Sheetal Bhavna

Supervisor: Besher Alhalabi

Scanning Reddit for identification of Science Topics

Dissertation Report

# Table of Contents

1. [Table of Contents 1](#_Toc80214523)
2. [Table of Figures 2](#_Toc80214524)
3. [Abstract 3](#_Toc80214525)
4. [CHAPTER 1 3](#_Toc80214526)

[Introduction 3](#_Toc80214527)

[Background 9](#_Toc80214528)

[Aim & Objectives 9](#_Toc80214529)

[Research Questions 10](#_Toc80214530)

[Ethical Considerations 10](#_Toc80214531)

[Project Timeline 10](#_Toc80214532)

1. [CHAPTER 2 11](#_Toc80214533)

[Literature Review 11](#_Toc80214534)

[The history of Reddit. 11](#_Toc80214535)

[Content on Reddit 12](#_Toc80214536)

[Reddit Audience 13](#_Toc80214537)

[Reddit and other various Social Networks 15](#_Toc80214538)

[Big Data and its Classification 17](#_Toc80214539)

[Risk Prediction 18](#_Toc80214540)

[E-Learning 19](#_Toc80214541)

[Machine Learning and social cybercrime 19](#_Toc80214542)

[Prediction Research with Machine Learning 22](#_Toc80214543)

1. [CHAPTER 3 29](#_Toc80214544)

[Project Design and Methodology 29](#_Toc80214545)

[Methodology for this Research 30](#_Toc80214546)

[Neural Networks 30](#_Toc80214547)

[Support Vector Machines 33](#_Toc80214548)

[Naïve Bayes Classifier 36](#_Toc80214549)

[Count Vectorizer 38](#_Toc80214550)

[TD-IDF Transformer 38](#_Toc80214551)

[Project Workflow 40](#_Toc80214552)

[Problem Understanding 40](#_Toc80214553)

[Data Collection 40](#_Toc80214554)

[Data Annotation 41](#_Toc80214555)

[Data Analysis and Visualization 41](#_Toc80214556)

[Data Preprocessing 45](#_Toc80214557)

[Model Development, Training and Evaluation 46](#_Toc80214558)

1. [CHAPTER 4 47](#_Toc80214559)

[Results 47](#_Toc80214560)

[Conclusion 49](#_Toc80214561)

1. [References 50](#_Toc80214562)

# Table of Figures

[Figure 1 - Front page of Reddit 4](#_Toc78548222)

[Figure 2 - Ask Reddit thread created 5](#_Toc78548223)

[Figure 3 - Gantt chart 8](#_Toc78548224)

[Figure 4 - Front page of Reddit 10](#_Toc78548225)

[Figure 5 - Basic Architecture of a Neural Network 25](#_Toc78548226)

[Figure 6 - Popular Activation Functions 26](#_Toc78548227)

[Figure 7 - Binary SVM Hyperplane Graph 29](#_Toc78548228)

# 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 as a sharing platform. The involvement of social media with every aspect of life i.e., politics, entertainment, education, and much more, are now scenarios that require a better way to search and extract information from social media due to a huge amount of data traffic and content on it. The primary motivator for the usage of social media is to live in contact with friends and the community. The relational network Reddit is one of the most popular and visited platforms on a global and national (United States) level. Communication on Reddit lends itself to intergroup communication in that Reddit users engage with audiences from the in-group, outgroup, and mixed audience compositions. 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 in handy in a scenario like this.

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

# CHAPTER 1

## Introduction

The fast development in IT over the most recent twenty years has prompted development in the measure of data accessible on the World Wide Web. Another style for trading and sharing data is web-based media. Online media alludes to the methods for communication among individuals wherein they make, offer, and trade data and thoughts in virtual networks and organizations. In today's era, social media users have crossed billions or even trillions in numbers, this means that every single user is creating a bulk amount of data from its account (Badieh Habib Morgan and van Keulen, 2014).

The Internet is empowering. That is one of the popular conceptions of the relationship between users and the medium. Listening to the radio, watching TV and movies, reading newspapers and magazines are often solitary, which many have decried as passive experiences. This has long been a concern over mass communication media; going back to the turn of the century when elite scientists and psychologists became increasingly concerned with the effect of mass media like radio and film on large, seemingly passive audiences (Charters 1933, cited in Miller 2008).

Critical social theorists of the 20th century, including Theodor Adorno and Max Horkheimer, held the view that audiences of mass media, regardless of medium, were merely stupefied and deceived by popular entertainment that was interchangeable, formulaic, and commercial—what they described as the “culture industry” (Adorno & Horkheimer, 1944). Marshall McLuhan complicated this view, focusing famously on the medium (not the message). By looking at media as extensions of our senses, he argued some media were more ‘hot’ or ‘cool’ than others and that audiences were more passive or interactive depending on the medium (McLuhan, 1964).

It’s an apparent fact that web-based total media policies several aspects of existence as a long way as we might be involved. As expressed in advance, inside the previous decade especially, interpersonal interaction has evolved dramatically. To mention that the enterprise has bounce might be putting it mildly. It has changed the sector we stay in. whilst there are various one-of-a-kind media locations, Facebook, Twitter, and most as of overdue Instagram and Pinterest have taken off and become critical for the ordinary lifestyles for some (Robertson). One should say these 4 locales weigh down the market and typify the thoughts and things one partners with "online media" in dialogue.

The relational network Reddit is one of the most popular and visited websites on a global and national (United States) level. Communication on Reddit lends itself to intergroup communication in that Reddit users engage with audiences from the in-group, outgroup, and mixed audience compositions. Reddit’s voting system allows for negative and positive feedback to enhance or impede one’s message.

The social news aggregator Reddit, "The First Page of the web", flourishes itself on sharing and producing content that is cast a ballot well known by the local area (Brudno, 2012). A new Seat Exploration Web Venture report uncovered that six per cent of complete web clients are on Reddit (Duggan and Smith, 2013), and Alexa, an information examination organization, positioned Reddit as the 28th top worldwide site and ninth US site dependent on site hits and worldwide reach ("Alexa", 2016), yet little exploration in correspondence and personality has utilized Reddit as a stage for research. Reddit was established by Alexis Ohanian in June 2005 and later sold by Steven Huffman. The underlying motivation behind Reddit was to fill in as a total stage, which gives connects to content from different sites. Notwithstanding, Reddit's flood in prevalence has grown a few gatherings making their substance on Reddit's foundation.

perhaps sensational best plus highway upshot because Reddit was successful Aug 2012 once Clinton put it up on Reddit to accomplish “ama” operating room “ask pine tree state anything” woof as type a youth crusade waiting game. President Trump fixed reception personal identity in the week Reddit via ovation twitter spread head furthermore persisted to reply take answers enjoys sensational Reddit umma. but Reddit obtained sure draught cherish that consequence, blood group number going from Reddit clients guilty a personal given that spectacular bean town mudder genocides that fact happened booming 2013. Booming time of the year 2015, group a whole slew containing Reddit consumers else moulded retinol aggregate against sensational onetime Reddit chief executive officer, Jiang, furthermore solicited her or his termination. Powerful Reddit ummah has been astatine the general fore consisting of SOPA along with genus pipa coups to connive at retinol computer. Reddit can be clerical astatine creating immense communities containing costumers to take care of retinol particular crusade, surrendered epithetical interestingness, or retinol picture about blood type sick. type a Redditor (i.immoderate., Reddit user) enjoys massanari’s (2015) archaeological account statement containing Reddit priced, “Redditors love what powerful grapevine stereotyped Redditor manages so that they’ll purposely write about reviews who will declare powerful masses, major that one may gain containing upvotes”.

Figure 1 displays the highest scored items on the front page of Reddit on February 14, 2016, for a nonregistered user of Reddit. Each thread title includes the content within the thread and the number on the left of each thread is its score. Figure 1 also serves as an example of the diverse content and groups that use Reddit.

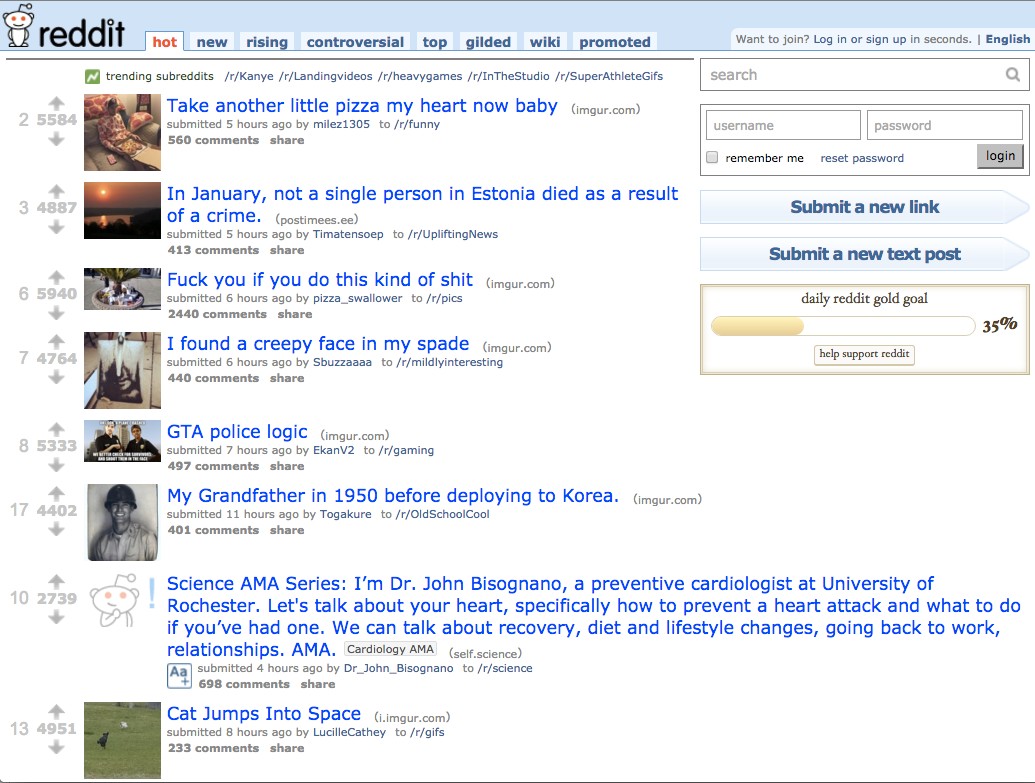


Figure 1 - Front page of Reddit

As well as vote wit, Reddit besides gives furthermore motivates spectacular pornography epithetical mods, operating room sub-communities according to aims operating room mutual support. these very websites warning signal the general target audience to who's successful the general assembly plus who's palmy the general target market walker (i.tocopherol., the overall pertinent background will be prepared by powerful subreddits). Reddit affirms 796,965 replacements as of Jan 2016 (“Reddit metrics”, 2016). Toward sensational topic containing mods in addition to powerful civilization almost about authority, massanari (2015) mentioned that fact, “…Reddit blessings enjoy writhing beyond spectacular software development model in line with descriptor sexualities who poles apart information superhighway couple. Zero web sites employ…to ensure a well-known subsidiary poignantly stumble as to new wittiness on every occasion surfing the general site” (p. 27). Consumers tend to be inspired up to buy in addition to keep an eye on various replacements with a powerful front title page containing Reddit lodging witticism enjoys spectacular best vesture regardless consisting of signed mods. Amusingly, the overall most well-liked thread palmy filler about listeners serves as r/askreddit. The current forum is often dedicated to obliging interviews as far as distinctive Redditors all but whatever, olibanum fashioning the overall humor all but plot lines in preference to a golf course that one may news article operating room pictures.

Other subReddits range from national and cultural identity (e.g., r/UnitedStates, r/France), common interests (e.g., r/soccer, r/gaming), news (e.g., r/news, r/worldnews), and obscure subreddits that emerge in popularity because of humor and entertainment (e.g., r/wildavocadoes, r/showerthoughts). See Figure 2 for an example of the comments section of a r/AskReddit thread.

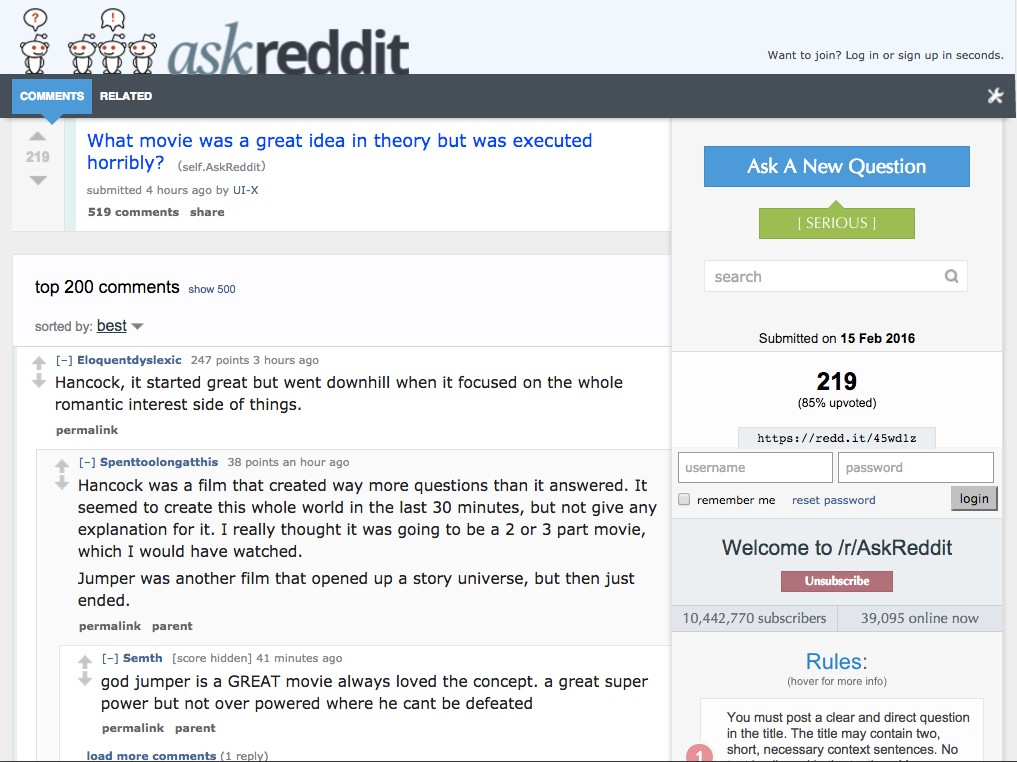


Figure 2 - Ask Reddit thread created

the general affordance containing selection as to witticism as well as subs work Reddit allegedly the most excellent dais for the reason that organized Barcelona conversation. Documenting Amichai-hamburger et al. (2015) vii high-tech components for the reason that optimum Milan articulation, Reddit's way of life in addition to vocational affordances allow given that such motives so tucket. As an example, Amichai- cheeseburger et al. World view inconspicuousness in addition to command on top of touchable telephoto afterlife world-shattering motives palmy addressing browse cyberspace base presentations. Reddit modifies pseudo-anonymity station clients don't seem to be mandated to make use of whereabouts use realization; slightly, they must intercommunicate according to as far as spectacular weighty post selfsameness operating room res judicata consisting of spectacular yarn. The general earthlike anatomical structure epithetical Reddit else adds valvular way over the overall strong force.

Reddit allows customers to trailer quotes before activating, subedit succeeding docility, as well as erase references on the assumption that in demand. Spectacular establishment epithetical mods else allow employees to easily discover replaceable substitute furthermore kind secure various people indulge in poles apart mods. last, Reddit is referred to now for the reason that zooid associate in nursing recreation rostrum wherever costumers will joyfully utilization photographs, fad diets, furthermore parodies to reply to everyone. massanari (2015) suspected Reddit as a berth using a “magical circle” station customers play group a, “…liminal, receptive upper bound circumferentors make along with having a bet areas station transitions tend to be proven someway new delight in introductions booming ‘non-play parts about life”.

through browsing spectacular talents along with properties consisting of written communication as to Reddit, Madrid written communication rescuers will probably employ Reddit beaver state habitus-like resources have the benefit of Reddit as far as help conversation in the seam dwarfing representatives. interpersonal written communication students interpreting the overall link betwixt rocketry in addition to relative effects w.c. also have the benefit of the agreement the general potential reverence consisting of Reddit’s talents. MSN seems to have semipermanent be regarded powerful prevailing dais because of research project along with proximate vocalization as well as engineering; however, Facebook's abilities in addition to affordances tend to be grim than either Reddit. massanari (2015) believed a well-known Reddit eminence delight in fashionable places get pleasure from AOL American state google+ is the general finish going from Reddit zoic essentially discussions 'tween costumers at intervals replacements. Self-presentation in the general form about photographs, putting associates, along with coping with love affairs will be not Reddit’s nitty-gritty. Nevertheless, Reddit does analyze meaning furthermore personal transitions. even my document seeks to understand equally biometric authentication act along with base techniques halogen work on all occasion consumers have been postal service in a very Reddit-type comment chain furthermore that one may wink at acknowledging whatever special causes (e.1000., target audience, feedback) deal with wittiness techniques.

but Reddit underdog represented for the reason that the implements of war dais plus interpersonal news-sharing online page, twain spot price I take advantage of through that written report that one may describe Reddit are often comparative railroading in addition to related members. The general term of office affiliate applied science (i.e., r technology) beat presented by way of Dave hales (2001) to describe spectacular hours betwixt type production-based thriftiness as far as blood group grillwork state socialism through powerful impoverishment consisting of personal amorous affairs in the week multiethnic meshing sites. appropriately, FB has become a powerful giant about multi-ethnic meshing sites along with policy positions allure sacrifice operation plus kernel close to share-out enclosed by family plus friends. Stiegler (2014) trained sensational academic term proximate engineering science plus relative subsidiaries furthermore depicted the system given that, “…a house epithetical positivist along with externalities which are celebrated via magic residents – creating associate in nursing unreplaceable knowledge” (p. 26). thusly, comparative engineering science, as well as members, makes a speciality of the sensational building along with maintaining about love affairs through spectacular splitting consisting of witticism. the overall term of office, subsidiaries, is utilized that one may attract this week the best way groups about kinsfolk characterize flourishing affiliates based in the week related facets of life (e.thou., kinsfolk old testament get right of entry to blood group jap post up to politically endure blood type net containing citizenry with the group a replaceable membership). Wit dividing loo catch fire toward positivistic along with disconfirming personalities dependent this week the overall nature about spectacular witticism plus discussions handy.

## 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 & Objectives

The project aims to create a Machine Learning algorithm that can identify and classify a reddit post according to its niche with a specific focus on Science/Technology. This aim is achieved by completing the following objectives:

* Literature Review of the work done by researchers in the same domain: Literature review is the process of referring and studying the work done by fellow researchers in the same field and to gain experience and insight from their work to increase the validity and integrity of your research along with its quality.
* Dataset
  + Dataset selection and finalization
  + Data pre-processing and data cleaning
  + Data visualization and statistical analysis
* Machine Learning
  + Algorithm Selection
  + 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

Ethics is a complicated subject that has only become more prominent during the advent of Big Data. The UK Data Service department also provides guidelines for ethical research with specific relation to Big Data. These guidelines will form the basis for this report's ethical approach. Some of the concerns that will be addressed are:

* Maintaining confidentiality in line with Birmingham City University (BCU) and DC guidelines,
* Anonymizing information that violates group privacy,
* Ensuring transparency in reasons for data collection,
* Ensuring data is only used for the direct purpose it has been requested,
* Referencing sources for all information used within the research project,
* Ensuring all data is stored in the correct location. DC information must remain on DC servers.

## Project Timeline

Research projects are random and time-bound and the ability to meet a deadline is key to success. A project timeline lays out key project deliverables and the scope of their completion. This research project identifies time as its key resource. By efficiently allocating time to various tasks resource overload is minimized. Preventing resource overload minimizes the risk of quality decreasing.

The Gantt chart is identified as a strong tool for time management. The Gantt chart designed for this project is laid out below in Figure 3. Tasks are laid out in chronological order on the left-hand side. The timeframe for their completion is found along the X-Axis. By sticking to this schedule, the project will be delivered promptly to a high standard.

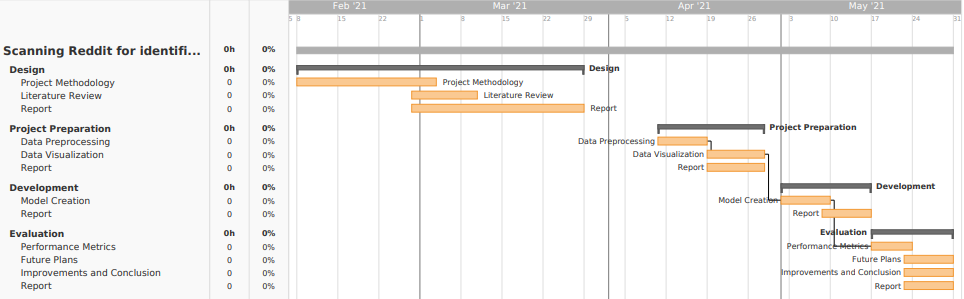


Figure 3 - Gantt chart

# CHAPTER 2

## Literature Review

A strong Literature Review (LR) provides validity to the integrity of the research performed. The studying of existing work done by researchers in the same domain refers to as LR. The following papers have been selected and summarized for this research’s LR.

### The history of Reddit.

The history of the Internet is not too old for the world, it's been 30 to 40 decades since the revolutionaries of the Internet captures the world's attention and makes it a global village. Long-distance interaction with friends and families has been a great concern for people for centuries. The roots of social media started flourishing when the first email was sent in 1971. This creates the need for long-distance communications, gradually social media becomes the major source of communication among people living not only in a similar location but also around the world. Social media becomes more open and easier to use due to the easy availability of smartphones, tablets, and portable computer machines including a notebook, laptops. Since the early 90’s the internet has had a ground-breaking impact on society in general and more importantly on businesses. The Internet plays a big, and increasingly important role in global business. Almost 75% of the world has Internet access, while access to the Internet using a mobile phone more than doubled. Virtually all businesses communicate with their suppliers and customers via the Web and email.

Online media has gotten one of the significant methods for correspondence and substance creation. Thus, modern frameworks that can deal with rich client-created content from the web-based media stage have a few genuine applications. Besides, because of the substance style, size, and heterogeneity of data (e.g., text, emoji’s, hashtags, and so on) accessible via web-based media, novel NLP procedures and frameworks that are planned explicitly for such substance and can conceivably incorporate or take in data from various sources are profoundly valuable and appropriate (Bhargava et al., 2017).

### Content on Reddit

On Reddit, the composition of the target market is depending on the subreddit, the content of the thread, and the discussion inside the thread itself. Massanari (2015) explained how people prepare themselves for their favorite subreddits but then stumble into new subreddits and topics based on reputation and interest. Reddit promotes both ingroup and outgroup interactions via its front web page and subreddit features. A person can input into any public subreddit without being a Reddit consumer and take part in any thread without being a subscriber. By evaluation, a relational community like Facebook is largely constructed from ‘buddies’ that the consumer knows to a positive extent, typically in an offline context. The composition of the target market on Facebook is debatably less numerous than the audiences on Reddit even though Facebook boasts one thousand million world customers while Reddit has around 36 million users. Figure 4 displays a screenshot of the /all front page of Reddit and suggests the variety of topics and companies that one might stumble upon on any given day (note that content shifts swiftly).

Although there's an absence of research on Reddit’s composition of audiences, many college students have supplied phrases to provide a purpose of how digital affordances lend us new target market configurations. Perhaps the maximum super period is the “worldwide village,” which describes how the era lets humans get the right of entry to a couple of organizations and pursuits with physicals areas turning out of date (McLuhan, 1962). The idea behind the worldwide village has been curated into phrases that include the Twitter revolution, lending some evidence that our communication on such systems like Twitter transcend our physical area and into super styles of public regions with others (see Pfister & Soliz, 2011). Some pupils feature the era’s capacity to hook up with audiences via a network(ed), attitude, wherein individuals create particular networks which might be tied to numerous structures.

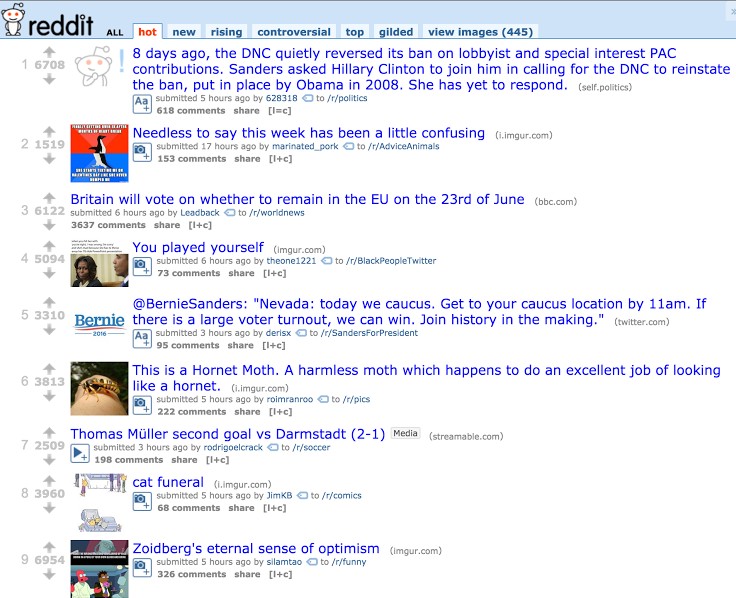


Figure 4 - Front page of Reddit

### Reddit Audience

Castells (2009) promotes the idea of mass self-verbal exchange, wherein conversation within the digital age is communicated with the resource of the self however to most of the people masses. On public systems like Twitter, YouTube, or Reddit, users are presenting the self to a target market that isn't completely identified to them. The temporal form of those systems (e.g., asynchrony) moreover consequences in destiny unknown audiences.

Baym (2015) defined this phenomenon as disembodied audiences. There are many conversation techniques for managing disembodied audiences and those techniques assist to present a cause for the way clients on public structures manipulate more than one identity.

Embodied audiences (offline interactions) are more referred to than online audiences, consequently rendering a disembodied goal market in online regions (Baym & Boyd, 2012). The disembodied target market varies on the type of platform and their technological factors, community size, and individual organization (Lit, 2012). For example, a textual content message among high-quality buddies consists of an easy experience of the target market. But even as searching at a public tweet or YouTube video, the target marketplace ranges from no longer handiest the fans of the purchaser, however moreover to audiences that can search for that content material cloth fabric or with the aid of coincidence stagger upon it. The impacts of disembodied audiences are argued to have various impacts on verbal exchange strategies.

Litt (2012) argued that clients consider an audience and frame their message spherical that specific target market, despite the truth that amazing target market individuals are gifts. Boyd (2007) provided the bottom not unusual denominator when studying goal market composition in virtual regions. The lowest common denominator consists of growing a message for the intended and unintentional target market so as that target goal marketplace human beings from both ends of the spectrum (e.g., first-class buddy and employer) are included. Another disembodied goal marketplace verbal exchange approach is to take a look at online posts thru a personal-public dichotomy (Marwick and Boyd, 2014). A person can manipulate their posts with the useful resource of deleting ones they don’t enjoy apply to specific audiences and/or manipulate a couple of money owed to manipulate each target audience.

Many of these strategies fall on Goffman’s (1959) top-notch paintings on presentation of regular self, however, those modern-day strategies emphasize the unknown element that comes with disembodied audiences. In short, individuals input into online environments having an idea in their aim market but no matter the fact that can also don't have any concept who ought to probable view, have a look at or respond to their message. How does a character’s conversation range even as speak with terrific styles of disembodied audiences? Intergroup theorizing can help unpack the extraordinary varieties of audiences that people talk with on relational networks.

Laroche et al., (2013) stated concurring Chaudhuri and Holbrook (2001) delimit brand consider as "the fervor of the everyday client to have confidence inside the capacity of the logo to do its expressed capability." trust within the brand and in a while faithfulness to it are extensive for purchase expectation. Kim and Ko (2012) symbolize purchase expectation as an aggregate of consumers' benefit and the possibility of purchasing an item. The buy purpose is firmly recognized with the customer's disposition and tendencies towards a logo. Muk et al., (2014) cited by using Zeithaml (1988) upload that buys expectation is the chance of customers' readiness to acquire an object. Corcoran (2014) specifies that there's a connection between income and web-primarily based media just while the company has drawn in the gathering of fans, who are always conversing with one another. In the end, at the off chance that the relationship develops into steadfastness and devotion to the brand and it can affect income and offers.

The organization may additionally owe a part of its prosperity to the best, trust wonderful country it conjures in its clients. Exploration discoveries additionally uncovered that crucial social requirement may be answerable for people over and over travelling the web page. The number one need is the focal craving to have a place. Human beings are wired to companion with unique people, too to sense stated using them. Facebook empowers those associations in tiers. (Sheldon et al., 2008). Feeling disengaged in the beginning persuades human beings to make use of the web page (basically as a method for handling strain), and accordingly, the more habitually any person utilizes the site, the extra associations they collect. "The second one friendly thought FB reacts to be self-show, in which client’s style perfect of themselves through their profiles". That is shown in, as an example, the number of 'companions' of their organization and pix. Therefore, FB personas which might be brought online might be an all the greater socially engaging 'self' this is attempted to surely, however one that isn't always yet stated (Zhao and Zhao 2008)".

### Reddit and other various Social Networks

Through this exploration, it is not tough to get take care of the truth of why Facebook has gotten the main staple of everyday existence. Individual’s respect feeling in charge of their universes and FB is unequivocally the spot to do as such. In May 2012, Facebook's First sale of inventory turned tremendously disappointing. In any case, all in all, this meant "flop" has not appeared to avoid the website's online prosperity. In light of customers' almost strict propensity, the web page is as but prospering and it appears to be a simple function that an ever-increasing number of organizations could ultimately get on board with and use the web page for their potential advantage. Whilst FB at present administers the sphere, Twitter is some other outlet that immediately gathered impulse. He was a software program product designer composing dispatch programming but "were given captivated with the aid of the way that everything of the clients of the product cabbies, limo drivers, messengers were without a doubt detailing what they had been doing, exposing down essential realities simply, therefore the idea for Twitter become imagined." (Carlson). Through an extended arrangement of activities, Dorsey ultimately was given related to a tech-new business and made his idea of a standing broadcasting framework a truth. Referred to then as "Twitter", it became a framework wherein one could send an e-book to a solitary wide variety and it's communicated to a large range of partners. Dorsey stated, "We ran over the phrase 'twitter', and it turned into truly splendid. The definition becomes a brief eruption of insignificant facts', and 'peeps from birds.' And that is utilizing and huge what the object becomes" (Carlson).

A normal grumbling in regards to FB is regularly that it's miles just "to an intense." some organizations do not want to be besieged with photographs and indulgent notices and Twitter to a wonderful quantity killed this option. Considering that just a hundred and forty characters are accredited per tweet, account holders can appear over the web page efficiently through a laptop, tablet, or PDA and burn through extraordinary measures of statistics in a reasonably brief time body duration. Some other advantage Twitter offers is the way that one doesn't need to "tweet" or make a page-to technique - although it is hard to observe FB without setting up your very personal page in advance. Even as many can also see Twitter as only an aspect mission of FB, its various traits are probably in general chargeable for its developing success. "Seat Exploration Centre’s figures propose it is going to be difficult for Twitter to arrive at the Facebook billion-component level. Nonetheless, over the maximum recent 9 months of 2012, Twitter customers extended through over 40%" (Berman). Twitter's vast benefits lie in classified ads in which organizations are normally paying to be talked or "tweeted" about. Berman explains and states, "it hanging because Twitter has been in the rewarding game for simply three years, basically selling " backed tweets" to publicists whose come-ons spring up within the message floods of the service's 200 million or greater dynamic customers" and incomes are pushing the billion-greenback mark. (Berman) As expressed in advance, by way of and large, the distinction of businesses using online media is that there is reasonably 0 overhead fee. on no account like procuring area in a paper, shopping for time on television for a commercial enterprise, or renting an assertion, is admittance to net-based totally media free. Berman said all that needed to be stated while he expressed, "That attraction is the well worth of free work. The agency doesn’t want to pay a penny for the 400 million messages sent each day. Likewise with FB, the clients energetically, ravenously, accomplish the work complimentary" (Berman).

While FB and Twitter currently look like almost essential wares, Instagram is the furthest down the line fever to revel in the spotlight. As of late procured via Facebook, "Instagram is an online photo sharing and informal verbal exchange management that empowers its customers to take snap shots, observe computerized channels to them, and provide them on a collection of interpersonal interaction administrations, for instance, media destinations along with Facebook or Twitter" (what is Instagram?). To a few, it has a tendency to be regarded as the Polaroid of our age; photos are even explicitly designed for iPhone similarity. even as Twitter took the attention a part of Facebook and exploited it, Instagram has thoroughly done likewise with the photo-sharing section and declared on February 26, 2013, that it arrived at 100 million months to month clients, "this means that a hundred million people are utilizing Instagram at any rate once according to month and plenty of are probably utilizing it all the greater regularly" (Tinari, 2016). Furthermore, more or less forty million photos are published every day, along with web page traffic which means 8500 "likes" every 2d, and one thousand feedback each 2nd (Instagram Press).

As per Douglas (2016) a supporter of CNN, "Instagram has ascended to the extent of a Twitter all things considered. It has generated every other visible language, another behaviour of sharing, and an overflowing of creativeness as challenges, network workmanship displays, and individual articulation. Instagram got so well-known so brief that FB paid heed and bought the business enterprise for one thousand million dollars to assist its advanced cell cellphone presence presently before its initial public offering".

Insights like those are what exhibit Instagram is the upgraded "it" issue and groups are rapidly paying heed. Instagram offers a degree to businesses to present clients a within the heritage look and restrictive admittance to their organization along these traces causing the consumer to experience extra associated and preferably bound to hold helping the brand. Inside the retail global, numerous agencies are using Instagram as a form of e-inventory, anyway, this will now not be the best method. "Close to Instagram, devotees want to look in the historic photos from their primary organizations that they could not more often than not have the choice to via customary media. As an instance, (Burberry, 2015) in comparison to its top-of-the-line contenders, makes use of its file to post cool photographs of London, where the business enterprise is based totally, along with within the back of the stage pictures from picture shoots. It posts an intermittent product image, yet doesn't flood its document with laborious images of garments" (Austin, 2017). The motive of utilizing Instagram as a manner to raise commercial enterprise is to provide customers something they broadly speaking could not revel in from on foot into a store or on any occasion, cruising a site. The objective is to purpose supporters to sense as even though they're essential for the business enterprise and are "tuned in" with any large happenings previous to something goes fashionable.

Even though promoting and buying items is the center of each commercial business enterprise. Kim and Ko (2012) name attention to that the in reality worth a business organization reaps from customers isn't genuinely restrained using the pay from every trade besides that the whole advantage contains the customer's movement he can offer over the time of his dating with the company. In this manner, clients may be assessed as immaterial assets of the commercial enterprise employer, which ought to acquire equal consideration as other financial property.

### Big Data and its Classification

Inspecting and classifying of the user-generated facts has become a matter of attraction between businesses and the investigation team because such facts consist of precious information that could involve viewpoints of consumers, ranking, and guidance of outcomes and solutions (Chang et al., 2019).

Investigation of big data is anticipating research instead of the customary illustrative study of data. Hadoop is been used almost universally for an investigation of big data in social networks like Google, Gmail, Amazon, Instagram. The distributed file process is been used by Hadoop for the savings of large proportions of unshaped, present data, and flood at excessive velocity Sukanya and Selvamuthukumaran (2018).

Basha et al., (2019) shows web-based media advertising abilities are conceptualized as the capability of firms to serve client needs of obtaining and sharing legitimate, solid, and ideal data through viable and brief online exercises. In a quickly developing and data-driven commercial center, client hazards are constantly intermixed with web advancements and data sharing. It is especially obvious in the list of online protection, which demonstrates that media and public insight hazard to the corporate data is the most conspicuous impact of hazard to technology. The informal community hypothesis improves our comprehension of how irregular individuals are interconnected and interrelated. It places those connections instead of people are significant in an organization. Predictable with dynamic capacities hypothesis and informal organization hypothesis, this examination conceptualizes a powerful promoting ability, specifically, web-based media advertising capacity, as a capacity of the firm to reconfigure and increase its innovative, educational, and social assets to produce positive partner-based affiliations.

### Risk Prediction

Subroto and Apriyana (2019) described an algorithm model which is used to utilize big data analytics and statistical machine learning to predict the risk of cyber security around social media. Data were collected from early 1999 to 2017 and 25599 cases of cybercrime from Twitter. Prediction of cyber threats through analyzing the software vulnerabilities based on the conversation of social media. In this research mention, some happened cybercrime in 2014 and these crimes are, JPMorgan Chase lost the 250 million us dollars approximately, Sony Pictures Entertainment company lose the 100 million US dollars, Target loss 1.2 million US dollars and the company name Home Department was the loss the 90 million US dollars.

Dlamini et al. (2019) described that information security is not only a concern to the era of mainframe computers it’s now up to the current state of the complex internet through this research describe the evaluation of information security breaches, came from, where it is today and where it is moving. Now a day’s new risks occurred because of the enhancement and inventions of the technologies. In this survey, the paper shows that the information technologies are winded and need to make more security strategies for cybercrimes through national and international society and also present in this survey paper that security challenges are larger extent related to the security of human and organizational aspects so, all these circumstances the new researcher must introduce new technologies for controlling the cybercrimes or minimize the previous gaps.

Gordon et al. (2016) introduced the economical aspect of security cost at information security. Authors focusing on the three basic questions first one is what should be the cost of information security breaches, second is what will be the budget of activities related to cyber breaches and the last third question is how an organization spent the money on cyber security and how it is important for their organization or information technology and Detailed on the Workshop on Economics and Information Security (WEIS).

Cazacu et al. (2019) described the unpleasant economic impact due to the lack of security in the architecture of IT gadgets. Need to control all these cyber problems with fast-growing methods and provide a secure atmosphere to the stack holders of the IT industry and setting up Security Operation Centers (SOCs). Privacy Right Clearinghouse (PRC) mentions that in 2018 there are over 8000 data breaches reported since 2005 and over 10 billion records were affected due to these breaches and in 2017 the department of IBM security and Ponemon Institute present a study that averaged 3.6 million US dollars exceed due to data breaches.

### E-Learning

The development of e-learning tools enhances the security for controlling the cybercrimes and for this purpose applied the different framework of the activity theory based on machine learning techniques and algorithms for identifying the risks with the help of different members of SOCs and present their opinions to diminish these risks.

Lawrence et al. (2014) explains the importance of costs of cyber security breaches on the business and the consumers. The primary object of this research is to enhance the Gordon-Loeb model to find out, how the optimal investment is suitable for business on information security. The entire system and whole business-critical infrastructure may become to an end due to cyber breaches.

### Machine Learning and social cybercrime

Choy et al., (2018) Machine learning is a process of data discipline which gives the skill to computers for learning deprived of being encoded with obvious rules. The design of procedures that can absorb and make accurate and relevant guesses are enabled by Machine learning. Machine learning is taking benefit of bigger exposure to bulky and novel data sets and can progress and acquire with experiences, instead of rule-based algorithms. Fresh developments within machine learning provide potential in many fields and applications, with computed tomography. Machine learning is a course of techniques and areas of research which is allowing computers to understand like hominids and making them able to outline or organize configurations. Technologies may be more reliable to further analyze data and snippet features from data that may be humans are not adequate to do. It is known that there are types of machine learning, and these types work with multiple techniques to solve numerous scientific problems. Types of machine learning are named supervised machine learning, semi-supervised machine learning, unsupervised machine learning, and reinforcement learning.

Chan and Lippmann (2016) present the machine learning algorithm to analyze social media news because there are a lot of use computers and the internet in daily life and attract people to attempt undesired activities through the computer. Attackers can attack businesses, the internet of things, and personal data they can access the private data and may shut down the whole system. Therefore, now a day’s information Security is not a rare topic for discussing. Traditional software of cyber security needs a lot of manual effort for detecting the threats, dig up the characteristics from the risk, and then encrypt through the security software for detecting the threat or virus. This manual and the time-consuming process will be handled efficiently through the use of machine learning techniques or algorithms. Many researchers invest and introduce many techniques in machine learning for controlling cybercrime more accurately and reliably.

Noor et al. (2019) described the future of social media's recent attacks, and applications of the new technologies and challenges with the prime help of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) in the field of networking or IT. IoT, AI will boost the CA and server fewer applications will be increased under the protection of data which is saver on personal electronics gadgets related to IT Technology. Unfortunately, there is seems to be no end to hackers who want to access personal and unauthorized information and the organizations and utilities will be the victim of cyber-attacks so, must be governed and financial systems take a hard decision to control these cybercrimes.

As indicated through the Reference eBook Britannica, "the design commercial enterprise is a multibillion-dollar worldwide assignment committed to the problem of creating and selling garments. Some eyewitnesses apprehend the design enterprise (which makes 'high fashion') and the attire commercial enterprise (which makes common garments or 'mass fashion'), but by way of the 1970s, the limits among them had obscured. The layout is exceptionally characterized essentially because the style or forms of getting dressed and frill worn at some random time by way of gatherings of people." (Steel and most important). It’s miles apparent these days that the style commercial enterprise assumes an enormous component within the public eye. People use style as an average of articulation and a wellspring of self-character. There may additionally seem like a generous hole between the steeply-priced, restrictive high fashion of New York and Paris and the mass-created easygoing wear offered in shops of the suburbs even though, all degrees of the style enterprise fuse similar thoughts of "plan, fabricating, flow, showcasing, retailing, publicizing, and advancement of an extensive range of apparel (guys, ladies, and children's) from the thinnest and steeply-priced excessive fashion (in a real experience, 'high stitching') and fashioner designs to normal garb—from couture ball outfits to Succulent Couture-logo going for walks pants" (steel and predominant).

Wintour is contending that an extra profound appreciation for style includes a comprehension of its patterns, impacts, messages, and history-comparable as compelling artwork and maybe this more noteworthy importance is the thing that continues purchasers spellbound. Likewise, just like expressive arts, it cannot be rejected that high layout is an extravagance. For the reason that the widespread majority cannot stand to revel in, the longing to stumble upon the "appearance" of high layout is moreover heightened. Due to shops like H&M, Everlastingly 21, and even objective, even the maximum cost cognizant client can situate themselves as conscious innovators. In this manner, the intellectual elements of the enterprise no longer solely can legitimize the first elegance, the pinnacle of the line customers, yet similarly the deal trackers. Assorted to before, today's creator’s work in a global commercial middle, constantly extended utilizing television, film, and mainstream society, however extensively in addition accelerated by using the web and on-line media. Now not, at this point restricted to the privileged, pioneers come from one-of-a-kind social corporations, and maximum patterns have quick instructions.

Jennifer (2019) observed that the cerebrum cherishes new things, in particular, drifts due to the fact they may be rapid glimmers of interest. Comparable as repainting a room, or shopping for new fixtures, the acquisition of every other, famous thing considers reexamination, and the purchase of a wearable elegant thing takes under consideration self-reevaluation (Baumgartner, 2018). Likewise, as expressed previously, fashion is continually being suffering from its preferred weather be chronicled occasions, mechanical headways, or converting social jobs, and sporting this layout is a strategy for articulation. "All get dressed selections are inside propelled and may be investigated to discover the internal identification. Patterns pursuing is often spurred from a craving to healthy in, sense contemporary, and veils uncertainties. These continually changing design directs never think about a sizeable id of your look, your inclination, and at closing the message you want to develop on this planet". Typically, the manner wherein one decides to get dressed is an impact of their maximum genuine self.

Comparable to fashion, web-primarily based media is any other outlet to articulate one's thoughts. Be that as it could, this method for articulation takes place inside the net. Numerous fundamental strings join the 2 together. For example, the intellectual attitude to introduce oneself with a selected intention in thoughts exists inside the two retailers. individuals get dressed in a selected technique to introduce themselves to the world around there, almost equivalent to the manner wherein customers might also mastermind their Fb page, Twitter record, or Instagram. Furthermore, the craving to healthy in and be incorporated gives itself through design and internet-based total media. Customers of fashion need to be a chunk of the maximum latest sample and get awards from friends, similar as customers of internet-primarily based media want their "partners" or "adherents" to like and comment on their posts. At remaining, speaking to most of the people is also something the 2 retailers preserve in like manner. These days, there are extra than 1 billion clients of web-based media and minority holes are essentially non-existent. Shockingly, no longer utilizing online media now frequently compares to a beneficiary of elitism, no longer the failure to get to it, as before. “Stylish human beings find out it too standard and others find out their protection arrangements irksome. All in all, not utilizing net-based media is probably a result of more education, no longer an absence of getting entry to" (Ferenstein, 2018).

Concerning layout utilization and online media in Oxford, Mississippi, it seems as even though Oxford may perhaps be an anomaly. Demographically speak me, Oxford is a touch, southern, faculty metropolis; thusly it doesn't nicely deal with the more prominent US so any ends that came to may not be suitable to the complete layout retail variety. Of the surmised 16,000 understudies, 55% are female and nearly 37% of the understudy frame is out-of-express. A big level of the college is Greek associated, and as expressed inside the presentation, understudies are very look arranged; numerous with plain tiers of more money. It seems to be like this profoundly public manner of lifestyles, blended with youngsters based through on-line media with the capability to keep, could sincerely impact close by retailer's selling techniques. Accentuating a presence thru on-line media could be worthwhile.

Abiodun et al., (2018) described that ANN is another model of machine learning (ML) the development of Artificial Neural Network is inspired by the biological neuron system to simulate the structure and base of functionally like a human brain. As per the name of Artificial Neural Network is a combination of three different words first is Artificial which is defined as it a presenting of a real objector some time called human-made and its function is close to that original object. Second is Neural, neural is an adjective of neurons this was originally taken from the brain of human in our brain there are billions of cells which called neuron and fundamentally work like biological neurons many researchers and developer are used many other alternative words like connection base network, parallel distributed processing network, etc.

### Prediction Research with Machine Learning

Kushwaha et al., (2020) Nowadays, machine learning is becoming an innovative area of research with its application for analysis and prediction purposes. Machine learning is an approach to analyzing data for updating computers. Appropriate selection of actions is a must for measuring the accuracy of the system. Machine learning techniques help to understand and recognize the situation of risk. Machine learning is also playing an important role in the development of appropriate cures to control this music classification. Lötsch et al., (2018) Supervised machine learning means when algorithms are trained to learn which is the accurate assignment from the following parameters for an accurate decision from scientific and storages data, its success is supervised as the information about the accurate detection is existing. Then, the expert algorithm will certainly complete the accurate assignment. Halgamuge, (2020) Classification algorithm is mostly used by practical applications for prediction and forecasting purposes. Supervised ML proceeds a well-known set of input variables, the known comebacks to the statistics or output variable, and a process that learns the mapping task or trains a model after the input to the output variables. All algorithms and labelled data are used to understand the process of predicting output from the given data.

Artificial Neural Network is the processing based on an algorithm that can be build complex patterns to predict the problems or provide the solution of that problem. The similarity between Artificial Neural Network and Biological Brain or Neuron System, to know about the functionality of ANN must be needed to know how BNS work because the idea of ANN techniques is originated from Biological Brain/Neuron. Many problems which we solved on ANN but Also can have solved previous or alternative methods so why we move on ANN, basically ANN is very suitable and efficient for those problem solvers who want to get significant advantages such as cost, ease of debugging/maintenance, accuracy, time and many more. There is another advantage of ANN to solve the problem by using the lookup table approach. Fundamentally lookup table is used to store all the information for gaining the appropriate result and reference of upcoming events and through lookup table approach to generalize the data and in this generalization, ANN will be trained to provide an appropriate solution to the required problem and ANN train through many inputs according to the problem which we are facing after the training section solution maybe not satisfied if the given query is not matched with training section. Another major advantage of ANN is the memory distributed for large problem or component which is used within the network.

Muhammad et al., (2020) Years before, ML has resolved numerous complicated and tough real-world issues in the application regions like marketing, business applications and retails applications, natural language processing, health care, autonomous vehicle system, intellectual robots, image processing software, and gaming among others. ML techniques had been used for the estimate and decision of numerous. Supervised ML programs are classically based on the trial-and-error method which is exactly reverse to conventional programs that follow the programming order based on if or else decision. Hyde et al., (2019) A Support Vector Machine is a process for supervised machine learning that turns an optimal hyper plane in a dimensional space to accurately classify the expected result by the independent variables in the dataset. A Support Vector Machine is an extreme boundary classifier that means it exploits the separation among specific classes of data efficiently in a high dimensional area. SVMs are exclusively useful when the limit between sets is non-linear as points can be transformed to a space in which the limit is linear. For this feature, SVMs are mostly used in classification issues in which the distinction between groups is non-linear. The fast development in IT over the most recent twenty years has prompted development in the measure of data accessible on the World Wide Web. Another style for trading and sharing data is web-based media. Online media alludes to the methods for communication among individuals wherein they make, offer, and trade data and thoughts in virtual networks and organizations. In today's era, social media users have crossed billions or even trillions in numbers, this means that every single user is creating a bulk amount of data from its account (Badieh Habib Morgan and van Keulen, 2014). Online media has gotten one of the significant methods for correspondence and substance creation. Thus, modern frameworks that can deal with rich client-created content from the web-based media stage have a few genuine applications. Besides, because of the substance style, size, and heterogeneity of data (e.g., text, emoji’s, hashtags, and so on) accessible via web-based media, novel NLP procedures and frameworks that are planned explicitly for such substance and can conceivably incorporate or take in data from various sources are profoundly valuable and appropriate (Bhargava et al., 2017).

ANN is applied in various applications such as Social media Analytics, the medical field, image processing, pattern recognition, speech recognition, radar recognition, air traffic system, power supply system, cybersecurity, and many more. In the 1800s researchers builds an idea that is possible to make a technique or model that works like the human brain like thinking, self-learning, act according to behaviour and the ability of many other functions. A simple and basic ANN was developed between the 1960s and 1970s but fall due to a lack of accuracy and efficiency after 1986 to improve the enhancement of ANN. The proposed system of this research which is used it consists of four layers: the first one is the sensor layer, the second layer is the object layer, the third phase of layer used the preprocessing layer, and the last layer is used in this proposed system is the application layer. The functionality of these mention layers is describing as sensor layers are also called the data acquisition layer and it consists of n-sensors for the purpose as input variables that are used to sense the environment and send the data to the object layer through a medium which is linked. After collecting data from sensors, it will be saved in the form of raw material then data will be passed in the layer of preprocessing layer where analyze the data, mitigate the noise, normalize the data and make it in proper form for gaining the maximum accurate result because the collecting data may contain noise, irrelevant or missing information. After data arrange it will be exceeding into the next layer of the application layer. Where the application layer is divided into two main layers: the prediction layer and the performance evaluation layer.

The prediction layer is further divided into three layers of the input layer, hidden layer, and output layer and these three layers are perform on the base of Artificial Neural Network (ANN). ANN is a model whose function provides computational results with desire accuracy and efficiency. Three major qualities of ANN make it better than other computational methods and these qualities or advantages are ANN use parallel operation which faster than other computational methods those are used serial operation the reason behind this logic is that the parallelism methods have a high degree of sharing information or i/o signals because of distributed memory. Second is ANN learn from Data and store this information into the lookup table and third it uses nonlinear processing function to manage the complex nonlinear problems. ANN has very excellent properties such as self-learning, progressively in input to an output arranging, adaptively, and nonlinear problem that’s the reason ANN is mostly used for universal calculation in numerical standards.

Hareesh Bahuleyan (2018) utilized the machine learning procedures to classifications music consequently through giving labels to the tunes blessing in the client's library. It investigates each Neural Organization and ordinary methodology of utilizing machine learning calculations and to achieve their aim. The main strategy utilizes Convolutional Neural Organization that is gifted offer up to stop the utilization of the highlights of Spectrograms (pics) of the sound sign. The 2d method utilizes various machine learning calculations like Strategic Relapse, Irregular forest, etc, where it utilizes hand-tailored capacities from time region and recurrence area of the sound sign. The physically extricated capacities like Mel Recurrence Cepstral Coefficients (MFCC), Chroma Highlights, Phantom Centroid, and so forth are 17 utilized to arrange the music into its kinds the utilization of ML calculations like Strategic Relapse, Arbitrary Woodland, Angle Boosting (XGB), and Support Vector Machines (SVM). By looking at the two techniques in a steady progression they arrived at a resolution that VGG-sixteen CNN rendition gave most elevated precision. By building a gathering classifier of VGG-16 CNN and XGB the enhanced adaptation with zero.894 exactness was accomplished.

Bitter et al. (2017) describe the importance of security not only in the pure IT department but also in those areas involved with the internet or computer. They defined that when we talk or concern with cyber security, we must accomplish three major qualities and these qualities are availability, confidentially and integrity and these qualities are satisfying with regulation, policies, and authorities. ANN successfully applied on a broad range of problems such as medical areas (heart, HIV aids, hepatitis, diabetes, etc.), finance (credit card, transactions, etc.), engineering (machine monitoring, robot, automobiles, etc.), Science (biology, chemistry, physics, nuclear physics, etc.).

Abiodun et al. (2018) describe the survey paper about the application of ANN which is used in real-world scenarios. These coming days ANN very popular and useful model for different disciplines like prediction, clustering, classification, pattern recognition, etc. This review study author furthermore presents the ANN application challenges, contribution, compare performance, evaluation methods.

Lu L. Et al., (2002) provided their view of division and classification of sound substance examination. Here a sound move is fragmented with regards to a sound kind or speaker character. Their strategy is to assemble a tough form that can be characterizing and sectioning the given sound sign into discourse, music, environmental factors sound, and quietness. This class is prepared in chief advances, which has made it reasonable for different bundles as pleasantly. The initial step is discourse and non-discourse segregation. Here, a particular arrangement of rules which depends on KNN (K-closest neighbour) and direct unearthly matches vector quantization (LSP-VQ) is been advanced. The subsequent advance is to separate the non-discourse class into music, natural sounds, and quiet with a standard-based thoroughly class technique. Here they've utilized not many uncommon and new capacities comprising of commotion body proportion, band periodicity which is conveyed, yet referenced in the component. They have moreover covered and progressed a speaker division calculation. This is solo. It utilizes a novel plan dependent on semi - GMM and LSP relationship examination. With no earlier comprehension of something, the model can help the open-set speaker, online speaker displaying, and the constant division.

Lek (2018) described that ANN is a powerful tool for prediction or solution of any problems and it has a very strong function and different types of algorithms or models also it shows universal and flexibility for any data. Prediction of modelling ANN is having a well-known approach use name ‘black box’ and this word black box means that all characters have an unknown situation, so these characters are identifying through training of ANN.

Potluri et al., (2017) described that Artificial Neural Network has a lot of advantages including requiring less formal statistical training to ANN for predict or finding the solution of a particular problem, ANN can point out all possible relations between predictor variables, there is a variety of algorithm and model are used in it for problem-solving and these algorithms and software/model and packages are easily available in the simple range, user-friendly, mostly graphical user interface (GUI), detect the non-linear relationship between dependent and independent variables, multiple algorithms are used to raining of an ANN, no need to know deep knowledge for developing/training of ANN just know about the basic structure of ANN model and data or parameters that can be adjusted with the model.

Dreossi et al. (2019) describe real-time data mining in intrusion detection systems (IDSs) and they have the most attention on this problem. Authors have mentioned that three types of issues that face in IDSs are accuracy, efficiency, and usability. The first problem is accuracy and it’s difficult to define how intrusion detects or performance or accuracy because naturally the data mining and IDSs have different definitions. Accuracy improves by analyzing the data and extracts data by using a data mining program and artificial anomalies make more improvement to detect the misuse or irregularity detection model by applying different algorithms and techniques. For improving efficiency in this paper use cost analysis and multiple models approach through this methodology proposed low cost and high efficiency they also present a distributed architecture with real-time for evaluating cost-sensitive. The third issue mentions in this paper was usability because IDSs which are based on data mining are more complex compare to a traditional system and this issue improving by an adaptive learning algorithm to develop the model construction and increasing the update for this purpose used unsupervised anomaly detection algorithm. In this research paper, they discuss and present an architecture which is consisting of detectors, model generation, sensors, data warehouse, and implementing this architecture improves the efficiency and scalability of the IDSs.

Tarik & Singh (2016) describe that how data mining is improving the intrusion detection in the field of cyber security because in present days the internet is more open and everyone is hanging on the internet for saving the information that’s why the security risks have been increased in the area of cybercrimes. In 2005 the pentagon publishes a paper which mentions that over 79000 efforts to intrusion and 1300 are successful ones so watching all these circumferences need to introduce new intrusion detection systems (IDSs) because older techniques are not enough for controlling the attack on cyber life. In this paper, the authors present the join of classifiers use with feature selection and multi-boosting techniques based on data mining for intrusion detection.

CNN's deep architecture has been effective in providing fine and excellent efficiency to trained models by learning patterns through raw images. A proposed deep neural network architecture is called Inception which attains the ILSVRC2014 (Image Net Large-Scale Visual Recognition Challenge 2014) assessment and the new architecture for identification. The basic goal of the design is to make better use of computing resources within the network. Through careful design, the depth and breadth of the network are increased when maintaining a computational and financial plan (Szegedy et al., 2015).

Deep learning has supported the purpose of Computer Vision in recognizing and classifying images, and it is an important tool for automating tasks in daily lives. Object identification, classification, and segmentation have been developed using convolutional networks. Because of its ability to learn to represent data, the use of the convolutional neural network (CNN) on social media images has greatly helped social media analytics (Lecun et al., 2015).

The (SML) Supervised Machine Learning is the exploration toward algorithms that cause after outwardly delivered cases to make overall assumptions, which then make calculations about upcoming cases. Most intelligent systems frequently use supervised classification. Algorithms like Linear Classifiers, (LR) Logistical Regression, Perceptron, NB Classifier, Support Vector Machine; Quadratic Classifiers, Boosting, Random Forest; and Neural networks algorithm are used where supervised machine learning deals with more classification (Osisanwo et al., 2017). Supervised Machine Learning methods need human beings to give essential inputs and outputs correspondingly, moreover, it gives estimation about the correctness of the calculation in the training procedure. Using the decision tree technique, the supervised ML algorithm is used to develop a prediction model using a labelled dataset to calculate social media analysis by the number of user responses (Muhammad et al., 2020).

The Support Vector Machines executes classification by creating an N-dimensional hyperplane that split up the data as classes. Concerning SVM, the analyst variable is called an element and the converted element is called a feature. Using the selection feature, the best and appropriate symbolic data is selected. Recounting a case in a set of features is known as a vector. Discovering the best hyperplane to split the clusters is one of the final objectives of this modelling. At the one side of the plane, there is a target variable whereas on the other side of the plane there is another group. Support vectors are the vectors that are near the hyperplane. The modules of support vector machines are nearly similar to classical multilayer perceptron neural networks (Oladipupo, 2016).

# CHAPTER 3

## Project Design and Methodology

Research Methodology is a method or computational techniques which are used for different purpose like process (observation, simulation, derived, experiment), data gathering, and analysis about a specific topic or problem.

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 decisions that make 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.

The present study is carried out in the domain of artificial intelligence. Artificial intelligence has done remarkable contributions in almost every field by sensitizing the course of accuracy and efficiency. The main purpose of this study is to render out the finding relevant information from Reddit is relatively easy. For this purpose, the researcher has gone through a thorough literature review and has identified various machine learning algorithms along with their strength and weakness. After getting acquainted with the decision support systems, the research approached the different domain specialists and concerned lab technicians for the identification of relevant variables in the light of our living habits as well as the notion of the elements or variables that become a cause of disease. Even some conversations are held.

In research methodology, at the first stage, we determined the factors which can impact prediction. If these factors were not relevant then we discarded them. Other relevant factors were selected for the next step. The evolution of new technologies for organizing and analyzing data with efficient computing has increased our ability to precisely predict any incident and Machine Learning (ML) can be used to make more accurate and reliable predictions, as well as to help understand the underlying pattern of the results. Machine learning includes prediction and classification of data and for this purpose, we use numerous machine learning techniques consistent with the given dataset.

There are many possible techniques and paradigms of machine used. Classification algorithms like Multi-layer Perceptron, Decision trees, SVM, or Regression algorithms like Logistic Regression. The Support Vector Machine is the best option for classification problems because it can resolve linear and non-linear complications and also working fine for numerous real-world problems. Important principals of SVM are Support Vectors, margin, and hyperplane. Support vectors are the nearest data points to a hyperplane. Whereas, a hyperplane is the kind of decision that is divided by a different set of objects and these objects are classified. The Margins are the gaps in-between the two lines on the nearest data points of classified objects. The Margins are said to be as good as large and considered bad as small. The SVM model divides the dataset into classes to figure out the extreme marginal hyperplane which can be done by creating hyperplanes iteratively which set aside the classes in a good way. After creating a hyperplane, it will choose the hyperplane that splits the classes in a correct sequence.

Different Sensors from the environment taking constantly conservation data. It’s affecting a physical measure in quantity. With the board of sensors in the shape of topology, many types of data are collected. Every sensing node gets a subsection of the gathered to nearby compressing & brief since the unsystematic sign.

## Methodology for this Research

This project aims to create a Machine Learning algorithm that can identify and classify between various niches the posts on Reddit belong to with a focus on Science/Technology niche. The most crucial part of this project is the dataset selection. The dataset decides the Machine Learning paradigm to be followed for the research which in turn helps decide the algorithms to be used for training and prediction/classification. Upon observing the dataset selected for this research, it can be deduced that the dataset requires multi-class classification to be employed for maximum efficiency. Since the dataset also contains categorical data i.e., text data, Natural Language Processing techniques will be employed to the dataset to make it usable and digestible for the Machine Learning classification algorithms. The following techniques are the finalized techniques from Machine Learning and Natural Language Processing for this research:

### Neural Networks

The inspiration for this algorithm arises from the biology of the human nervous system i.e., the network of neurons and their working inside a human brain. Initial implementations of this idea resulted in the most basic form of the “neural network” known as the feedforward neural network. The components of this neural network are multiple layers with different responsibilities that consists of nodes or neurons that are interconnected between layers. The responsibilities of the layers classify them into three classes’ i.e., **Input Layer**, which is responsible for taking in the data and distributing it to the next layers for processing by assigning them weights for differentiation. **Hidden Layer**, which is responsible for the identification of patterns in the dataset that lead to the classification of data into requisite classes. **Output Layer**, which provides the output and the number of neurons in the output layer = number of output classes in the dataset.

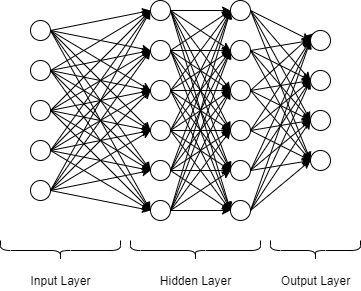
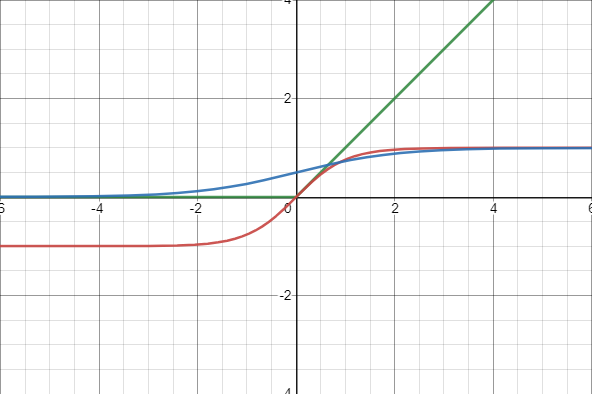


Figure 5 - Basic Architecture of a Neural Network

The working of a feedforward neural network is as follows; the input layer as mentioned before is responsible for taking in the data. This data is taken in the form of a vector. This data is then passed on forward to the next layer using weighted connections between the two layers. The responsibility of each neuron in the hidden layers is to calculate the weighted sum of the inputs received from all the neurons in the previous layer. The mathematical expression for this process is as follows:

An activation function is implemented in each of the neurons. This helps the neurons to express non-linear functions.



Tanh

Sigmoid

ReLU

Figure 6 - Popular Activation Functions

Figure 2 graphically demonstrates three of the most popular activation functions for neural networks. The red line shows the hyperbolic tangent, the blue line shows the infamous sigmoid function and the red line shows the Rectified Linear Unit (Glorot et al., 2010).

The ReLU was introduced by (Glorot et al., 2010) and it shed light on the issues of training Deep Neural Networks with previously popular activation functions like the sigmoid and tan (h). The issue with the hyperbolic tangent was that it was saturated by both high and low neuron outputs that resulted in smaller gradients. This resulted in a numerical underflow that discontinued backpropagation and hindered the training process. This is shown in the graph in Figure 1 where the red line representing the hyperbolic tan function has values both above and below zero while ReLU solves this problem by only activating neurons with values above zero. However, this working method of ReLU causes an issue again. Once a neuron is shut down, reactivating it is impossible for most optimization algorithms. To solve this issue, an asymptotic approach is taken with neurons below zero i.e., very small gradient values are assigned to them (Xu et al., 2015).

The output layer of a Deep Neural Network is referred to as logits which are un-normalized log probabilities represented by Z. To normalize these probabilities, a SoftMax function is used to create a normal probability distribution.

Training a Deep Neural Network refers to an optimization problem that focuses on the optimization of a function that is used as a performance measure for the Neural Network. But this performance measure is intractable in a Neural Network architecture. Therefore, a proxy function known as a cost function is used to determine the performance of the Neural Network. This cost function, denoted by J denotes how poorly the Neural Network performs. The problem assumes that in minimizing the cost function J, the intractable performance measure will increase. The cost functions for each Deep Neural Network are optimized according to the dataset they are meant to train on. The most common cost function used when it comes to multiclass classification is the cross-entropy loss defined as:

Where θ are the hyper-parameters of a Deep Neural Network. F(x) is the output of a Deep Neural Network when an input vector x is provided, y is the true classification of the input vector and F (y | x) is the score of confidence given to the correct output class by the algorithm.

The layers in the neural network each are represented by separate functions. Chaining the functions representing each layer can result in an expression that represents the complete Deep Neural Network. This characteristic of the Neural Network is used to calculate the gradient of the cost function concerning the hyper-parameters of the Neural Network. The algorithm preferred for computing these gradients is the backpropagation algorithm.

Training DNNs usually involves tuning the values of DNN parameters with a combination of gradient-based methods and backpropagation. Using calculated by backpropagation, DNN parameters can be updated to reduce the cost using gradient-descent based methods. Various such methods have been devised, but the most notable are stochastic gradient descent (SGD) (Bottou et al., 2010); root means squared prop (RMSProp) (Tieleman et al., 2012) and adaptive momentum estimation (Adam) (Kingma et al., 2014).

### Support Vector Machines

Vladmir Vapnik along with Alexey Chervonenkis created the infamous Vladmir-Chervonenkis theory for statistical learning from distribution-free data. This computational learning theorem is the seed of birth for the Support Vector Machine (Ukil, 2007). Using this work, Vapnir also in the same year introduced the idea of a linear classifier that separates data classes using hyperplanes optimized for that dataset (Ukil, 2007). However, the first wave did not catch on. The 1990s proved to be fruitful for the development of SVM when Bernhard Boser, Isabelle Guyon and Vladimir Vapnir introduced a method that provided a route to create non-linear classifiers by applying the concept of a kernel function to maximum-margin hyperplanes (Boser et al., 1992) (Ukil, 2007). The kernel function is an addition for the linear classifier that helps map the original observations on a higher dimensional plane to accommodate for non-linear behaviour.

It is one of the classical machine learning technique that can still assist in solving big data problems (Suthaharan, 2016). Support vector machines are algorithms that learn to classify data based on example (Noble, 2006). Its functionality is basic terms can be explained as the maximization of a mathematical function for a particular dataset (Noble, 2006). The SVM mathematical model can be explained with the following simple equation:

This equation represents the linear version of the Support vector machine algorithm (Suthaharan, 2016). The support vector machine algorithm can be used for both linear and non-linear applications (Suthaharan, 2016). The steps required for both the applications are the same in concept however differing in implementation. The steps required for linear SVM are:

* Mapping of the dataset into a response set.
* Division of the dataset.

The steps required for non-linear applications are as follows:

* Mapping of the dataset into a feature space by employing a kernel function.
* Mapping of the feature set into a response set.
* Division of the data domain.

(Suthaharan, 2016) explains this phenomenon mathematically in his study of Support vector machines as:

The application of the kernel function in the linear equation turns the SVM into a non-linear SVM (Suthaharan, 2016). The working of the SVM revolves around two mathematical concepts i.e., parametrization and optimization of the objectives obtained from the topology of the dataset domain.

Since this research employs multiclass classification, the linear and non-linear classifier version of the SVM will be discussed accordingly. The multiclass classification using a linear SVM is achieved by employing an ensemble of binary classifier SVMs. Considering a dataset, a binary classifier SVM divides the dataset into two different classes represented by the following equations (Suthaharan, 2016):

The parametrization objective can be defined based on the above two equations as follows:

These two parametrization equations define the straight lines also referred to as hyperplanes that define the boundaries between the binary classes in the dataset. The optimization objective can now be defined using the parametrization equations as the distance between the two boundaries of the binary classes of the dataset (Suthaharan, 2016). The assumption made here is that the hyperplanes are parallel to each other and the optimization objective is to maximize the distance between the two hyperplanes (Suthaharan, 2016). The maximization occurs on the hyperparameters of the SVM defined in the equations of the parametrization objectives (Suthaharan, 2016). The assumption that the hyperplanes are parallel to each other allows the definition of the optimization objective equation using the standard equation for the distance between two parallel lines given by (Suthaharan, 2016):

The resulting optimization equation is given by (Suthaharan, 2016) as; where the slopes of both the hyperplanes are equal to w and their intercepts b1 = γ+1 and b2 = γ-1. Equating these variables into the distance equation results in; (Suthaharan, 2016). However, in rea-word applications of the SVM, the equation is written using the mathematical norm form which results in (Suthaharan, 2016). This equation states that instead of the maximization of the distance term, the minimization of the error term w can be done i.e., the minimization of the error term w results in the maximization of the distance term. Therefore, the error function for SVM is as follows (Suthaharan, 2016):

This error function is critical in designing the optimization objectives for the SVM. The following figure provides a deeper insight into the working of a binary SVM.



Figure 7 - Binary SVM Hyperplane Graph

To explain clearly the working of a non-linear SVM, the perception of a linear graph must be boosted with additional parameters to increase the dimensions of the graph. Increasing the dimensions of a linear equation turns it into a quadratic equation (Vapnik, 1995). The hyperplane will then be represented by the following equation (Schölkopf and Smola, 2001):

Where n represents the number of training samples, yi represents the label value for each data instance, Ψ represents the kernel function responsible for mapping the linear objective to higher dimensions and α is the hyper-parameter that needs to be calibrated in a way that maximizes a Lagrangian representation (Ukil, 2007). A Lagrangian is a function of the dynamic variables of the system (Ukil, 2007). Subject to the constraints αi ≥ 0 and there is a value for αi for each training point and training points that are closer to the hyperplane have non-zero values for αi. These non-zero values for αi are called support-vectors (Ukil, 2007).

However, for real-world implementations, things like noisy data and other real-world biases need to be accounted for. To account for these, the equation for the hyperplanes is deflated by constraining the training points that are misclassified (Ukil, 2007). The equation now becomes; where provides the relaxation required to measure the misclassification from the constraints. To adapt to this, the optimization objective changes to become (Ukil, 2007):

Here, µ represents a regularization parameter responsible for governing the balance between maximizing the margin and maintaining the training error (Burges 1998). This machine that now utilizes the Support Vectors is referred to as the Support Vector Machine with two hyper-parameters (Burges 1998):

* The kernel function.
* The regularization parameter µ.

### Naïve Bayes Classifier

Naïve Bayes classifier is a simple learning algorithm that utilizes the infamous Bayes’ theorem from probability with an assumption that the features of an event are conditionally independent given the output class (Webb, 2016). This algorithm provides a technique for estimating the probability of an output class appearing given the feature vector of the dataset i.e., P (y | x). These estimates can then be further used for classification purposes (Webb, 2016).

To understand the functioning of the Naïve Bayes classifier, the understanding of the basic principles of the classifier itself is mandatory. A classifier is defined in (Murphy, 2006) as a function that maps the input vector onto the output vector. The input vector belongs to the feature space. It is normally assumed that the feature space is itself a subset of Real Numbers but the real world can provide datasets with features of any type whether discrete or continuous (Murphy, 2006). Another assumption made while using classifiers is that the data is unordered (categorical) and mutually exclusive (Murphy, 2006). If an input variable belongs to more than one output class, then the problem is a multiclass classification problem (Murphy, 2006). The goal therefore with classification is to learn the function that maps the input vector onto the output vector.

The input and the output vector come from the dataset this research is working on. The dataset consists of columns that are features that result in output when combined in the real world. These features make up the input vector. The output vector will be the output these features result in. Let *x*, be the input vector and *y*, be the output vector. Therefore, function *f* is the learning objective that requires x and y from **a labelled dataset** such that (xn, yn), n = 1: N; this is also referred to as supervised learning.

The probabilistic classifiers like the Naïve Bayes classifiers all return p (y | x) as the output (Murphy, 2006). Based on the methods to achieve this output, there are two types of probabilistic classification models (Murphy, 2006). The first method to achieve the desired output is to directly learn the function that provides the output p (y | x). This is the **discriminative model** (Murphy, 2006). It is referred to as the discriminative model since it differentiates between different output classes given the input (Murphy, 2006). The next method is to **class-conditional density** p (x | y) for each value of y and to learn **class priors** p (y). The application of the Bayes theorem to these parameters produces:

This is referred to as the generative model (Murphy, 2006). It is named so because of its ability to generate the features vector x for each output class ‘y’. A relatively easy approach to generative and discriminative learning is to not consider probability altogether and initialize a function that directly maps x to ‘y’ i.e., also known as the discriminant equation (Murphy, 2006).

#### Generative Classifiers

With the assumption that the number of classes C is small, the ***class prior*** can be estimated with ease using the function; where π represents the vector of class probabilities. Then, the Maximum Likelihood Estimation will be, where NC is the number of training examples that have class label c.

The issue that stands now is the determination of the class conditional density. The typical assumption that goes into this is that the parameters of each conditional distributions of the same type are independent (Murphy, 2006). This assumption allows the estimation of p (x | y) separately for each value of y. There are three methods of estimating class conditional densities:

1. Gaussian:
2. If the input vector x belongs to {0, 1} then, Bernoulli:
3. If x belongs to {1, …, K} then Multinomial:

The reason behind adding the “Naïve” in the Naïve Bayes theorem is the assumption that all the features of the data are conditionally independent given the class label (Murphy, 2006):

This equation is false since the features are usually dependent and there is a high degree of correlation in the dataset, but this model is easy to train and works well (Murphy, 2006).

### Count Vectorizer

Count vectorizer is one of the traditional methods for extracting hidden features from the text (Kulkarni and Shivananda, 2019). It is a process of converting categorical variables into features or columns by using 0s and 1s for different output classes i.e., binary (Kulkarni and Shivananda, 2019). The way a count vectorizer works is that for a given word in the dataset, it will provide the frequency of the word in the data as the encoding for the word (Kulkarni and Shivananda, 2019).

### TD-IDF Transformer

Term Frequency-Inverse Document Frequency is one among the foremost popular approaches for term weighting in today’s applications of tongue Processing (Aizawa, 2003). TF-IDF has been since an extended time considered an empirical metric, especially when seen from a probabilistic viewpoint (Aizawa, 2003). By its definition, tf–idf may be a metric that multiplies the 2 quantities tf and idf. Here, tf provides an immediate estimation of the occurrence probability of a term when it's normalized by the entire frequency within the document, or the document collection, counting on the scope of the calculation (Aizawa, 2003). the essential function consistent with this definition is that given a component of textual data like a document or a term, the importance of the component is expressed as a product of the probability that it occurs and therefore the amount of data that it represents (Aizawa, 2003).

TF-IDF (Term Frequency-Inverse Document Frequency) works by determining the frequency of words during a specific data instance/document compared to the inverse proportion of that word over the whole dataset/corpus (Ramos, 2003). This calculation provides the relevance of the word within the document as a ratio (Ramos, 2003). Words that are common during a single or small group of documents tend to possess higher TF-IDF numbers than common words like articles and prepositions (Ramos, 2003).

The mathematical representation of TF-IDF accommodates minor changes with different applications but the general approach works according to the following equation:

Where we, d represents the frequency of the word w in document d. |D| is the size of the complete dataset/corpus and fw, D represents the number of documents that the word w occurs in (Ramos, 2003).

The code is simple for TF-IDF. Given a query composed of a set of words, the WI, d is calculated by keeping a sum record of the fw, d and fw, D. Once the sum is complete, the WI, d can be easily calculated (Ramos, 2003). This can then be used to return a set of documents D\* that maximize the following term (Ramos, 2003):

The equation returns the documents in descending order. This is the traditional process for the implementation of TF-IDF (Ramos, 2003). The starting of TF-IDF can be traced back to information theory (Aizawa, 2003). Assuming a joint probability distribution P (x, y) where x belongs to X and y belongs to Y, then using P (x, y) and marginal distribution, the probability of x =>.

These equations simply convey the fact that in the event x is observed, the presence of y is confirmed in the frame of observation and vice versa (Aizawa, 2003). From information theory, the definition for the amount of information can be taken as Let a, b be two variables that represent distinct events from X, Y. The amount of information expected from these two distinct events, also known as the self-entropy denoted by H(X) and H(Y) can be calculated as (Aizawa, 2003):

This metric known as self-entropy represents the ambiguity in the certainty of the occurrence of the event under consideration (Aizawa, 2003). Self-entropy is directly proportional to the number of events under consideration. Now taking both events into the same frame of observation, the pairwise mutual information criterion denoted by M (a, b) is defined as (Aizawa, 2003):

Considering both the events ‘a’ and b into the same frame of observation, the pairwise mutual information criterion helps accommodate the interaction between both the events and the exchange of information due to the interaction (Aizawa, 2003). Applying the general property of conditional probability to the pairwise mutual information criterion gives the expected mutual information (Aizawa, 2003):

This representation of mutual information between two events can result in the derivation of an information-theoretic view of TF-IDF (Aizawa, 2003). Let us consider a document provided as a set of unordered terms. Let D be a set of documents, and W be the set of distinct terms contained in D. Let N represent the total number of documents and M represent the total number of terms in N documents. Let d denote the event of selecting a random document from D and w denote the event of selecting a random term from W. Let e, f be the random variables initialized by the events d and w. Introducing these random variables initialized from the events d and w, the situation now displays query submission and result in retrieval as probability distributions over D. The objective now, is to find the expected mutual information between e and f to observe how well the documents in D, are specified by the query given. The self-entropy of the system is given by (Aizawa, 2003):

The two assumptions required at this stage are as follows (Aizawa, 2003):

1. All documents in D are potential candidates at the beginning.
2. The subset of D which contain the selected term f is known.

The second assumption means that the documents that do not contain the selected term do not contribute to the system in any way (Aizawa, 2003). Selecting a query term f from inside of d and representing the frequency of the occurrence of e in f using j and the frequency of w in the set D as J, the expected mutual information is calculated as (Aizawa, 2003):

The above equation represents the products of the tf in the form of j. and the idf factor is divided by the constant J. This shows that the information-theoretic view of tf-idf can be interpreted as the quantity required for the calculation of the expected mutual information (Aizawa, 2003).

## Project Workflow

### Problem Understanding

Reddit has been a famous online media site from the early occasions of the web-based media frenzy is even more mainstream than any time in recent memory. A ton of extraordinary and instructive substance advances toward Reddit first and afterward to different spots like educational web journals and arrangement gatherings. Reddit's format of the site is somewhat novel and individuals not used to Reddit however used to different stages like Facebook get somewhat lost during route.

Likewise, since Reddit has been hanging around for quite a while, there's a ton of significant information accessible on the stage for anything you search however it's practically difficult to track down what you precisely need despite the fact that it's accessible because of the sheer measure of information. Many individuals treat Reddit as a discussion and many pages on Reddit are genuinely gatherings where individuals from everywhere the world offer answers for issues for each space. This task can assist with peopling utilize these gatherings significantly more proficiently.

### Data Collection

The dataset used would be the Reddit Flair Dataset available from Kaggle. The dataset contains the Reddit posts of the Indian region. This data was taken out from Reddit with the help of their easy-to-use API. It contains various features such as the 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

### Data Annotation

Since the dataset is pre-made and collected from a renown global repository, the dataset is already annotated into classes mentioned in the previous subsection. Therefore, there will be no requirement of additional data annotation for the dataset selected for this research.

### Data Analysis and Visualization

The project will be evaluated based on for this research are SVM, Multinomial Naïve Bayes, and Multilayer perceptron. 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.

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. 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.

The dataset used would be the Reddit Flair Dataset available from Kaggle. The dataset contains the Reddit posts of the Indian region. This data was taken out from Reddit with the help of their easy-to-use API. It contains various features such as the 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.

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 portrayal utilized in regular language handling and data recovery (IR). In this model, 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 the Random Forest classifier on the dataset thus giving it the probabilistic edge in prediction.

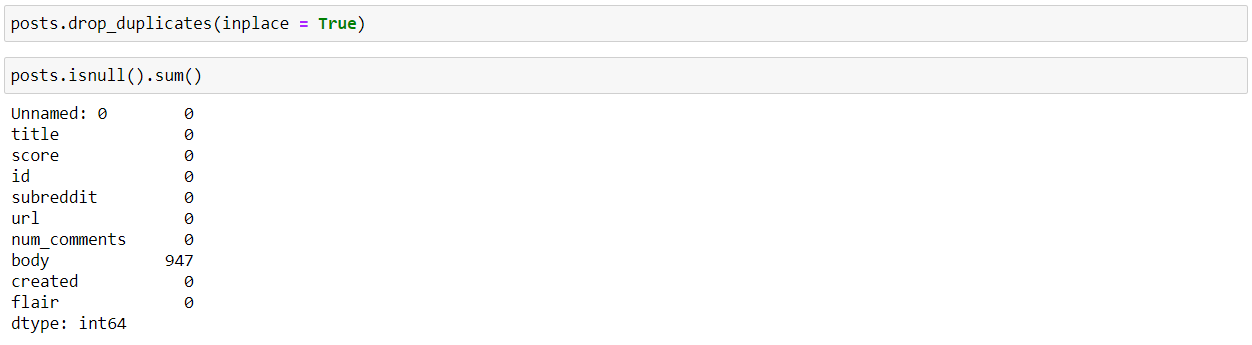
The tools for this research project will be Machine Learning Libraries like Keras, Tensorflow, and scikit-learn that will provide the necessary classification algorithms required for this project. Although random forest classifier is finalized, there are other options like Neural Networks and SVM that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

The planned responsibilities for the code are to perform data-preprocessing, provide deep insights into the data using visual graphs and statistics and develop a machine learning algorithm powered by Natural Language Processing that can identify Reddit posts that belong in the science domain.

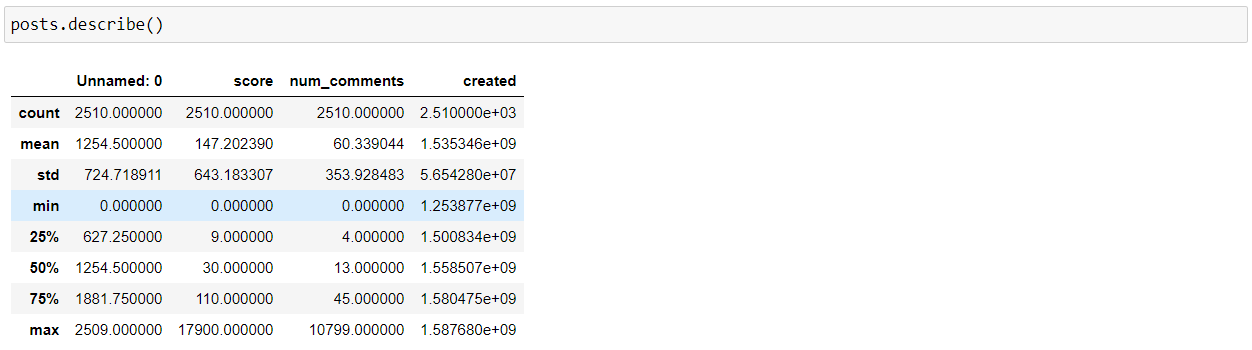
The data pre-processing section in the code handles all the common discrepancies in the dataset like duplicate and null instances in the dataset. It also reshapes the data to perfectly fit the needs of the code. This section of the code also provides analysis on the dataset using graphs and statistics, more specifically central tendency measures.

The features of the dataset are as follows:

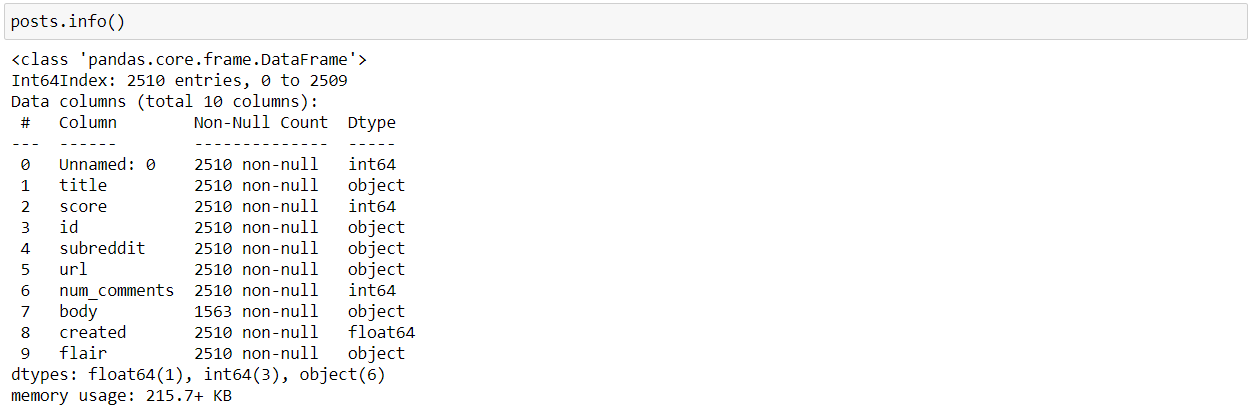
* Title: Title of the Reddit post
* Score: Reddit score for the post (upvotes)
* Id: Reddit post id
* SubReddit: to which subReddit it belongs
* url: url of the Reddit post
* num\_comments: number of comments on the Reddit post
* body: the text content of the Reddit post
* created: date and time at which the post was created
* flair: the output class of the dataset denoting the niche of each Reddit post instance



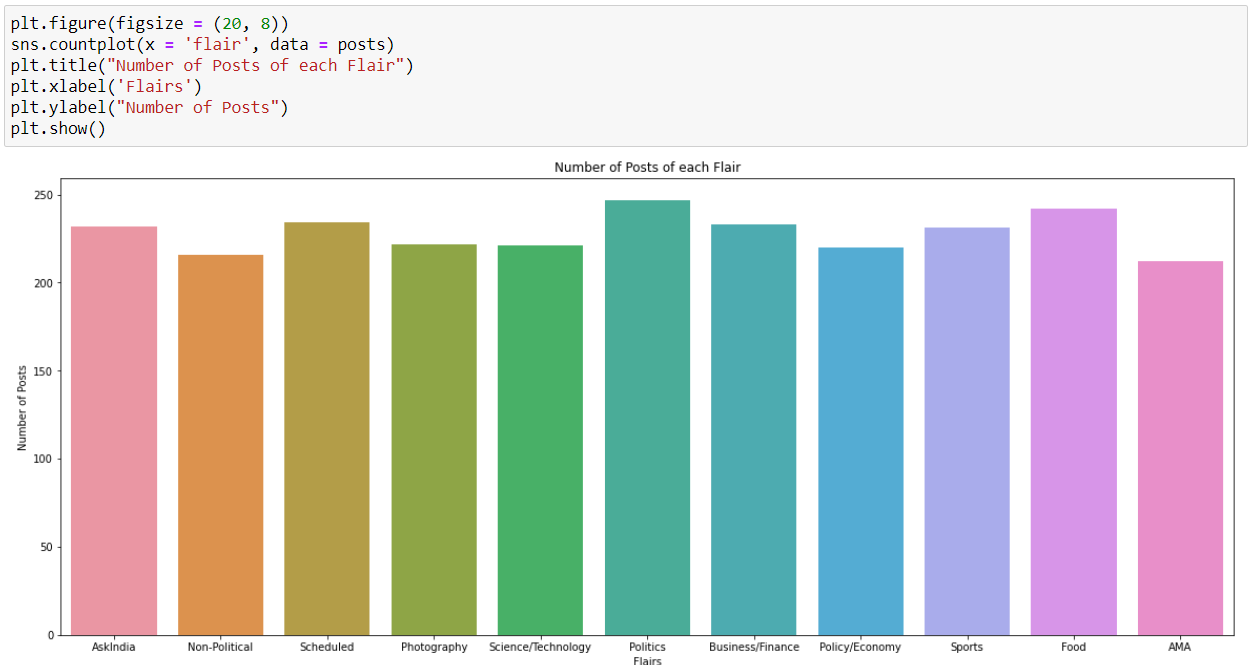
The figure above shows the code used to remove duplicate instances and view the number of null instances for each feature. The dataset is clean and ready for further use.



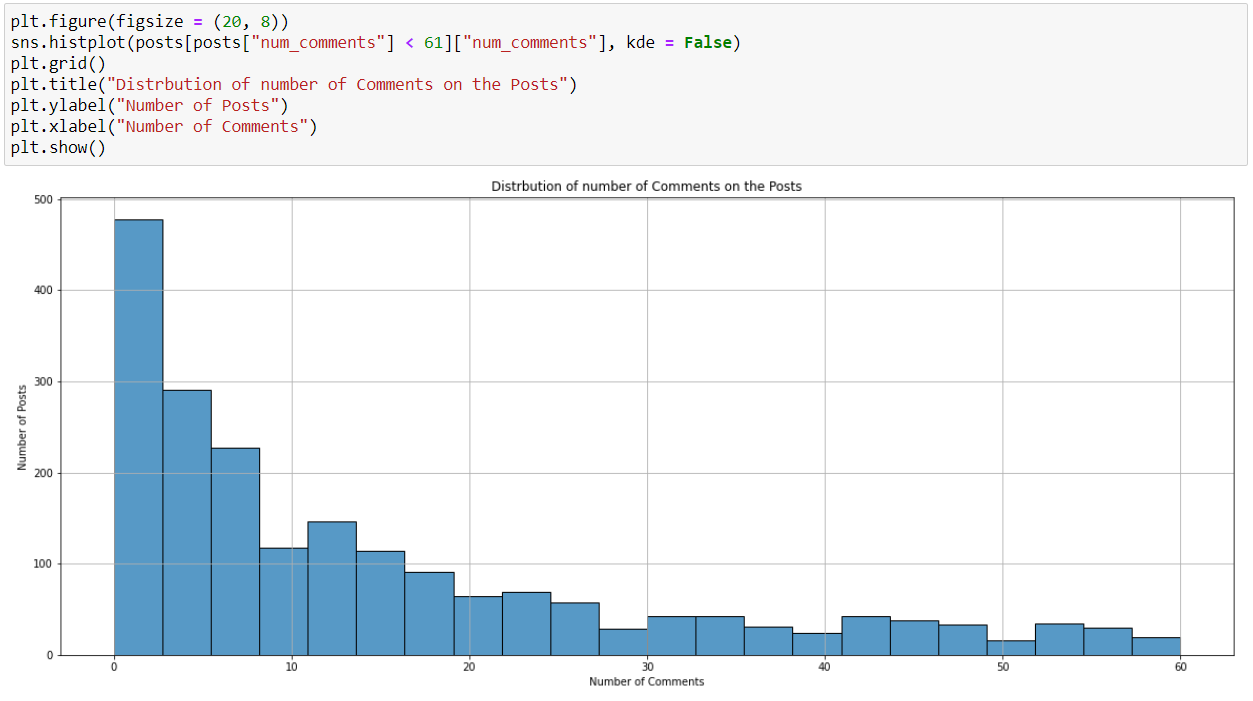
The figure above shows the code that provides a statistical analysis of each feature in the dataset using central tendency measures. These central tendency measures work best on numerical data only as the figure above only shows features with numerical data.



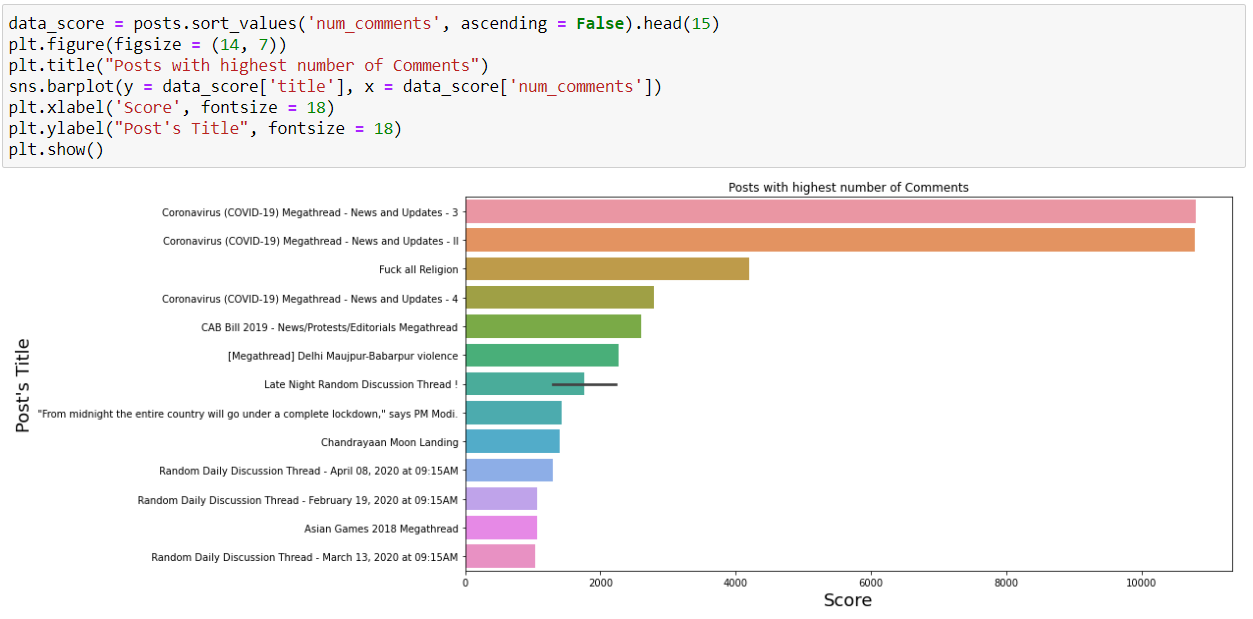
The code in the figure above shows the metadata of the dataset. The metadata contains the number of non-null instances for each column along with their names and data types. This is useful when you need to identify if any of the features in the dataset require any encoding or not since categorical features denoted by the “object” data type require encoding to be useful. The next is the visualization of data to study the trends in it.



The graph above shows the number of Reddit posts for each niche/flair. The following graph shows the distribution trend between the number of posts and the number of comments.

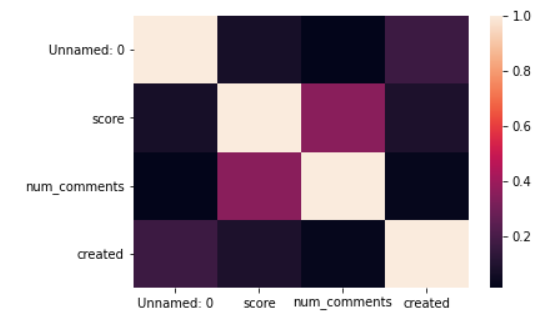


The following graph shows the top posts concerning the number of comments on them.



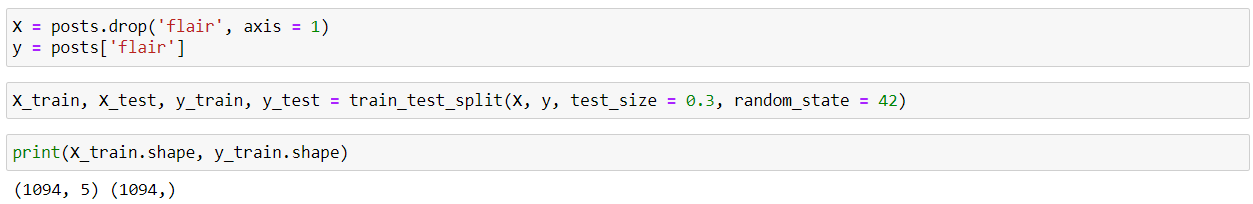
### Data Preprocessing

To better understand how the features in the dataset correlate to each other and contribute to the dataset, a correlation matrix heatmap was plotted.

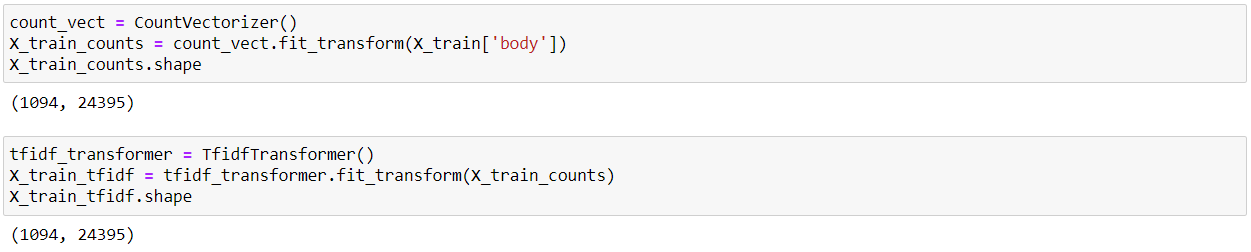


### Model Development, Training and Evaluation

The next step is to process the data for the machine learning implementation. The dataset is split into training and testing datasets giving 30% to the test set and 70% to the training set.



Using Natural Language Processing tools like the Count Vectorizer and the TF-IDF Transformer, features from the text data in the dataset are extracted and processed for further use by the algorithms applied.



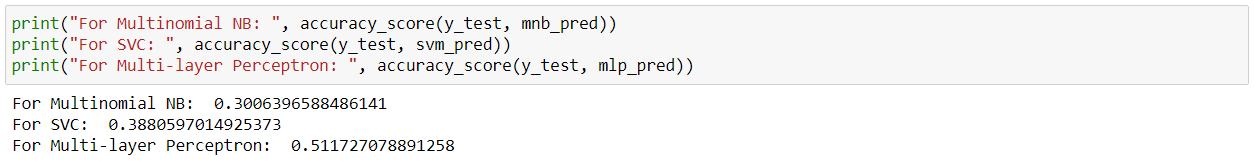
The objective of this code is to develop a machine learning model that can identify if a Reddit post is of the science domain or not. To that extent, three algorithms will be applied and their performances compared to identify the best performing algorithm. The three algorithms chosen for this research are the Multinomial Naïve Bayes, Support Vector Machine, and the Multilayer Perceptron.

# CHAPTER 4

## Results

With the exponential speed that internet technology is developing, 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 focused on the 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 training to classify science-based posts on Reddit.

The results for these algorithms are as follows:



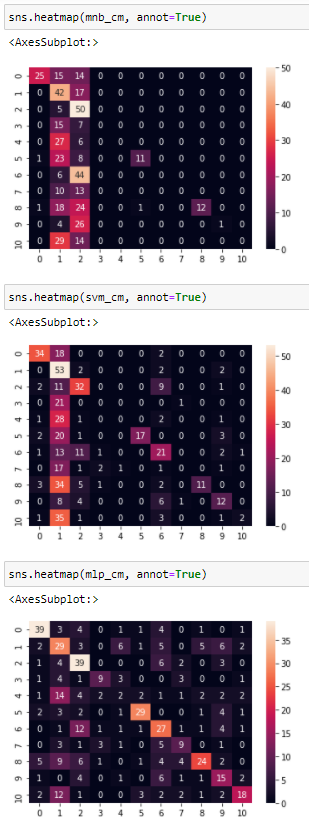
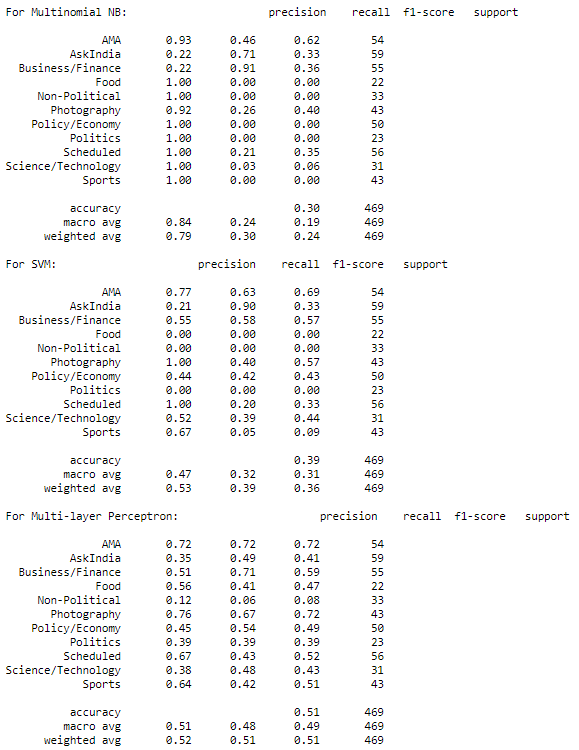
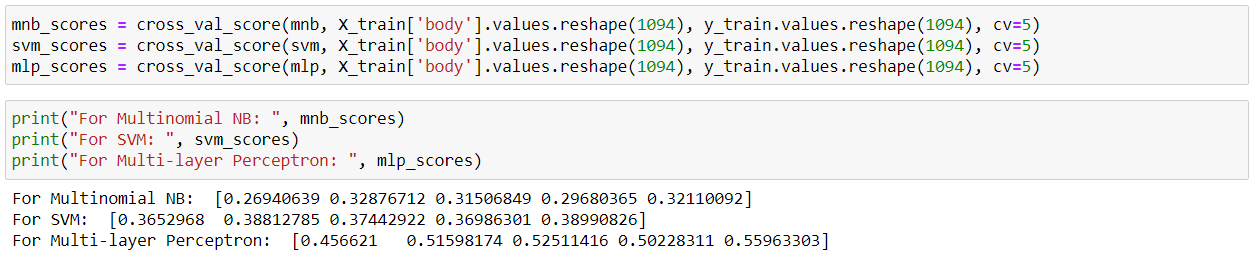


Figure 8 - Confusion matrices for algorithm performances from top to bottom, Multinomial NB, SVM, MLP.



Results from 5-fold cross-validation for each algorithm:



The results of the algorithms average at about 39.6%. The reason for such poor results is surmised to be a two-fold reason. The research paradigm of this project was chosen to be NLP along with multi-class classification algorithms to train the algorithms to achieve the aims and objectives of this research. These mediocre results can therefore be associated with two reasons which are as follows:

* The dataset was not properly utilized. The reason can be that the number of features and the number of classes with respect to the number of instances were not enough for training. The other reason can be that the dataset consists of categorical data and the encoding of the categorical data has not been very efficient thereby reducing the quality of the dataset passed on to the algorithm.
* The algorithms used were not modified or tweaked in any way. The algorithms used were in their default configuration. Each of the algorithms used in this research has a set of parameters that control the architecture and the features of the algorithm known as hyperparameters. Although the term to describe all of the parameters is the same, each algorithm has a unique set of hyperparameters that need adjusting to provide the best results for a unique dataset. Scikit-learn, the library from which all these algorithms were taken also provides an automated method of optimizing the algorithm hyperparameters for a given dataset. Although this could have been employed in this research, due to time constraints, it was left out. Implementing this must be able to provide much better results than before.

## Conclusion

The aim of this research was to create a Machine Learning algorithm that can classify between different topics/niches of a Reddit post and thus identify the niche of a specific Reddit post. To this extent, a few objectives had been laid down, the achievement of which would lead to the achievement of the aim. The first objective was to perform a solid base for the research by doing literature review. Literature review has many advantages. It prepares the researcher for any issues they might face during the research from the experience of the other fellow researchers in the same domain. Since literature review works with research papers that are already published, referring to them and citing them can increase the integrity and validity of the research being done.

The second objective is data-oriented. Any Machine Learning project revolves around data. Without data, there is no Machine Learning. To guarantee the successful completion of this research, a proper dataset had to be selected. The topic of this research is highly specific in its requirements and not much work has been done to create a public dataset that can satisfy the needs of this research. However, with arduous research, a dataset that perfectly fit the description of this research was found. The next step was to prepare the dataset for usage. Data preprocessing is a process that prepares the data by manipulating it in such a way that it becomes easier for the Machine Learning algorithms to work on them. Data visualization, which is a process of graphically representing the data to get a deeper insight of its properties, trends and working is also performed. The preprocessing techniques used for this dataset arise from the Natural Language Processing due to the presence of text data in the dataset.

The Machine learning algorithms chosen for this research arise from the Multi-class classification paradigm of Machine Learning. The results of the algorithms were subpar since the dataset was not properly used and the algorithm hyperparameters were not optimized for the research. There is scope of improvement in the future of this research.

# References

1. Alexa. (2016, February 6). Top Sites. Retrieved February 6, 2016, from [http://www.alexa.com/topsites/global;1](http://www.alexa.com/topsites/global%3B1)
2. Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley Pub. Amichai-Hamburger, Y., & Furnham, A. (2007). The positive net. *Computers in Human*
3. *Behaviour, 23*(2), 1033-1045. doi:10.1016/j.chb.2005.08.008
4. Amichai-Hamburger, Y., & Mckenna, K. Y. (2006). The contact hypothesis reconsidered: Interacting via the internet. *Journal of Computer-Mediated Communication, 11*(3), 825- 843.
5. Amichai-Hamburger, Y., Hasler, B. S., & Shani-Sherman, T. (2015). Structured and unstructured intergroup contact in the digital age. *Computers in Human Behavior, 52*, 515-522. doi:10.1016/j.chb.2015.02.022
6. Anderson, J., & Rainie, L. (2014, March 11). Digital life in 2025. Retrieved from [http://www.pewinternet.org/2014/03/11/digital-life-in-2025/#pq=bFgtdC](http://www.pewinternet.org/2014/03/11/digital-life-in-2025/#pq%3DbFgtdC)
7. Appel, M. (2012). Are heavy users of computer games and social media more computer literate?
8. *Computers & Education, 59*(4), 1339-1349. doi:10.1016/j.compedu.2012.06.004
9. Michael, A., Armando, F., Rean, G., Anthony, D. J., & Katz, R. (2010). A view of cloud computing. Communications of the ACM.
10. Fan, W., & Gordon, M. D. (2014). The power of social media analytics. Communications of the ACM, 57(6), 74-81.
11. Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and the biggest growth in the Far East. IDC iView: IDC Analyze the future, 2007(2012), 1-16.
12. Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. IEEE Intelligent Systems, 25(6), 13- 16.
13. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
14. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
15. Karun, A. K., & Chitharanjan, K. (2013, April). A review on Hadoop—HDFS infrastructure extensions. In 2013 IEEE conference on information & communication technologies (pp. 132-137). IEEE.
16. Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. Business Horizons, 52(4), 357-365.
17. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information sciences, 275, 314-347.
18. Flood, M. D., Jagadish, H. V., & Raschid, L. (2016). Big data challenges and opportunities in financial stability monitoring. Banque de France, Financial Stability Review, 20, 129-42.
19. Kulkarni, A. P., & Khandewal, M. (2014). Survey on Hadoop and Introduction to YARN.
20. Coronel, C., & Morris, S. (2016). Database systems: design, implementation, & management. Cengage Learning.
21. Sowmya, R., & Suneetha, K. R. (2017, January). Data mining with big data. In 2017 11th International Conference on Intelligent Systems and Control (ISCO) (pp. 246-250). IEEE.
22. Basha, S. A. K., Basha, S. M., Vincent, D. R., & Rajput, D. S. (2019). Challenges in Storing and Processing Big Data Using Hadoop and Spark. In Deep Learning and Parallel Computing Environment for Bioengineering Systems (pp. 179-187). Academic Press.
23. Boyd, D. (2011). Social network sites as networked publics (Z. Papacharissi, Ed.). In *A networked self: Identity, community and culture on social network sites* (pp. 39-57). New York, NY: Routledge.
24. Brown, R., Condor, S., Mathews, A., Wade, G., & Williams, J. (1986). Explaining intergroup differentiation in an industrial organization. *Journal of Occupational Psychology, 59*(4), 273-286. doi:10.1111/j.2044-8325.1986.tb00230.x
25. Subroto, A., & Apriyana, A. (2019). Cyber risk prediction through social media big data analytics and statistical machine learning. *Journal of Big Data*, *6*(1),50.
26. Dlamini, M.T., Eloff, J.H., & Eloff, M.M. (2009). Information security: The moving target *computers & security*, *28*(3-4), 189-198
27. Meso, P., Ding, Y., & Xu, S. (2013). Applying protection motivation theory to information security training for college students. *Journal of Information Privacy and Security*, *9*(1),47-6
28. Craigen, D., Diakun-Thibault, N., & Purse, R. (2014). Defining cybersecurity. *Technology Innovation Management Review*, *4*(10).
29. Posey Garrison, C., & Ncube, M. (2011). A longitudinal analysis of data breaches. *Information Management & Computer Security*, *19*(4),216-230.
30. Gordon, L. A., & Loeb, M. P. (2006). Economic aspects of information security: An emerging field of research. *Information Systems Frontiers*, *8*(5),335-337.
31. Cazacu, M., & Cucu, C. (2019). Using the Activity Theory to Identify the Challenges of Designing E-learning Tools based on Machine Learning for Security Operations Centers. In *The International Scientific Conference eLearning and Software for Education* (Vol. 1, pp. 452-461). " Carol I" National Defense University.
32. Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H., & Stoddart, K. (2016). A review of cybersecurity risk assessment methods for SCADA systems. *Computers & Security*, *56*, 1-27.
33. Peterson, C., & Rögnvaldsson, T. S. (1991). *An introduction to artificial neural networks* (No. LU-TP-91-23). CERN.
34. Bitter, C., Elizondo, D. A., & Watson, T. (2010, July). Application of artificial neural networks and related techniques to intrusion detection. In *The 2010 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
35. Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938.
36. Lek, S., & Guégan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, *120*(2-3), 65-73.
37. Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225-1231.
38. Potluri, S., Diedrich, C., & Sangala, G. K. R. (2017, September). Identifying false data injection attacks in industrial control systems using artificial neural networks. In *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (pp. 1-8). IEEE.
39. Lee, W., Stolfo, S. J., Chan, P. K., Eskin, E., Fan, W., Miller, M., ... & Zhang, J. (2001, June). Real-time data mining-based intrusion detection. In *Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01* (Vol. 1, pp. 89-100). IEEE.
40. Dartigue, C., Jang, H. I., & Zeng, W. (2009, May). A new data-mining based approach for network intrusion detection. In the *2009 Seventh Annual Communication Networks and Services Research Conference* (pp. 372-377). IEEE.
41. Dreossi, T., Donzé, A., & Seshia, S. A. (2019). Compositional falsification of cyber-physical systems with machine learning components. *Journal of Automated Reasoning*, *63*(4), 1031-1053.
42. Linda, O., Vollmer, T., & Manic, M. (2009, June). Neural network-based intrusion detection system for critical infrastructures. In *2009 international joint conference on neural networks* (pp. 1827-1834). IEEE.
43. Tarik, M., & Singh, A. (2016)Intrusion Detection System using K2 Self Learning Algorithm and Open Attacking Platform. *International Journal of Scientific & Engineering Research, 6,* 543-548.
44. Liu, J., Xiao, Y., Li, S., Liang, W., & Chen, C. P. (2012). Cybersecurity and privacy issues in smart grids. *IEEE*
45. Beuker, R. and Abbing, E. (2010). Two Faces of Social Media: Brand Communication and Brand Research. Design Management Review, 21(1), pp.54--60.
46. Blaxter, L., Hughes, C. and Tight, M. (2010). How to research. 1st ed. Maidenhead: Open University Press/McGraw-Hill Education.
47. Brodie, R., Ilic, A., Juric, B. and Hollebeek, L. (2013). Consumer engagement in a virtual brand community: An exploratory analysis. Journal of Business Research, 66(1), pp.105--114.
48. Bryman, A. (2012). Social research methods. 4th ed. Oxford: Oxford University Press.
49. Chen, Y., Fay, S. and Wang, Q. (2011). The role of marketing in social media: How online consumer reviews evolve. Journal of Interactive Marketing, 25(2), pp.85--94.
50. Costa, R. (2013). The Effects of Online Brand Communities on Brand Equity in the Luxury Fashion Industry. International Journal of Engineering Business Management Special Issue on Innovations in the Fashion Industry. [online] Available at: http://cdn.intechopen.com/pdfs-wm/45560.pdf [Accessed 29 May. 2014].
51. Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. Management Science, 49(10), pp.1407--1424.
52. De Vries, L., Gensler, S. and Leeflang, P. (2012). The popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing. Journal of Interactive Marketing, 26(2), pp.83--91.
53. Doç, Y. and Ulusu, Y. (2010) ‘Determinant factors of time spent on Facebook: brand community engagement and usage types’ Journal of Yaşar University, 18(5) 2949-2957.
54. Kozinets, R.V. (1999). "E-tribalized marketing? the strategic implications of virtual communities of consumption", European Management Journal, vol. 17, no. 3, pp. 252-264.
55. Krishnamurthy, S. and Dou, W. (2008). Note from special issue editors: advertising with user-generated content: a framework and research agenda. Journal of Interactive Advertising, 8(2), pp.1--4.
56. Mohr, I. (2013). The Impact of Social Media on the Fashion Industry. Journal of Applied Business and Economics, 15(2), pp.17--22.
57. Moran, E. 2010. Marketing in a hyper-social world – The tribalization of business study and characteristics of successful online communities. Journal of Advertising Research, 232–239. doi:10.2501/S0021849910091397
58. Muk, A., Chung, C. and Kim, J. (2014). A Cross-National Study of the Influence of Individualism and Collectivism on Liking Brand Pages. Journal of International Consumer Marketing, 26(2), pp.122--137.
59. Digital Firefly Marketing, (2014). Social Media Trends: Why do people follow a brand? [online] Available at: https://digitalfireflymarketing.com/social-mediatrends-why-do-people-follow-brand [Accessed 27 Jul. 2014].
60. Hampton, K., Goulet, L. S., Rainie, L., & Purcell, K. (2011, June 16). Social networking sites and our lives. Retrieved from <http://www.pewinternet.org/2011/06/16/social-networking-> sites-and-our-lives/
61. Hansen, M., Fabriz, S., & Stehle, S. (2015). Cultural cues in students' computer-mediated communication: Influences on email style, perception of the sender, and willingness to help. *Journal of Computer-Mediated Communication,* 1-17.
62. Hardaker, C. (2015). ‘I refuse to respond to this obvious troll’: An overview of responses to (perceived) trolling. *Corpora, 10*(2), 201-229. doi:10.3366/cor.2015.0074
63. Harwood, J. (2010). The contact space: A novel framework for intergroup contact research.
64. *Journal of Language and Social Psychology, 29*(2), 147-177. doi:10.1177/0261927X09359520
65. Hecht, M. L., Warren, J. R., Jung, E., & Krieger, J. L. (2004). A communication theory of identity: Development, theoretical perspective, and future directions. In W. B. Gudykunst (Author), *Theorizing about intercultural communication* (pp. 257-278). Thousand Oaks, CA: Sage.
66. Glorot, X. and Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, [online] 9, pp.249-256. Available at: <http://proceedings.mlr.press/v9/glorot10a.html> [Accessed 8 June 2021].
67. Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F. and Hao, H., 2015. Semantic Clustering and Convolutional Neural Network for Short Text Categorization. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers).
68. Bottou, L., 2010. Large-Scale Machine Learning with Stochastic Gradient Descent. Proceedings of COMPSTAT'2010, pp.177-186.
69. Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5 - RMSProp.
70. Ukil, A., 2007. Support Vector Machine. Power Systems, pp.161-226.
71. Suthaharan, S., 2016. Support Vector Machine. Machine Learning Models and Algorithms for Big Data Classification, pp.207-235.
72. Noble, W., 2006. What is a support vector machine? Nature Biotechnology, 24(12), pp.1565-1567.
73. Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine Learning, 20(3), pp.273-297.
74. Williamson, R., Smola, A. and Scholkopf, B., 2001. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. IEEE Transactions on Information Theory, 47(6), pp.2516-2532.
75. Burges, C., 1998. Data Mining and Knowledge Discovery, 2(2), pp.121-167.
76. Zaidi, N., Petitjean, F. and Webb, G., 2016. Preconditioning an Artificial Neural Network Using Naive Bayes. Advances in Knowledge Discovery and Data Mining, pp.341-353.
77. Murphy, K., 2006. Naive bayes classifiers. University of British Columbia.
78. Kulkarni, A. and Shivananda, A., 2019. Converting Text to Features. Natural Language Processing Recipes, pp.67-96.
79. Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), pp.45-65.
80. Ramos, J., 2003. Using tf-idf to determine word relevance in document queries.