Project Proposal

Extracting Fashion Related News from Facebook



|  |  |
| --- | --- |
| Name: | Supervisor: |
| Student ID: |  |
| Word Count: 4091 (Without References) |  |
| Date: |  |

Table of Contents

[Table of Figures 2](#_Toc68870631)

[Abstract 3](#_Toc68870632)

[Introduction 3](#_Toc68870633)

[Background 4](#_Toc68870634)

[Aim/Objective 4](#_Toc68870635)

[Research Questions 4](#_Toc68870636)

[Ethical Considerations 4](#_Toc68870637)

[Literature Review 5](#_Toc68870638)

[Domain Analysis of Information Extraction Techniques 5](#_Toc68870639)

[Methodology 5](#_Toc68870640)

[Results 9](#_Toc68870641)

[Information Extraction for Social Media 10](#_Toc68870642)

[Methodology 11](#_Toc68870643)

[Lithium NLP: A System for Rich Information Extraction from Noisy User Generated Text on Social Media 12](#_Toc68870644)

[Methodology 12](#_Toc68870645)

[Results 13](#_Toc68870646)

[Project Timeline 13](#_Toc68870647)

[Methodology 14](#_Toc68870648)

[Project Design 14](#_Toc68870649)

[Experimentation 14](#_Toc68870650)

[Mathematical Representation (Naïve Bayes): 14](#_Toc68870651)

[Mathematical Representation (TF-IDF) 15](#_Toc68870652)

[Dataset 15](#_Toc68870653)

[Project Evaluation 15](#_Toc68870654)

[Conclusion 15](#_Toc68870655)

[References 15](#_Toc68870656)

# Table of Figures

[Figure 1 - Popular Techniques for IE the Datasets they are used for in this research. 6](#_Toc68870657)

[Figure 2 - Drawbacks for each popular IE methodology based on the dataset it was applied on. 11](#_Toc68870658)

[Figure 3 - State-of-the-art vs Novel Proposed Methodology 12](#_Toc68870659)

[Figure 4 - Lithium NLP Pipeline Overview 13](#_Toc68870660)

[Figure 5 - Example working of Lithium NLP 13](#_Toc68870661)

[Figure 6 - Comparison of Lithium NLP IE with Traditional NLP methods 14](#_Toc68870662)

[Figure 7 - Comparison of Lithium NLP Language Support with Traditional NLP Methods 14](#_Toc68870663)

[Figure 8 - Gantt chart 14](#_Toc68870664)

[Figure 9 - Risk Management 15](#_Toc68870665)

# Abstract

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.

This research uses Machine Learning in the form of Naïve Bayes theorem and applies them to modern social media data. Specifically, it applies this to Facebook to extract information and posts from the domain of Fashion.

Social media is the new source of news for the world today. The requirement of having a better way to search and extract information from Facebook on Fashion has advantages for content creation and writing, or even just referring to a news article that went viral on Facebook.

From the extensive literature review performed for this research, it can be concluded that there have been only Pattern-based, Rule-based and Statistical approaches towards Information Extraction (IE). This paper explores the novel approach of applying probabilistic approaches like Naïve Bayes theorem to Social media IE.

**Keywords:** Machine Learning, Naïve Bayes Theorem, Social Media, Facebook, Information Extraction

# Introduction

The fast development in IT over the most recent twenty years has prompted a 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 (like Twitter and Facebook) (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 have the capacity to deal with rich client created content from web-based media stage have a few genuine applications. Besides, because of the substance style, size and heterogeneity of data (e.g., text, emojis, 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, Spasojevic and Hu, 2017).

But, there are a lot of resources available on Information Extraction (IE) on the internet. Techniques like NER, NLP, Statistical approaches have been used many times, However, to the extent of the Literature Review done for this research, there haven’t been any research papers focusing on probabilistic approaches for IE. This paper intends to explore the possibility of applying a probabilistic approach like Naïve Bayes Theorem to Facebook Data for IE.

The proposal will proceed as follows; the introduction section will continue defining the background of the project along with the objectives this project will achieve. The questions that this research will try to provide an answer for will also be discussed along with the ethical considerations of the project and the project timeline. The next section i.e., the methodology section will discuss the processes behind how the literature review will be performed and what the project design is. It will also discuss in detail the tools and theorems used for achieving the objectives this project has. The methodology section will end with a project evaluation that will convey the sections that the project comprises of and the order in which they need to be executed.

Following the former section will be the literature review which will discuss in detail the findings of related work in the domain. The literature review will summarize the methodology, experiments and findings of each research paper finalized for study in this project. The conclusion will be the last section summarizing the complete project followed by references, appendix and glossary.

## Background

The advent of social media as the harbinger of news to all people around the world effectively and irrespective of age has been realized unconsciously by all social media users and they make full use of it. Facebook in this regard has been thriving due to it being an early social media with a lot of accounts and following. With people posting the latest news on all trends and events as soon as they learn about it, news about the latest trends in fashion will also be on Facebook.

## Aim/Objective

Using Machine Learning, this project aims to automate the process of searching and extracting data on the latest trends and news in the Fashion domain from Facebook. This would be achieved through the following steps:

* Creating a machine learning algorithm/ensemble and training it to identify and extract information related to Fashion.
* Providing Data from Facebook for real-time testing and performance metrics.
* Analyzing the dataset created and providing statistical analysis on it.

## Research Questions

The following research question will give the project direction:

1. IE for social media has always been done using NER, NLP, statistical approaches. Will a probabilistic approach provide results on par with the state-of-the art techniques?

## Ethical Considerations

Ethics is a complicated subject that has only become more prominent during the advent of Big Data. For this research, Ethics brings the question of lawful Facebook Data consumption. Facebook is a publicly used social media platform and there are concerns regarding the usage of the data people put out on Facebook. There are existing datasets that are available on the internet that consist of Facebook data that is applicable to this research.

These datasets have been available for quite some time and are totally open-source and legal to use. For the requirement of data more specific and tailored, requesting Facebook developer section for the data is the ethical move. For this, you need to be registered with Facebook as a developer to access the data you need.

## Literature Review

Relevant literature is systematically gathered and analyzed to provide a basis for informing primary research. This information is used to justify the approach this paper takes to answering the research aim. Due to the availability of a large number of relevant papers, a selection criterion is finalized to increase the quality of LR. The selection criteria are as follows:

* The papers must be in English.
* Must be published in journals with a high impact factor.
* Complete and free access must be available for the journal.
* All the research datasets and code for the papers must be available for free.
* Papers earlier than a decade are not acceptable.

## Domain Analysis of Information Extraction Techniques

This paper studies the domain of Information extraction. This research provides a taxonomy of the tasks involved in IE. The paradigm for IE has also been studied and the popular algorithms for each paradigm have been studied. The weaknesses of each approach have also been studied.

### Methodology

For this research, the authors take data from four different domains for the domain analysis i.e., Social media, biomedical, chemical and unstructured data. The reason for choosing these domains was their violent development rate helpful outcomes appearing daily from them.

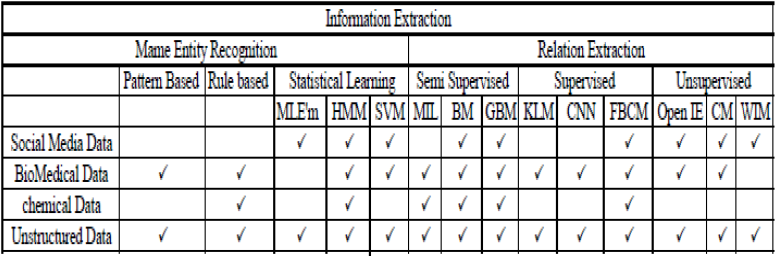


Figure 1 - Popular Techniques for IE the Datasets they are used for in this research.

There are two major tasks in IE. Both have popular approaches that are widely used for research purposes.

#### Name Entity Recognition (NER)

NER strives to locate and classify named entities from the content into predefined classes. The popular approaches used in this are as follows:

##### Pattern Based Method (PBM)

This is the earliest of methods for IE. Traditional IE methods use pattern recognition and word references to extract data from unlabeled data. PBM for IE is not easy to implement for social media. The reason behind this is that the data isn’t normalized. This is due to the risk of getting inaccurate results regarding patient diagnosis from the reports and patient data. PBM isn’t used in the chemical domain also. This is due to the very large data dictionary of chemical compounds, atoms and molecule names that isn’t possible for PBM to handle. Rule-based methods are used for data related to the chemical domain.

##### Rule-based Method (RBM)

This method has high congruency with Decision Trees (DT). The only difference is that in DTs overlapping is allowed which makes the training more robust.

The RBM has been used for protein and gene entity recognition in the biomedical domain. Using RBM makes it difficult to extract and label entities from unstructured data. But using word references i.e., seeds to kick-start that recognition process has worked in a few researches. RBM has also been used for question extraction and detection from twitter microblogs. In the chemical domain, RBM has been used to identify entities as well as normalize the text while identifying the boundaries of the sentences in the text.

##### Statistical Learning Method (SLM)

SML refers to the representation of the empirical laws that govern Natural Language Training with probability distribution. There are three popular approaches used in this method:

##### Maximum Likelihood Estimation Method (MLEM)

NER is easy for English due to the fact that this language tends to be comprehendible, the rules are clear and the sentences are created by combining words together. But a language like Chinese is difficult since sentence making in Chinese is the total opposite of English. MLEM has been used to successfully make sense of the unstructured Chinese data. MELM has been used for calculating probabilities in the biomedical domain. MLEM also helps increase the efficiency if MLEM is involved in the algorithm training process.

MLEM cannot be used for any work in the chemical domain for the same reason PBM cannot be used. The complexity of the data i.e., the equations, formulas is way too much. MLEM has been used for entity recognition from social media paired with the tag n-gram approach.

##### Hidden Markov Model (HMM)

HMM has been used successfully for NER in languages like Kannada that provide complex unstructured data. HMM was combined with the RBM approach for very impressive results.

Social media related datasets have a difficult time working with HMM due to irregular sentence structuring, message language and irregular capitalization. But results have been achieved in this domain by combining HMM with n-gram for a hybrid approach N-gram Language Markov Model (NLMM). A much-needed success in the chemical domain regarding extraction of information has been achieved by HMM. Linking chemical names of elements with their symbols i.e., the symbol for lead is PB.

##### Support Vector Machines (SVM)

SVMs use the power of linear algebra to create boundaries between different data classes. The intensity of separation is key in the results from SVM. This approach has been successfully utilized to extract individual information from the Swedish language like names, company location and time. SVM has also been utilized for biomedical data NER. The issue in dealing with social media data is that normal NER methods are created to deal with formal text properly written literature available in newspapers, etc. But social media doesn’t have this. Moreover, there is data in multiple languages on social media. But SVM has outperformed HMM in this regard. The only disadvantage of SVM is that it is slow.

#### Relation Extraction

The second task in IE, this extracts the relationship between entities and the content of the data. The following popular approaches are used for this:

##### Supervised Learning Models

The goal here is to identify patterns and learn from a labelled dataset for prediction. The following approaches are popular in this section:

###### Feature Based Classification Method (FBCM)

The difference between RBM and FBCM is the scope of data utilized. While RBM uses a focused, smaller scope of data for explaining the relations between entities and the context, FBCM uses a wide range of contextual data. Traditional FBCM is unfit when it comes to social media IE. Therefore, relation extraction, a technique to extract the relations between linked entities in a text is used.

The issue with FBCM is that occasionally information with unequivocal component vectors can't be addressed without any problem. In these cases, the extraction of qualities is an exceptionally perplexing errand and prompts vectors of extremely high measurement, which thusly prompts issues of computation. Portion based techniques endeavor to take care of this issue verifiably depend on scalar vector items are determined at extremely high Dimensional spaces with no sort of vector should be explicit.

##### Semi-supervised Learning Methods

Algorithms are trained on a dataset containing unlabeled and labelled data. This process is however expensive in terms of processing, time and data required. Using approaches that do not require much data is essential.

###### Multi-Instance Learning (MIL)

The preparation dataset for this methodology comes in areas, and each part is loaded up with numerous cases. MIL calculations discover events in the areas. At that point those events are utilized for preparing. The events are described by the segments to which they have a place.

###### Bootstrapping Method (BM)

BM used to separate connections, have acquired impressive consideration lately. These methodologies are built with a key suspicion that on the off chance that you a few words that relate with a particular goal in mind, phrases containing these words, these sorts of words probably make connections to communicate. Consequently, the sentences containing the word pair are utilized as preparing information for the connection extraction.

###### Graph Based Method (GBM)

BM are notable for the extraction of connections, mostly since they require a tiny bit of measure of human appraisal. Charts can address complex connections among classes and occasions. An equivocal occurrence; For instance, Usman Khawaja may be among the class of pilots and players. Today web-based media like Facebook beats the web. Scientists are chipping away at the investigation of online media for connection extraction from remarks, tweets just as posts. In online media, the proposed execution of this strategy is a powerful spiral chart to adapt to the impediments of past representation procedures. The capacity to accurately get a handle on both semantic and syntactic designs in biomedical information ends up being dynamically essential and permit serious comprehension of logical papers and clinical information.

##### Unsupervised Learning Methods

In unaided realizing there is no such managed or organized information and we simply have input data. The goal is to find predictabilities at the event.

###### Clustering Method (CM)

Because of the huge number of connections between substances, it very well might be expensive to cover an adequately huge measure of preparing information to successfully stamp each sort of relationship in each new space of interest. CM can be viewed as significant learning without managing the issue; actually, like any such issue, attempt to discover a construction in an assortment of unlabeled information just as it is vital unaided method. Substances are assembled by their absolute data. To utilize web-based media networks for the Semantic Web, a couple of surveys have dissected programmed connection extraction of online media.

###### Open Information Extraction (Open IE)

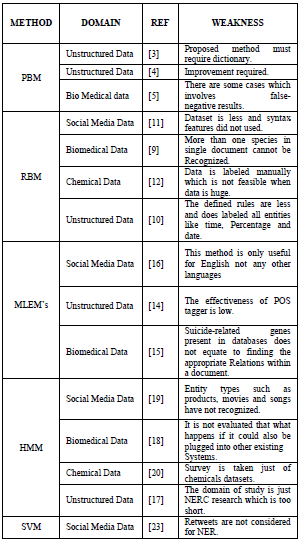
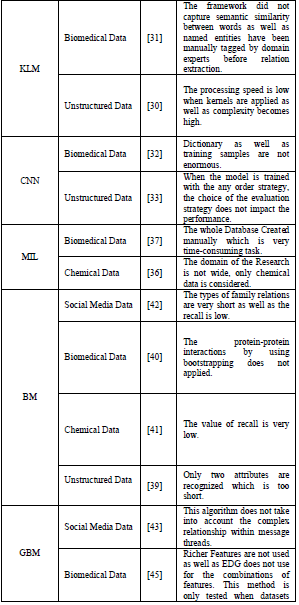
IE structures endeavor to remove the semantics of text in normal language associations; in any case, most systems use oversaw express occasions of the relationship to learn and in this way most of the way by the ease of use of planning data. Open IE framework, depends absolutely upon the information substance and its etymological qualities. Much more particularly make advisers for get these attributes of substance and from there on discrete relations.

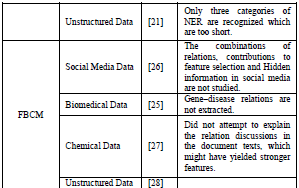
###### Wrapper Induction Method (WIM)

This procedure is an extraction method that houses a course of action of extraction rules and program code needed to execute these principles. WIM get familiar with the covering naturally. Given a succession of preparing information, the acceptance calculation sorts out some way to remove a covering objective information.

### Results

The weaknesses of each approach have been summarized in the following tables from the research paper.





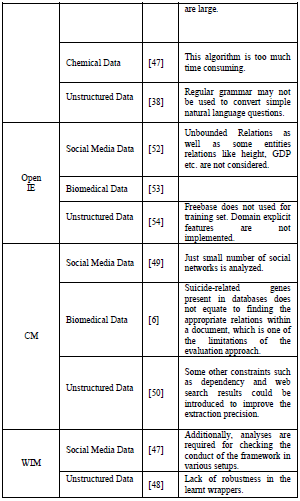


Figure 2 - Drawbacks for each popular IE methodology based on the dataset it was applied on.

## Information Extraction for Social Media

In this exploration paper, the makers propose a design for Information Extraction (IE) from unstructured customer delivered substance by means of online media. The design proposes answers for rout the IE challenges in this space like the short setting, the noisy meager substance and the problematic substance. To overcome the challenges standing up to IE from online media, State-Of-The-Art approaches ought to be changed in accordance with suit electronic media posts. The key sections and parts of the proposed framework are noisy substance filtering, named component extraction, named substance disambiguation, analysis circles, and weakness dealing with.

### Methodology

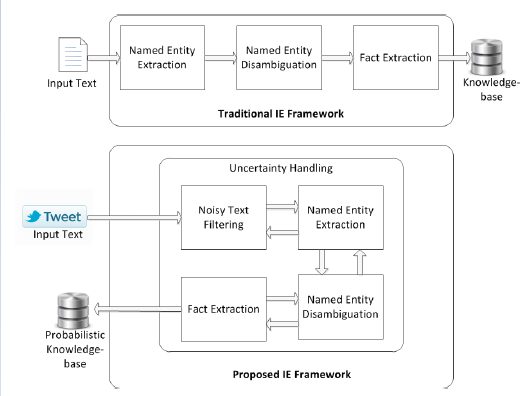


Figure 3 - State-of-the-art vs Novel Proposed Methodology

* **Noisy text Filtering**: It is needed to channel non-educational posts since there are a great deal of tweets made each day and not all are valuable in this unique circumstance. Sifting should be possible dependent on area or language or different standards to try to keep just applicable posts that contains data about the space should be handled.
* **Named Entity Extraction**
* **Named Entity Disambiguation**: Scientists regularly connect substances to Wikipedia articles or to KB passages. For online media posts, here and there this is preposterous as a large number of the referenced substances can't be connected to Wikipedia articles or a KB passage. Notwithstanding, regularly clients have home pages or profiles on an online media network. Besides, celebrations and nearby occasions additionally generally have home pages addressing these occasions. This exploration proposes an open world methodology for NED for tweets. Named elements are disambiguated by connecting them to a landing page or an informal community profile page in the event that they don't have a Wikipedia article. Target (tweets rotating around same occasion) are utilized to enhance the tweet setting and subsequently to improve the viability of tracking down the right substance page. Other metadata from client profiles could likewise be utilized to improve the disambiguation interaction.

## Lithium NLP: A System for Rich Information Extraction from Noisy User Generated Text on Social Media

Research done here displays the Lithium Natural Language Processing Framework – a tool obliged, high throughput and language-agnostic framework for data extraction from loud client produced text via online media. Lithium NLP removes a rich arrangement of data including substances, subjects, hashtags and notion from text.

### Methodology

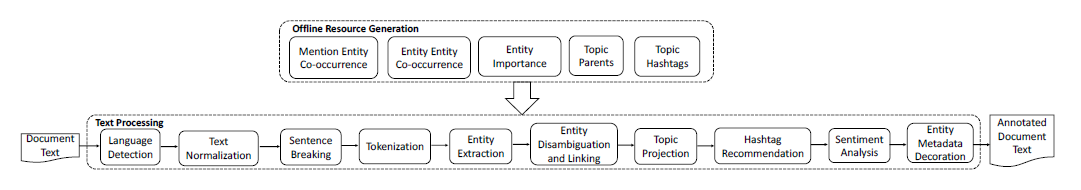


Figure 4 - Lithium NLP Pipeline Overview

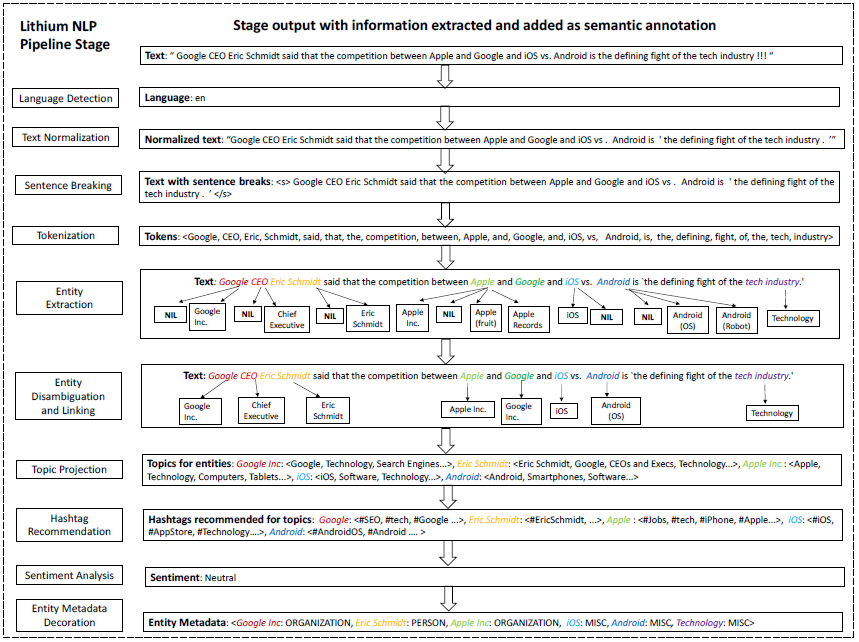


Figure 5 - Example working of Lithium NLP

### Results

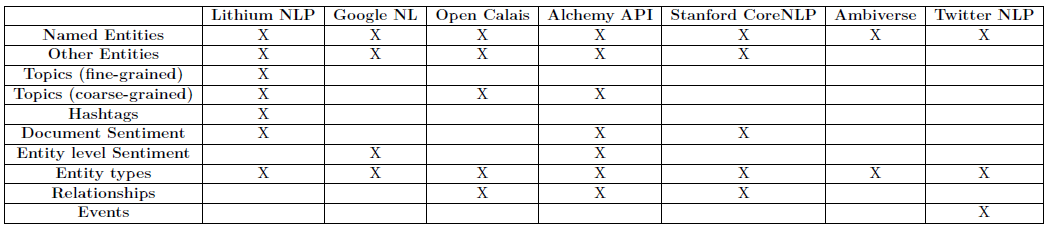


Figure 6 - Comparison of Lithium NLP IE with Traditional NLP methods

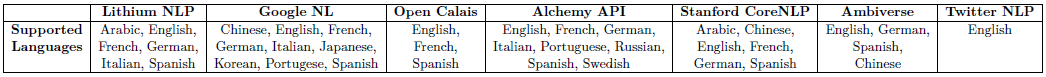


Figure 7 - Comparison of Lithium NLP Language Support with Traditional NLP Methods

## Project Timeline

Research projects are random and time bound and the ability to meet a deadline is key to success. A project timeline clearly 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 1*. 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 in a timely manner to a high standard.



Figure 8 - Gantt chart

There may be unexpected events that cause difficulty in maintaining the project timeline. These are risks. Contingencies for some expected risks are discussed below in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Risk Impact (1-5)** | **Risk Likelihood (1-5)** | **Contingency** |
| Data Mismanagement | 5 | 1 | Each iteration of the dataset will be synced with Version Control. |
| Personal Time Mismanagement | 3 | 3 | Regular Feedback from Supervisor for Continuous development. |
| Hardware Failure | 1 | 1 | Extra Supply of RAM and SSD. |

Figure 9 - Risk Management

# Methodology

As mentioned above, this project will use Machine Learning to identify and extract Fashion Related content from Facebook. This is achieved by the following stages:

## Project Design

This project is a research using Machine Learning to identify and extract Fashion Related Data from Facebook. This will consist of experiments and data analysis. Since this is an experimental research without a definitive quantitative goal, there will not be a Software Development Life Cycle (SDLC) followed. This is to allow freedom from constraints that come with an SDLC.

## Experimentation

Naïve Bayes is a famous probabilistic approach that is most popularly used for spam detection. In this research, the aim is to use Naïve Bayes theorem for IE. The key to this is to make congruent the approach to IE with Spam detection. Spam detection works on the basis of keywords. IE in this research will also work on the same principle. Based on specific keywords, Naïve Bayes can identify and thus help extract posts from Facebook that are Fashion related.

### Mathematical Representation (Naïve Bayes):

Naïve Bayes Theorem is represented by the following equation: ; where A and B are two events, and P () represents the probability of the occurrence of the event. Naïve Bayes helps determine the occurrence of event A provided event B has occurred.

Let,

* Event A be the occurrence of a specific keyword related to Fashion and,
* Event B be the identification of a FB post as a Fashion related post.

The probability that a post would be Fashion related provided the keyword has already occurred is what Naïve Bayes will be providing. Since the Naïve Bayes theorem requires a set of keywords using which it would identify Fashion related posts from the chosen dataset, the methodology for extracting keywords is required.

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. TF-IDF (term frequency-inverse document frequency) was invented for document search and information retrieval. It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don’t mean much to that document in particular.

### Mathematical Representation (TF-IDF)

TF-IDF score for a word *t* in a document *d* is calculated from the document set *D* as follows:

### Dataset

The dataset for this research will be the [Interactive Facebook Reactions Dataset](https://github.com/minimaxir/interactive-facebook-reactions). This dataset contains Facebook Posts of many popular pages and users in between June 20th, 2014 to June 24th, 2014. The dataset contains the status ID and the Facebook post itself.

# Project Evaluation

The project will be evaluated based on the following:

* Comparison of Naïve Bayes Theorem performance with state-of-the-art IE approaches from previous research papers.

The performance measure for Naïve bayes would be the confusion matrix that will display the precision, recall for the algorithm. This will be compared with the performance of other algorithms.

# Conclusion

The research will be exploring the effects and advantages of using probabilistic approaches for IE, in particular the Naïve Bayes theorem for IE. The literature review provides us with a list of popular algorithms and approaches used by the researchers but the mention of probabilistic approaches isn’t there. The experiment would be to provide the keywords related to Fashion for the Naïve Bayes using the TD-IDF approach. Then the Naïve Bayes will be trained using the selected dataset and the keywords. The final performance will be measured using confusion matrices for evaluation.

# References

* Badieh Habib Morgan, M. and van Keulen, M., 2014. Information Extraction for Social Media. Proceedings of the Third Workshop on Semantic Web and Information Extraction.
* Bhargava, P., Spasojevic, N. and Hu, G., 2017. Lithium NLP: A System for Rich Information Extraction from Noisy User Generated Text on Social Media. Proceedings of the 3rd Workshop on Noisy User-generated Text.
* Alam, T. and Awan, M., 2018. Domain Analysis of Information Extraction Techniques. ResearchGate.