Tweets were inspected from every client, keeping up the proportion among positive and negative posts. The acquired EDUCATION dataset comprises of 443,476 tweets produced by 7,208 clients, and the JOB dataset contains 176,043 tweets created by 1,772 clients.

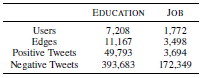


Figure 6 - Statistics for EDUCATION and JOB datasets

A diagram is developed as inner self organizations, i.e., when data around one client is extricated, a subgraph is shaped by the client and his/her immediate neighbors. Every hub compares to a Twitter client, who is addressed by the arrangement of posted tweets. Edges are characterized by the followed by interface, under the presumption that associated clients are bound to come from a similar college or organization.

#### Results

GraphIE beats the SeqIE benchmark in many ascribes, and accomplishes 1:2% improvement in the miniature normal F1 score. It affirms that the advantages of utilizing design chart structure in visual data extraction.

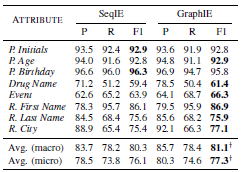


Figure 7 – Results

### Information extraction from text messages using data mining techniques

The point of this paper is to gauge the self-destructive propensities of an individual by applying information mining strategies to the instant messages an individual ships off the related individuals. By breaking down the segments of the instant messages (catchphrases and emojis) the self-destructive inclinations of an individual can be assessed so vital advances can be taken to save the existence of the subject.

#### Dataset

The information is gotten by removing all the instant messages send by the subject. This can be accomplished from numerous sources like Facebook, WhatsApp, and so forth Every one of the messages send through these informing administrations are put away in an information base where the model can be applied and investigation for the notions should be possible. The informational index will contain text type of information and emojis. No other type of information, for example, pictures will be investigated through the model.

#### Model

##### Sentiment Analysis

In this part the information is allocated an opinion, for example, positive or negative and the degree of it by performing information pre-preparing utilizing SVM calculation. Text pre-preparing is required due to the nature of the dataset. The model consists in total of the following components:

* Tokenization
* Data Standardization
* Emoji Conversion
  + Positive Emoji
  + Negative Emoji
* Stop-word Removal
* Stemming
* N-gram
* Term Frequency
* Inverse Document Frequency
* SVM
* K-NN

#### Results

The outcome got from the proposed model gives the assessed feeling forecast of the subject dependent on the instant messages sent by the client. The subsequent yield can be utilized much of the time, the psychological issues and feeling of anxiety is assessed and in this way in the event of "basic" suppositions the companions and relatives of the subject can make moves to support, persuade and elevate the passionate height of the subject accordingly bringing about the congruity and significant serenity of the subject. Consequently, such conclusion investigation models are a prerequisite for forming the general public into an occurrence place.

## Project Timeline

Keeping track of the project by assigning deadlines to different modules and submodules in a project can help digest and comprehend the project easier and is crucial to its completion. To this extent, the Gantt chart is a fantastic tool to give life to the project plan and timeline. Using time as the X-Axis, the project module and submodule timespans are stretched onto the graph as shown below in *Figure 1* for this project.

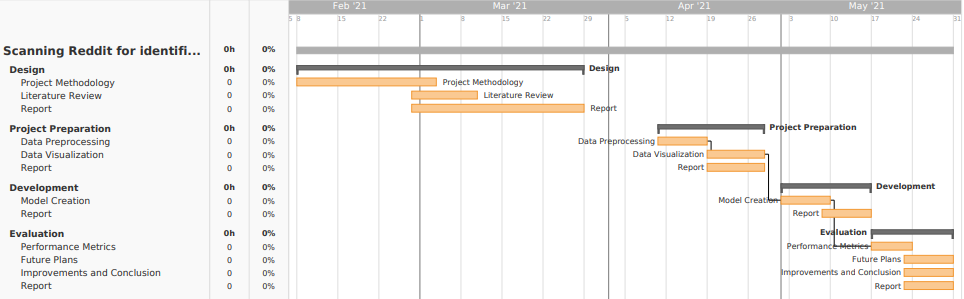


Figure 8 - Project timeline with Gantt chart

# Methodology

The different parts that are needed to procedurally direct the task towards fruition make up the Methodology. Execution of this strategy gives a bunch of decides that makes advancement more smoothed out and centered towards the primary goals.

The purpose of this research is to experimentally prove that with Machine Learning, finding relevant information from Reddit is relatively easy. To this extent, the following is done.

## Project Design

AI situated tasks include weighty experimentation and result examination. These examinations change as indicated by the information and the outcomes accessible. Research projects like these don't follow any Development Life Cycle, accordingly there won't be an improvement life cycle for this research.

## Experimentation

Another factor that needs to be considered before proceeding with any Machine Learning project is the dataset. The dataset also plays an essential role in deciding what algorithm will be used for the project. All of this means that the dataset must be handled properly and separately before proceeding with the machine learning part of the project.

### Dataset

The dataset used would be the Reddit Flair Dataset available from [Kaggle](https://www.kaggle.com/hritik7080/reddit-flair-dataset). The dataset contains the Reddit posts of Indian region. This data was taken out from Reddit with the help of their easy to use API. It contains various features such as post's title, URL, description, flair etc. It contains approx. 220 posts for each of the following flair:

* AskIndia
* Non-Political
* Scheduled
* Photography
* Science/Technology
* Politics
* Business/Finance
* Policy/Economy
* Sports
* Food
* AMA
* Inspiration

The main target is to develop a prediction model accurate enough for predicting the flair of a Reddit post. Although the dataset is designed with a different aim in mind, the requisite features that we require for the project are available in this dataset and the dataset will be molded to suit the needs of the project.

### Machine Learning

This project will take the route of Natural Language Processing (NLP) along with binary classification. The essence of language processing comes in this project from the use of the “nltk” library for creating a "Bag of Words". The bag-of-words model is an improving on portrayal utilized in regular language handling and data recovery (IR). In this model, a content (like a sentence or an archive) is addressed as the bag (multiset) of its words, dismissing language and even word request yet keeping assortment. This will make it easier to apply Random Forest classifier on the dataset thus giving it the probabilistic edge in prediction.

The tools for this research project will be Machine Learning Libraries like Keras, Tensorflow and scikit-learn that will provide the necessary classification algorithms required for this project. Although random forest classifier is finalized, there are other options like Neural Networks and SVM that can provide better results. All of this can be finalized after testing. NLP tools mentioned above is the library ‘nltk’.

# Conclusion

With the exponential speed that internet technology is developing with, the popularity of social media will only soar higher. The amount of data this brings to the platform is huge. To make better use of this data, Information Extraction techniques are being researched for application on social media data. This research focuses on identification and extraction of Science related posts from Reddit. This is achieved using Natural Language Processing. A bag of words is created using natural language processing. This bag of words is then passed onto a Random forest algorithm for the purpose of training to classify science-based posts on Reddit.

# Project Evaluation

The assessment of this examination will be finished utilizing the accompanying data:

* The yield of the presentation measurements subsequent to estimating the calculation execution.
* The measurable examination of the dataset.

# References

1. Bontcheva, K. and Derczynski, L., 2016. Extracting Information from Social Media with GATE. Working with Text, pp.133-158.
2. Brambilla, M., Ceri, S., Della Valle, E., Volonterio, R. and Acero Salazar, F., 2017. Extracting Emerging Knowledge from Social Media. Proceedings of the 26th International Conference on World Wide Web.
3. Mishra, S., 2019. Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets. Proceedings of the 30th ACM Conference on Hypertext and Social Media.
4. Ahmad, S. and Varma, R., 2018. Information extraction from text messages using data mining techniques. Malaya Journal of Matematik, S(1), pp.26-29.
5. Qian, Y., Santus, E., Jin, Z., Guo, J. and Barzilay, R., 2019. GraphIE: A Graph-Based Framework for Information Extraction. Proceedings of the 2019 Conference of the North.