Democratic and Popular Republic of Algeria
Ministry of Higher Education and Scientific Research

University of Adrar
Faculty of Sciences and Technology
Department of Mathematics and Computer Science

جامعة أحمد دراية.أدرار-الجزائر
Université Ahmed Draia.Adrar -Algérie

A Thesis Presented to Fulfill the Partial Requirements of Master's degree in Computer Science

Option: intelligent systems

Title

Big Data Analytics Using fuzzy logic approach

Prepared by:

LOURDJANE Halima

BENLARIA Hosnia

Supervised by:

Dr. OMARI Mohammed

Academic Year: 2018/2019

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Abstract

In the era of big data, we are facing with an immense volume and high velocity of data with complex structures. Data can be produced by online and offline transactions, social networks, sensors and through our daily life activities. A proper processing of big data can result in informative, intelligent and relevant decision making completed in various areas, such as medical and healthcare, business, management and government. Fuzzy sets have been employed for big data analyzing due to their abilities to represent and quantify aspects of uncertainty. In this thesis, we have implemented and analyzed big dataset using a fuzzy logic where four rule groups are used and their results have been analyzed and compared. Results showed that the number of rules affects the false negative and the false positive rates when final decisions are compared to those of experts.

Key words: Big data, big data analytics, FIS, Fuzzy set, Crisp set.

Résumé

À l'ère du Big Data, nous sommes confrontés à un volume immense et à une grande vitesse de données avec des structures complexes. Les données peuvent être produites par des transactions en ligne et hors ligne, des réseaux sociaux, des capteurs et par le biais de nos activités de la vie quotidienne. Un traitement correct des données volumineuses peut aboutir à une prise de décision informative, intelligente et pertinente dans divers domaines, tels que les domaines médicaux et le soin de santé, les entreprises, la gestion et le gouvernement. Les ensembles flous ont été utilisés pour l'analyse de données volumineuses en raison de leur capacité à représenter et à quantifier les aspects de l'incertitude. Dans ce mémoire, nous avons implémenté et analysé de grands ensembles de données en utilisant la logique floue, dans laquelle quatre groupes de règles sont utilisés avec l’analyse et la comparaison de leurs résultats. Ces résultats ont montré que le nombre de règles affecte les taux de faux négatifs et de faux positifs lorsque les décisions finales sont comparées à celles des experts.

Mots clés : Big Data, Big Data Analytics, FIS, ensemble flou, ensemble impeccable.

ملخص:

في عصر البيانات الكبيرة، نواجه كميات هائلة وبيانات عالية السرعة بهياكل معقدة. يمكن إنتاج البيانات عن طريق المعاملات عبر الإنترنت وغير المتصلة والشبكات الاجتماعية وأجهزة الاستشعار ومن خلال أنشطتنا اليومية. يمكن أن تؤدي المعالجة الصحيحة للبيانات الضخمة إلى اتخاذ قرارات غنية بالمعلومات وذكية وذات صلة في مجموعة متنوعة من المجالات، مثل الرعاية الطبية والصحية، والأعمال التجارية، والإدارة، والحكومة. تم استخدام مجموعات غامضة

لتحليل البيانات الكبيرة بسبب قدرتها على تمثيل وتحديد جوانب عدم اليقين. في هذه المذكرة، قمنا بتطبيق وتحليل مجموعة بيانات كبيرة باستخدام المنطق الغامض حيث يتم استخدام أربع مجموعات قواعد وتم تحليل نتائجها ومقارنتها.

أوضحت النتائج أن عدد القواعد يؤثر على المعدلات السلبية الكاذبة والمعدلات الإيجابية الخاطئة عند مقارنة القرارات النهائية بعدد الخبراء.

الكلمات المفتاحية: البيانات الكبيرة، تحليلات البيانات الكبيرة، FIS، مجموعة ضبابية، مجموعة لا تشوبها شائبة.

Acknowledgements

"In The Name Of Allah the Most Gracious the Most Merciful"

*First and foremost, all praises go to Almighty **Allah** (SWT), the Most Gracious the Most Merciful. Peace upon the prophet **Mohammed S.A.W.** we would like to thank **God** almighty for giving us the strength, knowledge, ability and opportunity to undertake this research study and to preserve and complete it satisfactorily. Without his blessings, this achievement would not have been possible. Alhamdulillah.*

*We are deeply grateful to our ideal supervisor **Dr.OMARI Mohamed** for his patience, guidance, advice, and encouragement throughout the course of this work.*

We would like to thank the members of the jury for having accepted to assess our work, we are also thankful to all teachers of our department of Mathematics and Computer Science for their dedication and hard work.

There are no words to express our gratitude and thanks to our beloved parents, family members, and friends for always standing by us. Their love has been the major spiritual support in our life.

Dedicates

Praise be to Allah, Lord of the universe and peace and blessings be upon the seal of the prophets and messengers.

My Dear father ***Mohammed,***

As a sign of love and gratitude for all the support and sacrifices that he has shown me.

For my dear mother ***Djelloula,***

My reason for being, my reason for living, the star that illuminates my way.

For my dear brothers and my fiancé,

*For my friend and my dear sister "**LOURDJANE Halima**"*

For my friends,

To all my dear colleagues of study especially of 2nd year master promotion 2018/2019

In testimony of the sincere friendship and unwavering support, you had given me.

To all my family and all my best teachers.

I dedicated this work.

BENLARIA Hosnia

Dedicates

My Special Dedicates Go To :

My dear mother, who supported me, encouraged me to finish my studies and were eager to live this day,

And my dear father May God have mercy on him,

All my brothers, sisters and all the members of my family,

For my sister and my partner in this discuss: Hasna,

For my friends and groups and for everyone who helped me during my studies.

For the supervisor who helped us a lot, thank you,

To all my teachers and my friends in the study,
For everyone I love in the world.

LOURDJANE Halima

Table of Contents

CHAPTER I: Introduction and overview of Big Data analytics

CHAPTER II: Big data Analytics using fuzzy logic

CHAPTER III: Implementation of Fuzzy Inference System

CHAPTER IV : Results and Analysis

List of tables

List of figures

Abbreviations

ADBMS: Active **D**ata **B**ase **M**anagement **S**ystem

BI: Business **I**ntelligence

CSV: Comma **S**eparated **V**alues

FIS: Fuzzy **I**nference **S**ystem

FL: Fuzzy **L**ogic

FP: False **P**ositive

FPR: False **P**ositive **R**ate

FN: False **N**egative

FNR: False **N**egative **R**ate

GPS: Global **P**ositioning **S**ystem

HDFS: Hadoop **Distributed File S**ystem

NOSQL: No Structured **Q**uery **L**anguage

RDBM: Reliable **D**ata **B**ase **M**anager

TB: Tera **B**yte

TN: True **N**egative

TNR: True Negative **R**ate

TP: True **P**ositive

TPR: True **P**ositive **R**ate

URL: Uniform **R**esource **L**ocator

XML: Extensible **M**arkup **L**anguage

General Introduction

Big data analytics is the study of huge amounts of stored data in order to extract behavior patterns. More than 2.5 trillion bytes of information are generated every day through our smartphones, tablets, GPS devices, sensors spread all over our cities and bankcards. What can be done with all this information? This where Big Data analytics comes into play a combination of high technology systems and mathematics which together are capable of amazing all this information and providing it with a meaning of great values for companies governments.

Big data analytics helps companies or public administrations to understand the users better find previously unnoticeable opportunity provide a better service.

Big data can be analyzed by very different ways and methods, including Fuzzy logic. Our objective from this work is to propose a fuzzy logic model for sentiment analysis of social media network data (a Facebook posts).

This thesis consists of three chapters:

- The first chapter entitled **“Introduction and overview of Big Data analytics”** presents the general concepts related to the field of big data analytics.
- The second chapter entitled **“Big data Analytics using fuzzy logic ”** highlights the method used in our work which is the fuzzy logic
- The third chapter entitled **“Results and Analysis”** gives an overview of
- MATLAB environment, the simulation results of the selected method as well as comparison between different ways of analysis.

Finally, a conclusion is given to summarize our main contributions in addition to a section that describes our ideas of the future work.

Chapter I:

Introduction and overview of big data analytics

1.8 Background

In this chapter, Big Data analytics is indeed a revolution in the field of Information Technology, so we discussed in details for general understanding of the concept of this study. In a first part, the evolution and sources of data. Next, we highlight the concepts of Big Data, we also made a comparison between Big and small Data, and we explained the data structure. At the end of this part, we highlight the term of Big Data Analytic and it's concepts and tools. In the second part, we explain how social media generate Big Data.

1.8.1 **Data Evolution**

To better understand what Big Data is and where it comes from, it is crucial to first understand some history of data storage, repositories and tools to manage them. As shown in Figure 1, there has been a huge increase of data volume during the last three decades.

As we can see in the decade of 1990s the data volume was measured in terabytes. Relation databases and data warehouses representing structured data in rows and columns were the typical technologies to store and manage enterprise information.

Subsequent decade data started dealing with different kinds of data sources driven by productivity and publishing tools such as content managed repositories and networked attached storage systems. Consequently, the data volume was started being measured in petabytes.

As it is shown, the decade of 2010s has to deal with the exponential data volume driven by variety and many sources of digitized data. Almost everybody and everything is leaving a digital footprint. Some applications that generate a lot of data are shown in Figure 1. Due to much data, it has been considered to start measuring data in exabytes[1].

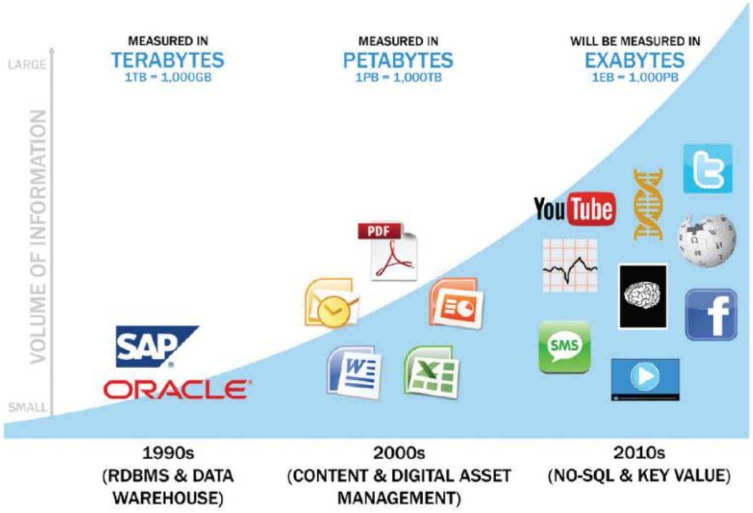


***Figure 1.1:** Data evolution and the rise of Big Data sources [1].*

1.8.2 Sources of Data

The sources of increasing the amount of data can be divided into a few categories of data generation:

- Machine.
- Human Interaction.
- Data Processing.

Typically, the first type is bound up with the spread of machines digitalization, which is related to sensors integration, the connectivity increase, and devices recording sounds, images or videos and to machines communication between each other. Particularly, devices such as cameras recording videos, cell phones collecting geospatial data, machines in production lines of industrial systems, are exchanging important information while processing their activities.

The next category concerns information exchange among people. For instance, some examples of social networks can be Facebook, Twitter, LinkedIn, etc. Every second these systems generate a huge amount of data shared by millions of people.

Data processing is concerned to deal with raw or already handled data to get an output or to get to another process phase that is involved in a particular project [1].

In summary, there are many sources generating a lot of data that can be stored for further analyses and explorations.

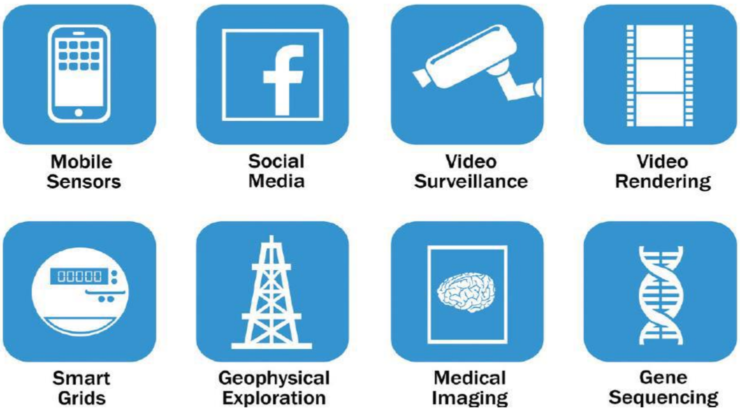


Figure 1.2: *What is driving the data [1].*

1.9 Definitions of Big Data

Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions [2].

Another definition of Big Data comes from McKinsey [3]:

*"*Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value."

The most known and used definition is originally described as "3V": Volume, Velocity and Variety (see Figure 3).

- ➢ Volume: Large amounts of data, from datasets with sizes of terabytes to zettabyte.
- ➢ Velocity: Large amounts of data from transactions with high refresh rate resulting in data streams coming at great speed and the time to act on the basis of these data streams will often be very short. There is a shift from batch processing to real time streaming.
- ➢ Variety: Data come from different data sources. For the first, data can come from both internal and external data source. More importantly, data can come in various formats such as transaction and log data from various applications, structured data as database table, semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more. There is a shift from sole structured data to increasingly more unstructured data or the combination of the two.

Some make it 4V's: (Volume, velocity, variety and veracity).

➢ Veracity: data in doubt uncertainty due to data inconsistency & incompleteness ambiguities, latency, deception, model approximations

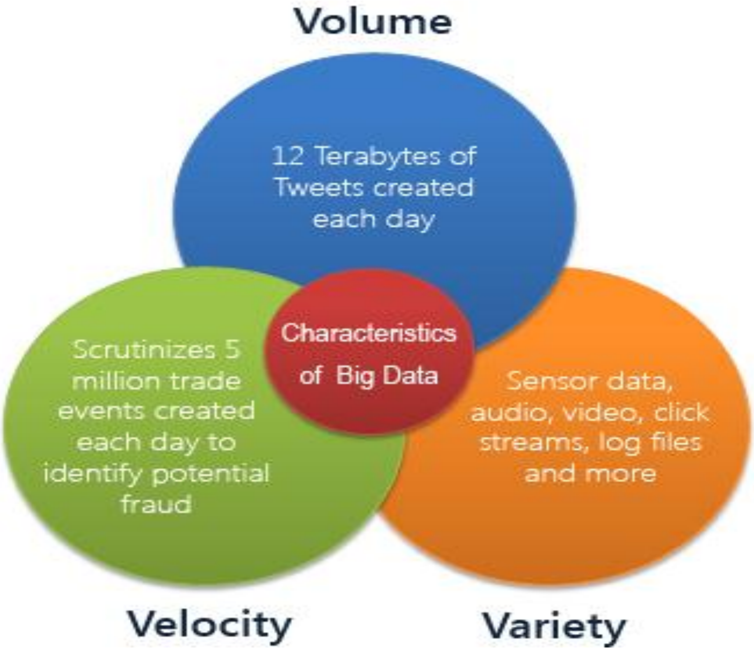

Volume
12 Terabytes of Tweets created each day
Characteristics of Big Data
Scrutinizes 5 million trade events created each day to identify potential fraud
Sensor data, audio, video, click streams, log files and more
Velocity
Variety

Figure 1.3: *Big Data characteristic [46].*

In relation to the definitions, the reason why there is intense complexity in processing Big Data is shown. Along with the Big Data there also exists ambiguity, viscosity and virality (see Figure 4).

- Ambiguity: emerges when there is less or no metadata in Big Data.

- Viscosity: this term is often used to describe the latency time in the data relative to the event of being described.

- Virality: describes how quickly data is shared throughout a network among people who are connected. The measurement result is the rate of spread of data in time.

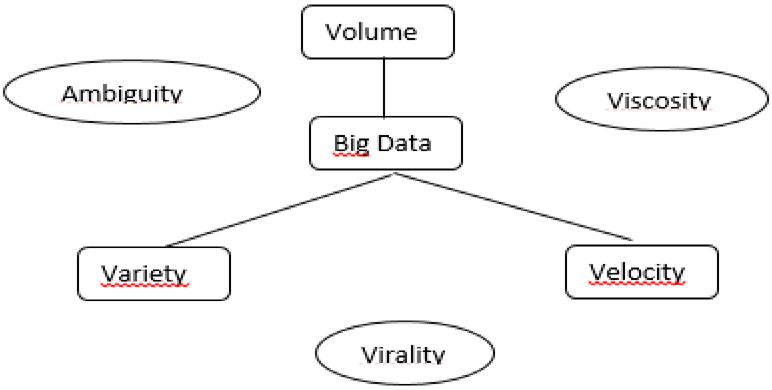


Figure 1.4: *Big Data characteristics derived from '3 V' definition [46].*

1.10 Big Data/Small Data

Big Data is not simply small data that has grown. There are more aspects that define differences between these two categories.

	Small Data	**Big Data**
Goals	Usually designed to answer a specific question or to achieve a particular goal	Nobody knows what the exact output of the project is. Usually it is designed with a goal in mind, but the goal is flexible
Location	Typically, small data is contained within one institution, often on one computer or even in one file	Big data are located throughout the company network or throughout the Internet. Typically, it is kept onto multiple servers, which can be everywhere
Data structure	Ordinarily contains highly structured data. Commonly, the data domain is restricted to a single discipline or its sub-sequence. The typical forms of its storage are uniform records or spreadsheets	Has to be capable of absorbing unstructured data such as text documents, images, sounds and physical objects. The subject of disciplines can vary throughout the data
	Small Data	**Big Data**
Data	In many cases, the data is prepared	The data is collected from many different

preparation	by its own user for his own purposes	sources. People who the data comes from are rarely the same people who use the data
Longevity	The data is kept for a limited time (academic life). After few years when the data project is finished, the data is usually discarded	A Big Data project typically contains data which has to be stored in perpetuity
Measurements	Typically, the data is measured using one experimental protocol	Many various types of data are measured in many different electronic formats
Reproducibility	The projects are typically repeatable. If there is a problem with the data quality, the entire project can be repeated	Replication of the project is seldom feasible. There is nothing more than optimism that the bad quality data is found and flagged, rather than be replaced by a better one
Stakes	Project costs are limited. The institution can usually recover from small data failure	Big data projects can be really expensive. A failed Big Data project can lead to bankruptcy
Introspection	Data points are identified by rows and columns within a spreadsheet or a database. It enables to address a particular data point unambiguously	It is more difficult to access the data. The organization and the context can be inscrutable. Access to the data is achieved by a special technique referred to as introspection
Analysis	In most cases, all the data in the project can be analyzed together	Big data is typically analyzed in incremental steps. It is extracted, reviewed, reduced, normalized, transformed, visualized, interpreted and reanalyzed with different methods

Table 1.2: Big Data versus Small Data [1].

1.11 Data Structure

As it is argued by [4] that approximately 80-90% of future data growth is coming from non-structured data types.

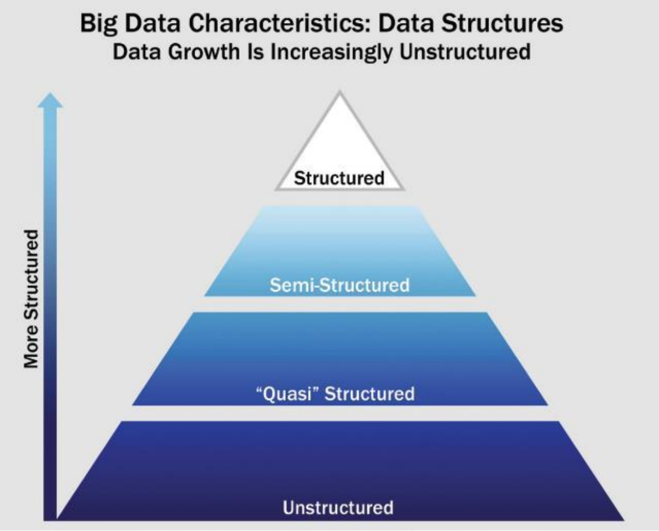

Big Data Characteristics: Data Structures
Data Growth Is Increasingly Unstructured
More Structured
Structured
Semi-Structured
"Quasi" Structured
Unstructured

Figure 1.5: *Big Data Formats [1].*

1.11.1 **Structured**

It is data that is stored in a structure that defines its format. Typically, structured data is categorized into groups which define how they will be stored (data types: number, text, etc.), processed, accessed and restricted (a set of possible values: male, female). Due to its structure, this type of data can be easily stored, processed and queried.

1.11.2 **Semi-structured**

Although it is a type of structured data, it has no strict model structure (XML for example).

1.11.3 **Quasi-structured**

In spite of not being commonly mentioned, this group can be added to these three categories as the one in between the semi-structured and the unstructured.

1.11.4 **Unstructured**

It is data that cannot be easily classified which makes it the opposite of structured data. They do not have any inherent structure.

1.12 Managing and Analyzing Big Data

The most important question that arises at this point of time is how do we store and process such huge amount of data; most of which is raw, semi structured, and may be unstructured. Big data platforms are categorized depending on how to store and process them in a scalable, fault tolerant and efficient manner [5]. Two important information management styles for handling big data are relational DBMS products enhanced for systematic workloads (often known as analytic RDBMSs, or ADBMSs) and non-relational techniques (sometimes known as NOSQL systems) for handling raw, semi structured and unstructured data. Non-relational techniques can be used to produce statistics from big data, or to preprocess big information before it is combined into a data warehouse.

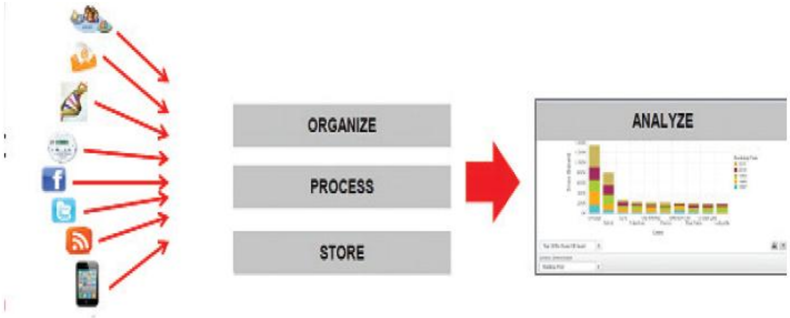


Figure 1.6: *Big Data management [6]*.

1.13 Big Data analytics

Big Data Analysis mainly involves analytical methods of big data, systematic architecture of big data, and big data mining and software for analysis. Data investigation is the most important step in big data, for exploring meaningful values, giving suggestions and decisions. Possible values can be explored by data analysis. However, analysis of data is a wide area, which is dynamic and is very complex [7].

1.13.1 **Traditional Data Analysis**

Traditional data analysis means the proper use of statistical methods for huge data analysis, to explore and elaborate the hidden data of the complex dataset, so that value of data can be maximized. Data analysis guides different plans of development for a country, predicting demands of customers, and forecasting the trends of market for organizations. Big data analysis may be stated as a technique of analysis of a special data. So, most of the traditional methods are still used for big data analysis [7].

1.13.2 Big Data Analytics tools

a) Hadoop [11]

Is an open source software project that enables the distributed processing of large data sets across clusters of servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Hadoop was derided from Google's MapReduce and the Google File System. Yahoo! was the originator, has been a major contributor, and uses Hadoop across its business. Other major users include: Facebook, IBM, Twitter, American Airliness, LinkedIn, The New York Times and many more.

A key aspect of the resiliency of Hadoop clusters comes from the softwares ability to detect and handle failures at the application layer. Hadoop has two main subprojects: first MapReduce, the framework that understands and assigns work to the nodes in a cluster, and secondly, HDFS, a distributed file system that spams all the nodes in a Hadoop cluster for data storage.

HDFS links together the file systems on many local nodes to make them into one big file system. HDFS assumes nodes will fail, so it achieves reliability by replicating data across multiple nodes.

Hadoop is supplemented by an ecosystem of Apache projects, such as; Pig, Hive and Zookeeper, that extends the value of Hadoop and improve its usability.

Hadoop changes the economics and the dynamics of large scale computing. it's impact can be boiled down to four characteristics:

1. **Scalable**: New nodes can be added as needed without needing to change: data formats or how data is loaded, or how jobs are written or the applications on top.

2. **Cost effective**: Hadoop brings massively parallel computing to large clusters of regular servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes analyzing all of our data affordable.

3. **Flexible**: Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways, enabling deeper analyses than other system can provide.

4. **Fault tolerance**: when we lose a node, the system redirect works to another location in the cluster and continues processing without missing a beat. All of this

happens without programmers having to write special or be aware of the mechanics of the parallel processing infrastructure.

With Hadoop, we can take data management and analytics to a whole new level.

b) OpenRefine [12]

OpenRefine is a sophisticated tool for working on big data and perform analytics. OpenRefine is able to perform various tasks on data. the tasks are, cleaning data, transformation of data from one format into the other format, and also extend with web services and data that are external.

The big idea behind choosing OpenRefine as our tool is to provide a tutorial by which users can have a free and an open source tool to manipulate their data sets. OpenRefine provides the flexibility to choose from a variety of data set functionalities, which makes it even more users friendly. Users can use this tool to get a big view of their data in terms of statistically curved graphs. They can play with messy data without worrying about risks, since they can undo their activity at any time. Cleaning, transforming and fetching URLs for a dataset can be easily done by simply having the application downloaded in the system.

A messy, unstructured, inconsistent dataset can be explored using open refine. In general, it will be very difficult to explore data through redundancies and inconsistencies. But, OpenRefine gives several functions through which one can filter the data, edit the inconsistencies, and view the data. It’s a tool to clean the data.

Spreadsheets can also refine a dataset but they are not the best tool for it as Openrefine cleans data in a more systematic controlled manner. While using historical data, we come across issues like blank fields, duplicate records, inconsistent formats and using Openrefine tool can help to resolve such issues.

Now data analysis plays an important role in business. Data analysts improve decision making, cut costs and identify new business opportunities. Analysis of data is a process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Therefore, to ensure the accuracy of our analysis, we have to clean our data [12].

1.13.3 **The importance of big data analytics [3]**

Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.

Big data analytics applications enable big data analysts, data scientists, predictive modellers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. That encompasses a mix of semi-structured and unstructured data -- for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the internet of things.

1.14 How Social Media Companies Use Big Data

If there is one thing that social media companies specialize in, it is data. And this they have a lot of it, thanks to their tendency to get users to share information about each waking minute. The large body of data at the disposal of social media companies mirrors how people interact with each other, and at the heart of these interactions lies invaluable information about what individuals and societies hold important. This volume of data, together with the fast rate of data flow for which social media is well known for, represents the essence of **big data**.

By applying **analytics** to social media data, **big data applications in different industries** go beyond the mechanics of interaction to seeing how the content contained in the interactions will affect business performance and people’s view of a brand. **Content analytics** allows companies to zero in on actionable information from the messages that users post. For instance, **analytics** tools can be programmed to track negative or positive sentiment about a brand as this could threaten reputation and revenue [9].

1.14.1 Big data is a driving factor behind every marketing decision

Just like any other industry, social media companies find **big data** useful for analyzing markets and predicting consumer behavior. In 2012, Jay Parikh, engineering VP at Facebook, revealed that Facebook handles over 500 terabytes of data every day, 300 million photos daily, 2.6 billion ‘likes’ and 2.5 billion content uploads. All this data is processed in mere minutes giving Facebook insight into user reactions and the ability to roll out or modify its offering.

What’s more, correlating content to demographics of users age, gender, marital status, geographic location, income levels, educational achievement, inclination to purchase certain products allows a company to know more about the people it's dealing with. Such analysis also reveals how adverts are doing among different customer segments. This is very beneficial as advertisers can react in near real time and adjust campaigns to make more revenue. Analysis of social media data collected by a retailer could for instance reveal that unmarried females between 25 and 35 are suitable candidates for a discount offer on gym equipment.

Based on this information, the retailer might decide to target these candidates with discount offers through Twitter, Facebook and other media. If analytics show the uptake and comments are bad, the offer can be refined to improve performance.

A lot of companies appreciate the powerful nature of social media for personal-level interaction with their customers. Though product customization existed since before social media, the extent and granularity to which it's done by businesses that collect social media data is astounding. Through social media analytics tools, these companies can make data-driven decisions by the minute.

Furthermore, social media analytics tools mean that businesses can look beyond the chatter contained in unstructured data to find meaningful information that can guide decisions and action. Through analysis of statistical data such as impressions per post, audience distribution, interactions on mobile versus desktop, responses (e.g retweets), click-through rates for URLs embeds, and transactional history, a company can measure the effectiveness of its social media **strategy** for promoting brand recognition and loyalty.

Big data also makes it possible to gain insight into the roles people play within social media groups. Users with a large number of followers for instance, can be considered to be influencers. By singling out such people, a company can monitor trends in discussion threads and even participate in such discussions [9].

1.14.2 The Future: Big Data Will Continue to Accelerate the Intrusion of Social Media Companies into People's Privacy

As Facebook, Twitter, Instagram and Pinterest continue to monetize their offerings, it would appear that the benefits that big data will have for social media in the future will become even more personalized. A study published by researchers from Cambridge and Stanford Universities shows that Facebook can use its data to predict people's personality with more accuracy than close friends and families. Every like, share, follow and comment, is data that tells social media companies what you like or dislike, what your actions will be, which cause or brand you will support and what you're likely to buy. Not to mention, any action you take on browsers and search engines today will most likely link back to your social media profile, leaving behind a long trail of digital footprint that can be used for detecting your next moves. This situation will only intensify as people become more reliant on social media platforms for accessing and sharing information [9].

1.15 Conclusion

The availability of Big Data, and new information management and analytic software have produced a unique moment in the history of data analysis. In this chapter, we presented the basic concept of data, then the Big Data, Big Data Analytics, and managing. We had also explain the use of Big Data in social media companies. In the next chapter, we will go deeper in Big Data Analytics espatially by using fuzzy logic.

Chapter II:

Big data analytics using fuzzy logic

2.6 Background

Thousands of products and services are being advertised daily over the web. It is estimated that 75,000 new blogs emerge daily with 1.2 million new posts each day covering many consumer opinions on products and services. Over 40% of people in the modern world depend on opinions and reviews over the web to buy products and apply for various services and express their opinions on diverse issues (Kim 2006, Khan et. al. 2009). As a result of such information wealth, consumers are faced with tough choices to make a rationale decision to select products with good features and competitive prices. In additions suppliers, service providers and manufacturers view customer reviews and opinions about their products and services in order to enhance their products and services. Hence, there is a big demand to develop automated opinion retrieval systems in order to allocate, mine, and classify these thoughts and opinions in some reasonable and presentable fashion (Yao , et al. 2008) [14].

This chapter defines the sentiment analysis and fuzzy logic, highlights the steps that cover the sentiment analysis using fuzzy logic illustrated by an example. Then, a brief description is given related to the method used in our study.

2.7 Sentiment analysis

Sentiment analysis is the determination of a piece of text to be positive, neutral or negative in meaning. When applied to the comments of social media network users, we can attribute their views as being positive, neutral or negative. This is an important piece of detective work when we are an organisation wondering what our customers, or stakeholders are saying about us [13].

2.8 FuzzyLogic

The term “Fuzzy” mean things which are not very clear or vague. In real life, we may come across a situation where we can't decide whether the statement is true or false. At that time, fuzzy logic offers very valuable flexibility for reasoning. We can also consider the uncertainties of any situation.

Fuzzy logic algorithm helps to solve a problem after considering all available data. Then it takes the best possible decision for the given the input. Although, the concept of fuzzy logic had been studied since the 1920's. The term fuzzy logic was first used with 1965 by Lotfi Zadeh a professor of UC Berkeley in California. He observed that conventional computer logic was not capable of manipulating data representing subjective or unclear human ideas [14].

Fuzzy logic has been applied to various fields, from control theory to AI. It was designed to allow the computer to determine the distinctions among data which is neither true nor false. Something similar to the process of human reasoning. Like Little dark, Some brightness, etc [15].

Defining fuzzy sets for such words needs to be based on some expert opinions Since opinions are fuzzy in nature and meaning of opinion words can be interpreted differently, Fuzzy logic is an effective technique to be considered here to properly extract, analyze, categorize and summarize opinions. This due to the following reasons:

- Fuzzy logic is conceptually flexible, easy to understand and it is build to handle imprecise data like opinion words
- Fuzzy logic is based on natural language and hence very suitable to resolve the fuzziness in human expressed phrases .
- Fuzzy logic is an intelligent control Technique which relies on human-like expert knowledge using IF-THEN reasoning rules. Such rules are based on sets that have flexible membership functions rather than just the normal crisp binary logic.
- Fuzzy logic allows better classifications of sentiments with proper strength assigned to each opinion level. This will help to increase the accuracy of classifications [14].

However, fuzzy logic is never a cure for all. Therefore, it is equally important to understand that where we should not use fuzzy logic. Here, are certain situations when we better not use Fuzzy Logic:

- If you don't find it convenient to map an input space to an output space
- Fuzzy logic should not be used when you can use common sense
- Many controllers can do the fine job without the use of fuzzy logic [15].

2.8.1 Architecture

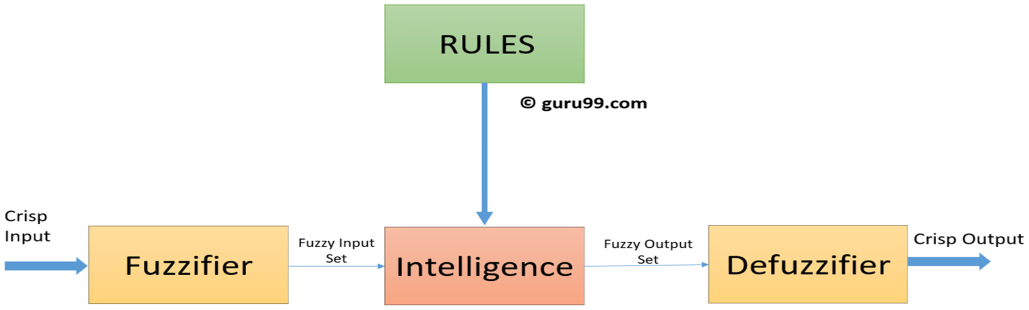


Figure 2.1 : Fuzzy Logic Architecture [15].

Fuzzy Logic architecture has four main parts as shown in the diagram:

- **Rule Base:** It contains all the rules and the if-then conditions offered by the experts to control the decision-making system. The recent update in fuzzy theory provides various methods for the design and tuning of fuzzy controllers. This updates significantly reduce the number of the fuzzy set of rules.
- **Fuzzification:** Fuzzification step helps to convert inputs. It allows us to convert, crisp numbers into fuzzy sets. Crisp inputs measured by sensors and passed into the control system for further processing. Like Room temperature, pressure, etc.
- **Inference Engine:** It helps us to determines the degree of match between fuzzy input and the rules. Based on the % match, it determines which rules need implment according to the given input field. After this, the applied rules are combined to develop the control actions.
- **Defuzzification:** At last the Defuzzification process is performed to convert the fuzzy sets into a crisp value. There are many types of techniques available, so we need to select it which is best suited when it is used with an expert system [15].

2.9 Sentiment analysis using fuzzy logic

2.9.1 Conceptual model of social data

Based on the theory of social data, we present the conceptual model of social data below:

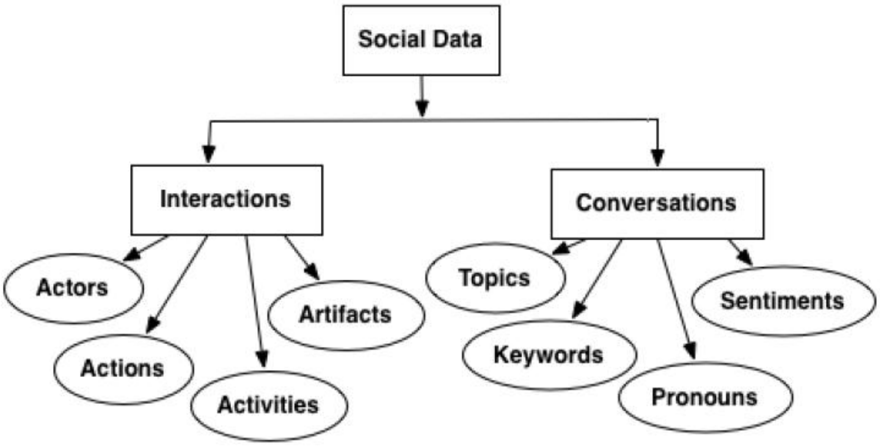


FIGURE 2.2: CONCEPTUAL MODEL OF SOCIAL DATA [16].

In general, Social data consists of two types: Interactions (what is being done) and Conversations (what is being said). Interactions refer to the first aspect of socio-technical interactions constituted by the perception and appropriation of affordances. Conversations relates to the second aspect of socio-technical interactions: structures and functions of technological inter subjectivity. Interactions consists of the structure of the relationships emerging from the appropriation of social media affordances such as posting, linking, tagging, sharing, liking etc. It focuses on identifying the actors involved, the actions they take, the activities they undertake, and the artifacts they create and interact with. Conversations consists of the communicative and linguistic aspects of the social media interaction such as the topics discussed, keywords mentioned, pronouns used and emotions expressed. Figure 2.2 presents the conceptual model of social data [16].

2.9.2 **Research methodology**

The research methodology is shown in Figure 2.3 and is described below:

1) Systematically collect big social data about organizations from Facebook, Twitter etc using the Social Data Analytics Tool [17], [18] developed in the Computational Social Science Laboratory (http://cssl.cbs.dk) and other research and commercial tools.

2) Technically combine organizational process data with business social data so that the resulting dataset legally compliant, ethically correct, privacy adherent, and data security ensured

3) Big Social Data Analytics: Phase One: Adopt current methods, techniques and tools from Computational Social Science to model and analysis.

a) <u>Interaction Analysis</u>:

- Who is doing what, when, where, how and with whom?
- Social media users and organizational stakeholders (like consumers) liking pictures of cute puppies posted by Walmart on its of facial Facebook wall every third Sunday according to its social media marketing strategy.

b) <u>Conversation Analysis:</u>

- What are the things human actors (and fraudulent accounts/robots) saying?
- Social media users and organizational stakeholders (like consumers) commenting on those pictures of cute puppies by discussing/mentioning various topics/keywords of organizational /societal relevance/irrelevance and expressing their subjective feelings etc.

4) Applying set theoretical methods and techniques drawn from crisp sets, fuzzy sets, and rough sets and random sets [19], [20], [21], [22].

5) Software realization of the empirical findings from traditional and novel (set theoretical) approaches to Computational Social Science as tools for Organizations.

6) Publication of research findings in peer-reviewed conferences, journals and edited books.

7) Generation of instrumental bene fits for Organizations in terms of meaningful facts (sensible data), actionable insights (applicable information), valuable outcomes (constructive knowledge) and sustainable impacts (wisdom).

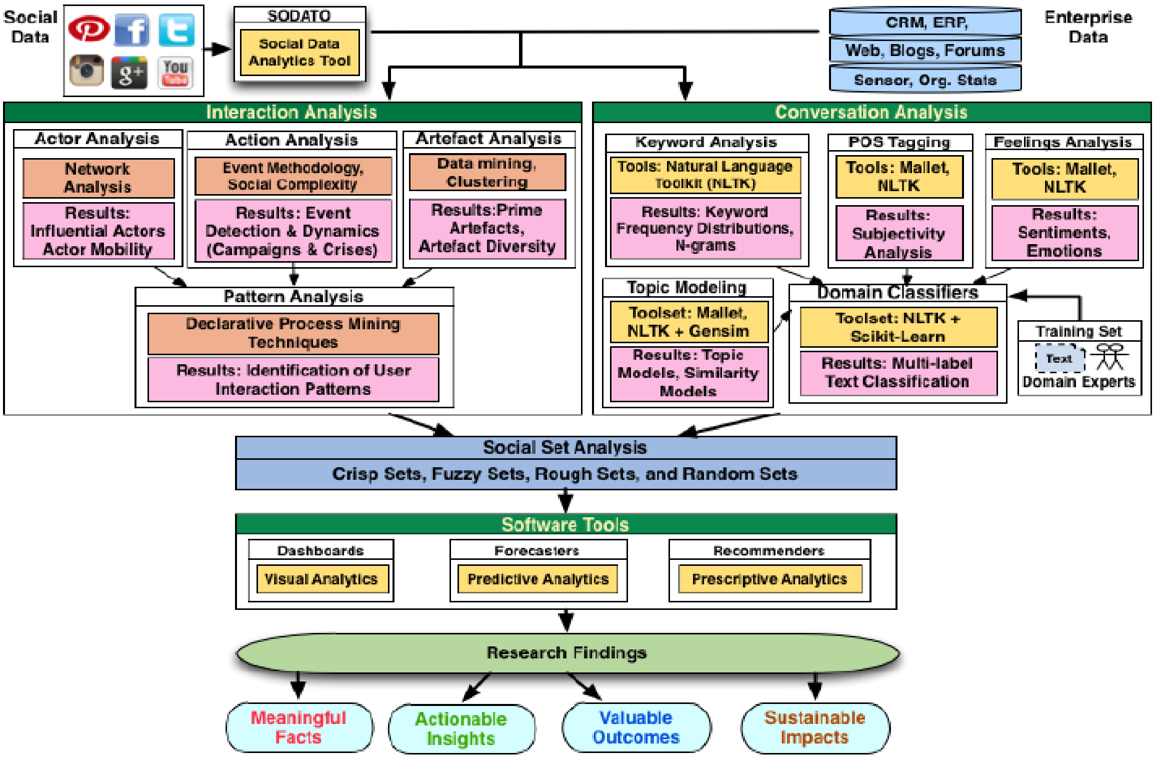

Social Data
SODATO
Social Data Analytics Tool
CRM, ERP,
Web, Blogs, Forums
Sensor, Org. Stats
Enterprise Data
Interaction Analysis
Actor Analysis
Network Analysis
Results: Influential Actors Actor Mobility
Action Analysis
Event Methodology, Social Complexity
Results: Event Detection & Dynamics (Campaigns & Crises)
Artefact Analysis
Data mining, Clustering
Results:Prime Artefacts, Artefact Diversity
Pattern Analysis
Declarative Process Mining Techniques
Results: Identification of User Interaction Patterns
Conversation Analysis
Keyword Analysis
Tools: Natural Language Toolkit (NLTK)
Results: Keyword Frequency Distributions, N-grams
POS Tagging
Tools: Mallet, NLTK
Results: Subjectivity Analysis
Feelings Analysis
Tools: Mallet, NLTK
Results: Sentiments, Emotions
Topic Modeling
Toolset: Mallet, NLTK + Gensim
Results: Topic Models, Similarity Models
Domain Classifiers
Toolset: NLTK + Scikit-Learn
Results: Multi-label Text Classification
Training Set
Text
Domain Experts
Social Set Analysis
Crisp Sets, Fuzzy Sets, Rough Sets, and Random Sets
Software Tools
Dashboards
Visual Analytics
Forecasters
Predictive Analytics
Recommenders
Prescriptive Analytics
Research Findings
Meaningful Facts
Actionable Insights
Valuable Outcomes
Sustainable Impacts

FIGURE 2.3: RESEARCH FRAMEWORK FOR SET-THEORETICAL BIG SOCIAL DATA ANALYTICS [16].

2.9.3 Formal model of the conceptual social data

In this section, we will provide formal semantics for the concepts of social data, which is based on social data model that was initially presented in [38] and [39].

Notation: For a set A we write P (A) for the power set of A (i.e. set of all subsets of A) and P_{disj}(A) for the set of mutually disjoint subsets of A. The cardinality or number of elements in a set A is represented as |A|. Furthermore, we write a relation R from set A to set B as R □A×B. A function {defined from a set A to set B is written as {:A→B, where if f is a partial function then it is written as {:A→B.

First, we define type of artifacts in a socio-technical system as shown in Def. 1[16].

Definition 1: Formally, Social Data is defined as a tuple D = (I, c) where

i) I is the Interactions representing the structural aspects of social data as defined further in Def. 2.

ii) c is the Conversations representing the content of social data and is further defined in Def. 3[16].

Definition 2: The Interactions of Social Data are defined as a tuple I = (U, R, Ac, r_{type}, $\rhd$, $\rightarrow$post, $\rightarrow$share, $\rightarrow$like, $\rightarrow$tag, $\rightarrow$act) where

i) U is a (finite) set of actors (or users) ranged over by u,

ii) R is a (finite) set of artifacts (or resources) ranged over by r,

iii) Ac is the activities set which is also finite,

iv) $r_{type} : R \rightarrow R$ is typing function for artifacts that maps each artifact to an artifact type,

v) $\rhd: R \rightarrow R$ is a partial function mapping artifacts to their parent artifact,

vi) $\rightarrow$post : $U \rightarrow P_{disj}(R)$ is a partial function mapping actors to mutually disjoint subsets of artifacts created by them;

vii) $\rightarrow$share $\subseteq U \times R$ is a relation mapping between users to their artifacts (shared by them),

viii) $\rightarrow$like $\subseteq U \times R$ is a relation mapping users to the artifacts liked by them,

ix) $\rightarrow$tag $\subseteq U \times R \times (P(U \cup Ke))$ is a tagging relation mapping artifacts to power sets of actors and Keywords indicating tagging of actors and keywords in the artifacts, where Ke is set of keywords defined in Def. 3,

x) $\rightarrow$act $\subseteq R \times Ac$ is a relation from artifacts to activities[16].

Definition 3: In Social Data D = (I, c), we define Conversations as C= (To, Ke, Pr, Se, $\rightarrow$topic, $\rightarrow$key, $\rightarrow$pro, $\rightarrow$sen) where

i) To, Ke, Pr, Se are finite sets of topics, keywords, pronouns and sentiments respectively,

ii) $\rightarrow$topic $\subseteq R \times To$ is a relation defining mapping between artifacts and topics,

iii) $\rightarrow$key $\subseteq R \times Ke$ is a relation mapping artifacts to keywords,

iv) $\rightarrow$pro $\subseteq R \times Pr$ is a relation mapping artifacts to pronouns,

v) $\rightarrow$sen $\subseteq R \times Se$ is a realtion mapping artifacts to sentiments[16].

Definition 4: In Social Data, let $T: (u, r, ac) \mapsto N$ be time function that keeps tracks of timestamp ($t \in N$) of an action ($ac \in ACT$) performed by an actor ($u \in U$) on an artifact($r \in R$).

2.9.3.1 **Operational semantics**

Operational semantics of Social Data model are defined how actors perform actions on artifacts. As formally defined in Def. 5, the first action is post, which accepts a pair containing an actor and a new artifact (*u*; *r*). First, the actor will be added to the set of actors (i) and then the new artifact will be added to the set of artifacts (ii). Finally, the post relation (→post) will be updated for the new mapping (iii)[16].

Definition 5: In Social Data D = (I, c) with Interactions I = (U,R, Ac, r_{type}, ⊳,→post , →share, →like, →tag, →act), we define a post operation of posting a new artifact r (r ∉ R) by an user u as D⊕p(u, r) = (I',c) where I' =(U', R', Ac, r_{type}, ⊳, →post', →share, →like, →tag, →act),

i) U' = U ∪ {u}

ii) R' =R ∪ {r}

iii) →post'= $\begin{cases} \rightarrow \text{post (u)} \ \cup \ \{r\} \text{ if } \rightarrow \text{post (u) defined} \\ \qquad \rightarrow \text{post } \cup \{\{u, r\}\} \text{ otherwise} \end{cases}$

Definition 6: In Social Data D = (I, c) with Interactions I =(U,R, Ac, r_{type}, ⊳,→post ,→share, →like, →tag, →act), the comment operation on an artifact r_p (r_p ∈ R) by an user u for a new artifact r is formally defined as D⊕c(u, r, r_p) = (I', c) where I' = (U',R', Ac, r_{type}, ⊳', →post', →share, →like, →tag, →act)

i) D⊕ p(u, r) D (I'', c) where I''= (U',R', Ac, r_{type}, ⊳, →post', →share, →like, →tag, →act),

iii) ⊳ ′= ⊳ ′ ∪ {r, rp}

Definition 7: Let Social Data be D = (I, c) with Interactions I=(U,R, Ac, r_{type}, ⊳, →post , →share, →like, →tag, →act), then we define the share operation on an artifact r by an user u as D ⊕ $_s$(u, r) = (I', c) where I'=(U ∪ {u},R, Ac, r_{type}, ⊳,,→post ,→share ∪ {(u, r)},→like, →tag, →act)[16].

Definition 8: In Social Data D = (I, c) with Interactions I= (U, R, Ac, r_{type}, ⊳ →post, →share, →like, →tag, →act), we define the like operation by an user u on an artifact r as

D⊕$_l$ (u, r) = (I', c) where I'= (U ∪ {u}, R, Ac, r_{type}, ⊳, →post, →share, →like ∪ {(u, r)}, →tag, →act).

Similarly, we define the unlike operation on D=(I, c) with Interactions I = (U,R, Ac, r_{type}, ⊳, →post ,→share, →like, →tag, →act), as D⊕$_l$(u, r)=(I', c) where I' = (U,R, Ac, r_{type}, ⊳,→post ,→share, →like n {(u, r)},→tag, →act)[16].

Definition 9: In a Social Data D= (I, c) with Interactions I= (U,R, Ac, r_{type},⊳,→post ,→share, →like, →tag, →act), we define the tagging operation by an user u on an artifact r with a set of hash words t ∈ P(U ∪ Ke) as D $\oplus_t$ (u, r, t)= (I', c) where I' = (U ∪ {u},R, Ac, r_{type}, ⊳, →post ,→share, →like, →tag ∪ {(u, r, t)},→act)[16].

2.9.3.2 **Illustrative example**

In this section, we exemplify the formal model by taking an example post from the Facebook page of McDonald's.

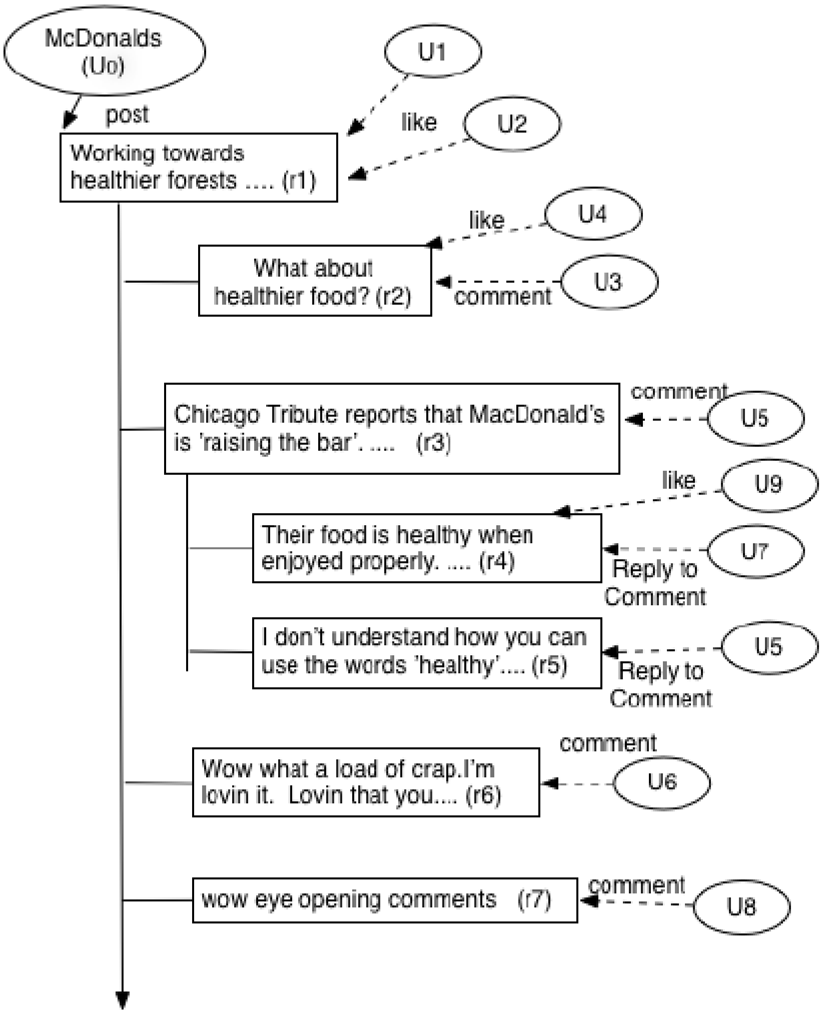

McDonalds (U0)
post
U1
like
U2
Working towards healthier forests .... (r1)
like
U4
What about healthier food? (r2)
comment
U3
Chicago Tribute reports that MacDonald's is 'raising the bar'. .... (r3)
comment
U5
like
U9
Their food is healthy when enjoyed properly. .... (r4)
Reply to Comment
U7
I don't understand how you can use the words 'healthy'.... (r5)
Reply to Comment
U5
comment
Wow what a load of crap.I'm lovin it. Lovin that you.... (r6)
U6
wow eye opening comments (r7)
comment
U8

Figure 2.4: Facebook post example in formal model[16].

As shown in the figure 2.4. In order to enhance the readability of the example, the artifacts have been annotated as r1, r2 etc. and the annotated values will be used in encoding the example using the formal model. The following are some of the texts extracted from a sample post [54] from Facebook page of McDonald's Food/Beverages.

r1 = Working towards healthier forests through more sustainable packaging.

r2 = what about healthier food?

r3 = Chicago Tribute reports that MacDonald's is `raising the bar'. You mean bars with nails in them to beat live chickens with. MacDonald's is one big lie. Do not believe them. Next, they will tell you their food is healthy.

r4 = their food is healthy when enjoyed properly. Their beef is amazing and that is what they move a lot of. The fattier menu items, if you have any modicum of a pallet, you will notice are sides and not to be enjoyed in such an amount as whole meals themselves, but hey, I know some people who think raw sugar is a treat.

r5 = I do not understand how you can use the words `healthy' and MacDonald's in the same sentence. They manufacture (and I use that word deliberately) to have a perfect balance of salt, sugar and fat to hook children with their `Happy Meals'. Sorry Keith, but healthy does not contain GMO's, Factory Farmed Animals, Chicken beaks, feathers etc, wood cellulose, fat, sugar and salt.

r6 = Wow what a load of crap. I love it. Loving that you are losing business and closing stores.

r7 = wow eye opening comments[16].

The example shown in Fig. 2.4 can be encoded as follows, the social Data D= (I, c) contains two components:

I = (U, R, Ac, r_{type}, $\rhd$, ⟶post,⟶share,⟶like, ⟶tag, ⟶act) is the Interactions and c = (To, Ke, Pr, Se, ⟶topic, ⟶key, ⟶pro, ⟶sen) is the Conversations.

Initially, let us assume that the sets of activities, topics, keywords, pronouns and sentiments will have the following values.

Ac = {promotion},

To = {healthy food, sustainable packaging},

Ke = {healthy, sustainable, beef, chicken, . .}

Pr = {We, I}, Se = {c, 0, -},

U = {u0, u1, :::}

R = {r1}

⟶act = {(r1, promotion)}

Post action by u0

D $\oplus$ p(u0, r1) = D1 = (I1,c) where I1 = (U1,R1, Ac, r_{type}, $\rhd$,⟶post 1,⟶share,⟶like,⟶tag, ⟶act) with the following values U1 = U∪ { u0 }, R = R∪ { r1 } and ⟶post1=⟶post∪ {(u0, { r1 })} like action by u2 and u1 Let's imagine that the post was liked by user u2 first and then liked by user u1. D1$\oplus_l$ (u2, r1) $\oplus_l$ (u1, r1) = D2 = (I2, c) where I2 = (U2, R, Ac, r_{type}, $\rhd$, ⟶post 1, ⟶share, ⟶like 1, ⟶tag, ⟶act) with the following values

$U2 = U1\cup \{u2\} \cup \{u1\}$, and $\longrightarrow$like 1 $=\longrightarrow$like$\cup \{(u2, r1), (u1, r1)\}$

Comment action by u5 on the post r1

Let us imagine that the user u5 posted a comment (r3) on the Facebook post and let D3 be the social data before the comment action.

D3 $\oplus$ C(u5, r3, r1) = D4 = (I4,c) where I4 = (U4,R3,B1, r_{type}, Ac, $\longrightarrow$post 3,$\longrightarrow$share,$\longrightarrow$like 1, $\longrightarrow$tag, $\longrightarrow$act) with the following values $U3 = U3\cup \{ u5 \}$, $R3 = R2\cup \{ r3 \}$, $\longrightarrow$post 3 $=\longrightarrow$post 2$\cup \{(u5, \{ r3 \})\}$ and $B 1 = B\cup \{(r3, r1)\}$.

Reply to comment by u7 on the comment r3 etc.

The rest of the operations shown in Fig. 2.4 can expressed similarly in the formal model.

2.9.4 Fuzzy set based sentiment analysis

First, we will recall necessary basic definitions of Fuzzy sets [51].

Definition 10: If X is a set of elements denoted by x, then a fuzzy set A over X is defined as a set of ordered pairs $A = \{(x, \mu A(x)) | x \in X)\}$ where $\mu_A : X \longrightarrow [0, 1]$ is the membership function. Each member or element of a fuzzy set A is mapped to real number between 0 and 1 ([0, 1]), which represents the degree of membership of an element in the fuzzy set. A membership value of 1 indicates full membership, while a value of 0 indicates no membership[16].

Definition 11: The support of a fuzzy set A is a crisp set of all $x \in X$ such that $\mu A(x) > 0$. The crisp set of elements that belongs to fuzzy set A at least to a degree μ is called α-level or α-cut is defined as $A_\alpha = \{ x \mid x \in X \wedge \mu_A(x) > \alpha\}$[16].

Definition 12: In Social Data = D (I, c), fuzzy Interactions is defined as a tuple I = (U, R, Ac, r_{type}, ⊳, $\longrightarrow$post, $\longrightarrow$share, $\longrightarrow$like, $\longrightarrow$tag, $\longrightarrow$act) where

i) U,R, Ac, r_{type}, ⊳, $\longrightarrow$post , $\longrightarrow$share, $\longrightarrow$like, $\longrightarrow$tag are same as defined in 1,

ii) $\longrightarrow$act $= \{((r, a),\mu\longrightarrow act\ (r, a)) \mid r \in R, a \in Ac\}$ is a fuzzy relation mapping artifacts to activities with membership function $\mu_{\rightarrow act} : R \times Ac \longrightarrow [0, 1]$

Definition 18: In Social Data = D (I, c) we define fuzzy Conversations as c = (To, Ke, Pr, e, $\longrightarrow$topic, $\longrightarrow$key, $\longrightarrow$pro, $\longrightarrow$sen) where

i) To, Ke, Pr, Se are the sets of topics, keywords, pronouns and sentiments respectively as defined in 2.

ii) $\longrightarrow$topic $= \{((r, to),\mu_{\rightarrow topic}\ (r, to))| \ r \in R, to \in To\}$ is a Fuzzy relation mapping artifacts to topics with membership function $\mu_{\ \rightarrow topic} : R \times To \longrightarrow [0, 1]$,

iii) $\longrightarrow$key $= \{((r, ke),\mu_{\rightarrow key}\ (r, ke)) \mid r \in R, ke \in Ke\}$ is a Fuzzy relation mapping artifacts to keywords with membership function $\mu_{\ \rightarrow key} : R \times Ke \longrightarrow [0, 1]$,

iv) ⟶pro = {((r, pr), $\mu_{\rightarrow pro}$ (r, pr)) | r ∈ R, pr ∈ Pr} is a Fuzzy relation mapping artifacts to pronouns with membership function $\mu_{\rightarrow pro} : R \times Pr \longrightarrow [0, 1]$,

v) ⟶sen = {((r, se), $\mu_{\rightarrow sen}$ (r, se)) | r ∈ R, se ∈ Se} is a Fuzzy relation mapping artifacts to sentiments with membership function $\mu_{\rightarrow sen} : R \times Se \longrightarrow [0, 1]$[16].

2.9.4.1 **Illustrative example**

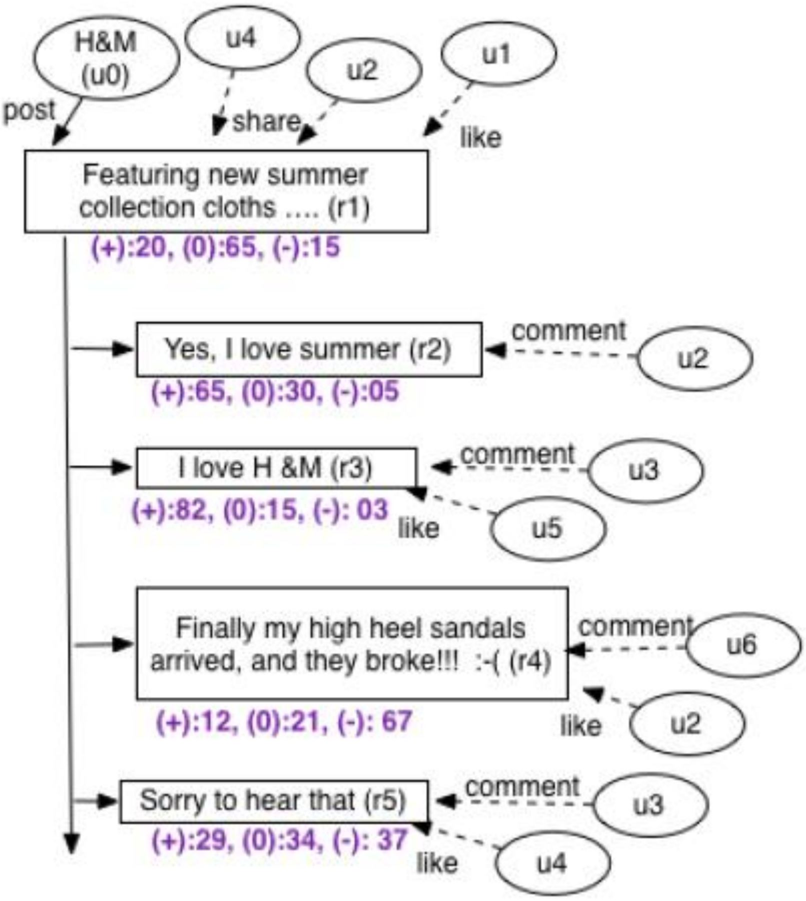


Figure 2.5: Example in formal model[16].

In this section, we will exemplify the formal model with fuzzy sets by taking an example post from the Facebook page of H&M cloth stores as shown in the figure 2.5. In order to enhance the readability of the example, the artifacts (e.g. texts) have been annotated as r1, r2 etc. and the annotated values will be used in encoding the example using the formal model[16].

Moreover, as our focus is to mainly to demonstrate sentiment analysis, we will abstract away from the details of the sets which are not directly involved in the sentiment analysis. As shown in Figure 2.5, the sentiments of the artifacts (e.g. (+).20, (0).65, (-).15) are represented in the boxes below the artifacts.

Example 2: The example shown in Fig. 2.5 will be encoded as follows,

D = (I, c) where I is the Interactions and c is the Conversations.

Initially, the sets of actors, artifacts and other relations have the following values.

U = {u0, u1, u2, u3, u4, u5, u6...}

R = {r1, r2, r3, r4, r5 …}

⊳ = {(r2, r1), (r3, r1), (r4, r1), (r5, r1)...}

⟶post = {(u0, {r1...}), (u2, {r2}), (u3, {r3, r5}), (u6, {r4})...}

⟶share = {(u4, r1), (u2, r1)…}

⟶like = {(u1, r1), (u5, r3), (u2, r4), (u4, r5)...}

Se = {+, 0, -}

After the artifacts are analyzed for the sentiments, the sentiment relation becomes a fuzzy set contain the pairs of artifacts and sentiment labels with the sentiment score as membership value as shown below,

⟶sen = {((r1,+), 0.20),((r1, 0), 0.65),((r1,-), 0.15),((r2,+), 0.65),((r2, 0), 0.30),((r2,-), 0.05),((r3,+), 0.82),((r3, 0), 0.15),((r3,-), 0.03),((r4,+), 0.12),((r4, 0), 0.21),((r4,-), 0.67),((r5,+), 0.29),((r5, 0), 0.34),((r5,-), 0.37)}

From the sentiment fuzzy set, one can extract different crisp sets (R_{α}^{se}) for artifacts based different values of µ-cuts.

For example for a value of µ = 0.4, the artifact sets for + and - will be

$\mathrm{R}_{\alpha=\,0.40}^{+}$ = {r2, r3} and $|\mathrm{R}_{\alpha=0.40}^{+}|$=2

$R_{\alpha=\,0.40}^{-}$ {r4} and $|R_{\alpha=0.40}^{-}|$=1

On the other hand, if someone wants a fine-grained analysis of the data, they could use a lower value for µ- cut, which will include more elements into the analysis.

$R_{\alpha=\,0.20}^{+}$ = {r1, r2, r3, r5} and $|R_{\alpha=0.20}^{+}|$ = 4

$R_{\alpha=\,0.20}^{-}$ = {r4, r5} and $|R_{\alpha=0.20}^{-}|$ = 2

Similarly, we can also compute the actor sets ($U_{R_{\alpha}^{se}}$) that are associated with the artifact sets as follows.

$U_{R_{\alpha=\,0.40}^{+}}$ = {u2} ∪ ∅ ∪ ∅ ∪ ∅ ∪{u3} ∪ ∅ ∪ ∅ ∪ {u5}

= {u2, u3, u5}

$U_{R_{\alpha=\,0.40}^{-}}$ = {u6} ∪ ∅ ∪ ∅ ∪{u2} ∪ ∅ ∪ ∅ ∪ {u4}

= {u6, u2, u3, u4}

Notice that, here we have an advantage due to fuzzy set modelling that an can an actor can be present in more than one set (e.g. $U_{R_{\alpha=\,0.2}^{+}}$ and $U_{R_{\alpha=\,0.2}^{-}}$), as an actor can express more than one sentiment by performing the actions on artifacts in reality. When once crisp sets for artifacts (R_{α}^{se}).

2.9.4.2 **Inferred sentiment and actor profiling:**

The inferred sentiment for actors can be calculated in the similar line as above. In this example, we will show how one can compute inferred sentiment for the actor u2, where we take union of fuzzy sets containing artifacts with sentiment labels for the artifacts posted, shared and liked by actor u2 as follows.

u_2^+ = {((r2, +), 0.65)} ∪ {((r1, +), 0.20)} ∪ {((r4, +), 0.12)}
= {((r2, +), 0.65), ((r1, +), 0.20), ((r4, +), 0.12)}
u_2^- = {((r2, -), 0.05)} ∪ {((r1, -), 0.15)} ∪ {((r4, -), 0.67)}
= {((r2, -), 0.05), ((r1, -), 0.15), ((r4, -), 0.6)}

After computing the fuzzy sets as above, one could apply µ-cut with the required granularity to get crisp sets similar to the sentiment analysis of the artifacts. After that many such sets can be computed for a given time intervals and can be plotted on a time scale to analyze how the sentiment of an actor varies in the time frame[16].

2.9.4.3 **Conversation analysis**

Google Prediction API [53] was utilized in order to calculate sentiments for the posts and comments on the wall. Configuration for computation of sentiment began with the setting up a model which was trained with the manually labelled data subset from the H&M data corpus fetched by SODATO[18][17]. This training dataset consisted of 11,384 individual posts and comments randomly selected from H&M data corpus and their corresponding sentiment labels as coded by five different student analysts. Training data was labeled Positive, Negative or Neutral. The sentiment results for each individual post/comment returned by the Google Prediction API were [16].

Sentiment	α_cuts				
	≥0.1	≥0.3	≥0.5	≥0.7	≥0.9
+	17,752	25,949	30,343	25,869	19,974
-	9,166	14,503	16,577	13,494	10,397
0	12,566	21,607	26,826	24,312	21,830
+∩-	5,661	5,184	2,067	1,489	913
+∩0	16,674	14,401	8,550	7,069	4,673

-∩0	10,017	9,984	6,541	5,381	3,892
+∩-∩0	39,001	19,209	6,512	4,567	2,739
Total artifacts	110,837	110,837	97,416	82,181	64,418

TABLE 2.5 : PARENT ARTIFACT(POST) SENTIMENT DISTRIBUTION[16].

Sentiment	α_cuts				
	≥0.1	≥0.3	≥0.5	≥0.7	≥0.9
+	36,114	57,653	77,551	80,540	83,378
-	19,433	32,472	42,145	35,388	28,310
0	37,511	62,037	85,334	80,404	79,981
+∩-	28,788	33,929	49,315	109,785	237,156
+∩0	94,094	99,339	119,822	141,158	297,516
-∩0	54,756	56,527	50,176	44,660	35,520
+∩-∩0	16,537,774	13,810,588	11,477,670	10,189,815	7,742,858
Total artifacts	16,808,470	14,152,545	11,902,013	10,681,750	8,504,719

TABLE 2.6 : TOTAL ARTIFACT (POSTS + COMMENTS + LIKES) SENTIMENT DISTRIBUTION[16].

Sentiment	α_cuts				
	≥0.1	≥0.3	≥0.5	≥0.7	≥0.9
+	331,891	441,290	549,159	563,600	555,964
-	211,783	311,861	382,082	317,912	199,815
0	1,074,602	1,335,933	1,469,989	1,413,921	1,168,176
+∩-	67,496	92,901	111,438	76,491	51,868
+∩0	647,315	712,828	523,046	511,667	508,537
-∩0	411,821	248,707	149,532	122,645	66,889
+∩-∩0	979,718	581,106	400,186	338,565	231,158
Total artifacts	3,724,626	3,724,626	3,585,432	3,344,801	2,782,407

TABLE 2.7: ACTORS SENTIMENT WITH DIFFERENT A- CUTS[16].

2.9.4.4 Finding

Compared to existing sentiment analysis methods and tools in academia and industry, the set theory and fuzzy set theory approach that we demonstrated in the tables (1, 2 and 3) and figures (2.6, 2.7 and 2.8) above reveal the longitudinal sentiment profiles of actors and artefacts for the entire corpus[16].

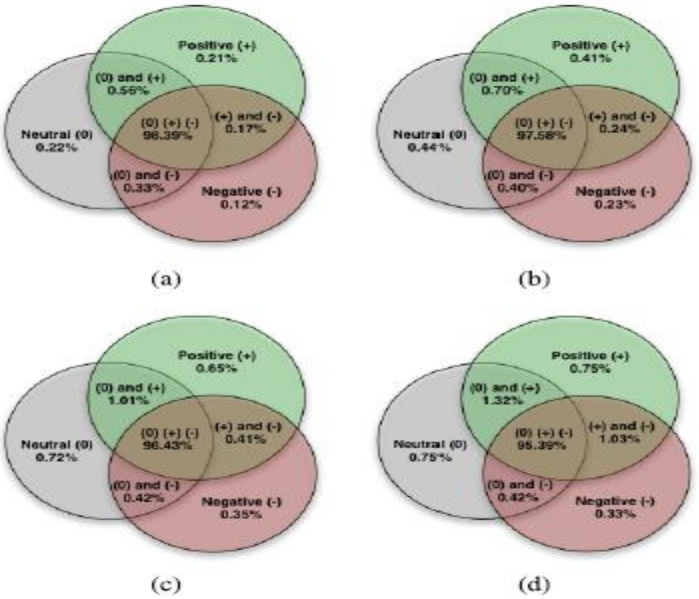


FIGURE 2.6 : ARTIFACT SENTIMENTS. (A) A≥ 0.1. (B) A≥ 0.3. (C) A≥ 0.5. (D) A≥ 0.7[16].

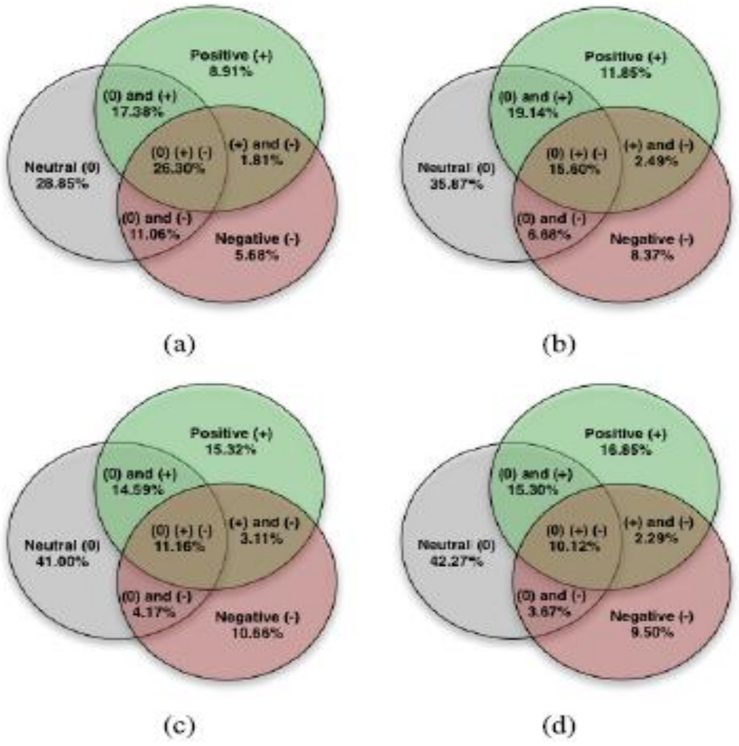


FIGURE 2.7: ACTOR SENTIMENTS. (A) A ≥ 0.1. (B) A ≥ 0.3. (C) A≥ 0.5. (D) A≥ 0.7[16].

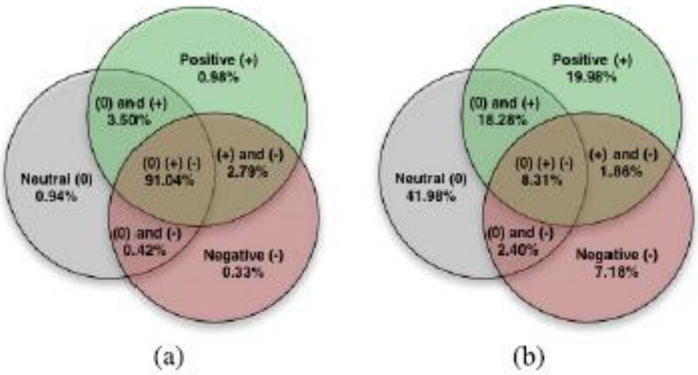


FIGURE 2.8 :ARTIFACT AND ACTOR SENTIMENTS FOR A≥ 0.9. (A) ARTIFACT SENTIMENTS A≥ 0.9. (B) ACTOR SENTIMENTS A≥ 0.9[16].

2.10 Conclusion

One of the contributions of this chapter is to demonstrate the suitability and effectiveness of fuzzy logic for analyzing big social data from content-driven social media platforms like Facebook. In the last chapter we will describe our fuzzy logic model, its implementation.

Chapter III:

Implementation of Fuzzy Inference System

3.3 Background

The implementation of big data analysis is the most important phase of our study. In the literature, we found many simulation environments and network simulators that are available for network performance measurement. In our study, we selected the MATLAB environment because it is very simple and has easy ways to create FIS (**F**uzzy **I**nference **S**ystem).

In this chapter, we briefly present the MATLAB environment as well as the developed interface that we used for this implementation

3.3 **MATLAB environnement**

MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. Using the MATLAB product, we can analyze a huge volume of data faster than with traditional ways.

We can use MATLAB in a wide range of applications, including signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology. Add-on toolboxes (collections of special-purpose MATLAB functions, available separately) extend the MATLAB environment to solve particular classes of problems in these application areas.

MATLAB provides a number of features for documenting and sharing work. We can integrate MATLAB code with other languages and applications [40].

MATLAB Features include:

- ✓ High-level language for technical computing,
- ✓ Development environment for managing code, files, and data,
- ✓ Interactive tools for iterative exploration, design, and problem solving,
- ✓ Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration,
- ✓ 2-D and 3-D graphics functions for visualizing data,
- ✓ Tools for building custom graphical user interfaces,
- ✓ Functions for integrating MATLAB based algorithms with external applications and languages, such as C, C++, FORTRAN, Java™, COM, and Microsoft Excel.

3.4 Data set description

In our study, we use a data set downloading from a web site [41]; it is treat a topic, which is popular recently. It has a high volume of data: 23 posts and 10755 comments. For each post, using Facebook Graph API, all comments have been collected during the first 30000 s. Data is stored in flat table format (e.g. CSV file) which is easy to save in distributed file system. The header of CSV file contains the following columns: [Data time], [Topic], [Post], [Comment], [Positive], [Negative].

The topic was chosen is “United States presidential election 2016”which is popular recently [42], as we show in the figure 3.1

3.5 Fuzzy Inference Systems

3.5.1 Definition

Fuzzy inference is the process of formulating input/output mappings using fuzzy logic. Fuzzy Logic Toolbox software provides command-line functions and an app for creating Mamdani and Sugeno fuzzy systems[43].

Fuzzy inference systems have been successfully applied in fields such as automatic control, data classification, decision analysis, expert systems, and computer vision[44].

3.5.2 Architecture

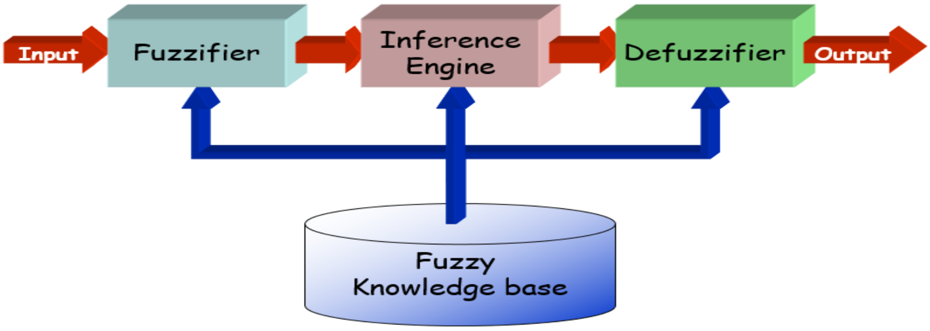


FIGURE3.1: THE ARCHITECTURE OF FUZZY INFERENCE SYSTEMS[44].

The steps of fuzzy reasoning (inference operations upon fuzzy IF–THEN rules) performed by FISs are:

1. Compare the input variables with the membership functions on the antecedent part to obtain the membership values of each linguistic label. (this step is often called fuzzification.)
2. Combine (usually multiplication or min) the membership values on the premise part to get firing strength (deree of fulfillment) of each rule.
3. Generate the qualified consequents (either fuzzy or crisp) or each rule depending on the firing strength.

4. Aggregate the qualified consequents to produce a crisp output.
5. Defuzzification [44] .

3.6 Building Systems with the Fuzzy Logic Toolbox

There are five primary GUI tools for building, editing, and observing fuzzy inference systems in the Fuzzy Logic Toolbox: the Fuzzy Inference System or FIS Editor, the Membership Function Editor, the Rule Editor, the Rule Viewer, and the surface viewer. These GUIs are dynamically linked, in that changes you make to the FIS using one of them, can affect what you see on any of the other open GUIs. You can have any or all of them open for any given system.

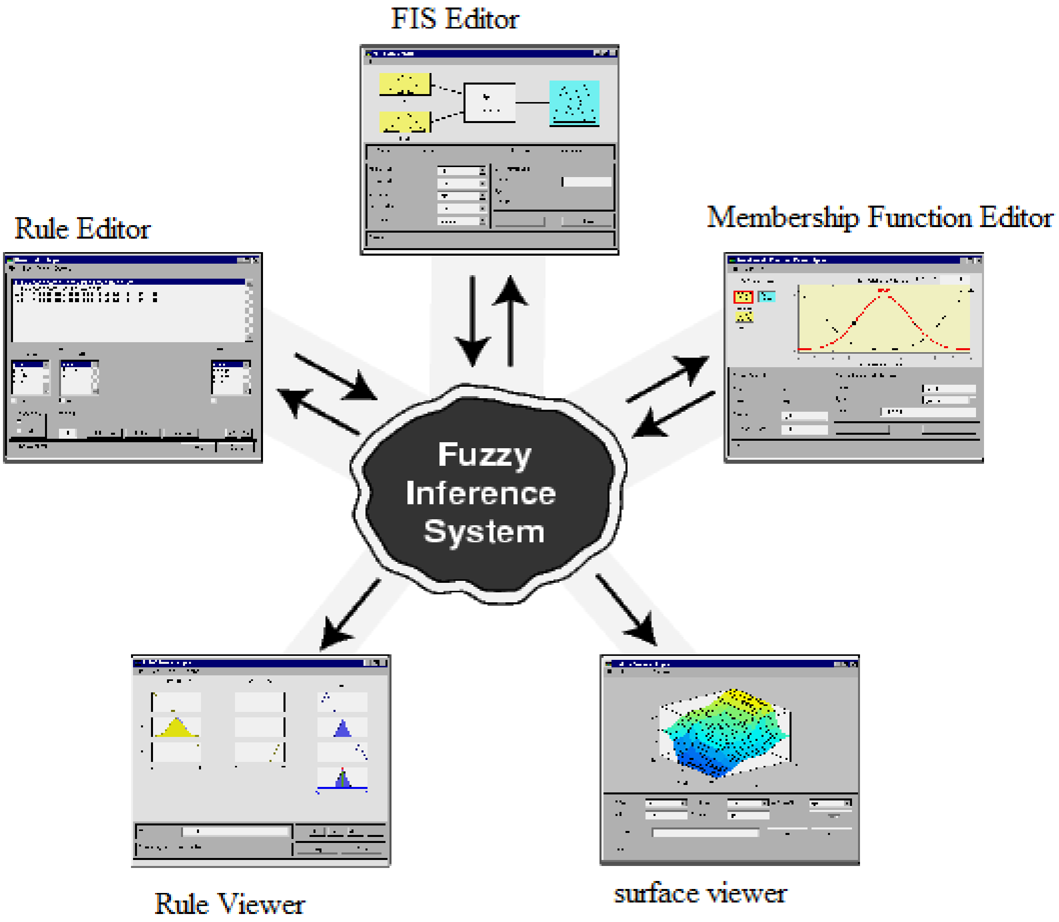


FIGURE 3.2: BUILDING SYSTEM WITH FUZZY LOGIC TOOLBOX [54]

The FIS Editor handles the high-level issues for the system: How many input and output variables? What are their names? The Fuzzy Logic Toolbox doesn't limit the number of inputs. However, the number of inputs may be limited by the available memory of your machine. If the number of inputs is too large, or the number of membership functions is too big, then it may also be difficult to analyze the FIS using the other GUI tools.

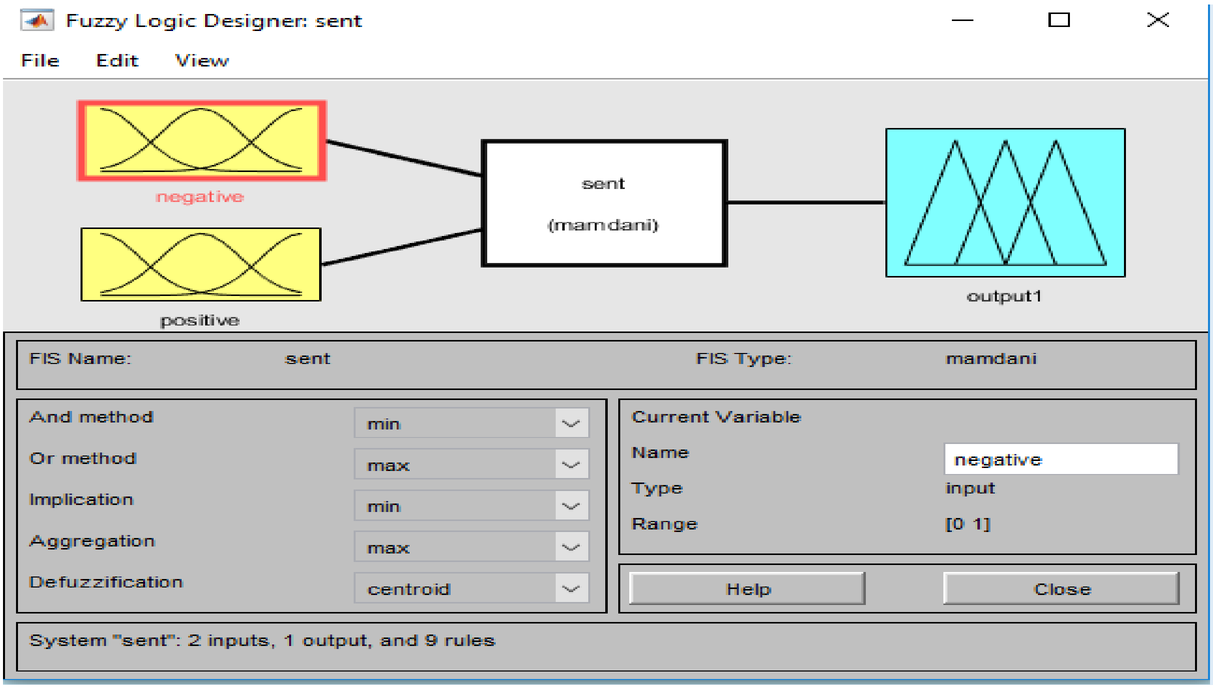
The Membership Function Editor is used to define the shapes of all the membership functions associated with each variable.

The Rule Editor is for editing the list of rules that defines the behavior of the system.

The Rule Viewer and the Surface Viewer are used for looking at, as opposed to editing, the FIS. They are strictly read-only tools. The Rule Viewer is a MATLAB based display of the fuzzy inference diagram shown at the end of the last section. Used as a diagnostic, it can show (for example) which rules are active, or how individual membership function shapes are influencing the results. The Surface Viewer is used to display the dependency of one of the outputs on any one or two of the inputs that is, it generates and plots an output surface map for the system.

3.7 Implementation with Fuzzy Inference System

In our implementation, we used MATLAB environment to implement and simulate the Fuzzy Inference System as shown in the figure 3.2, we has as inputs a matrix of the negative and the positive columns from data set which has downloading



FIS Name: sent

FIS Type: mamdani

And method: min

Or method: max

Implication: min

Aggregation: max

Defuzzification: centroid

Current Variable

Name: negative

Type: input

Range: [0 1]

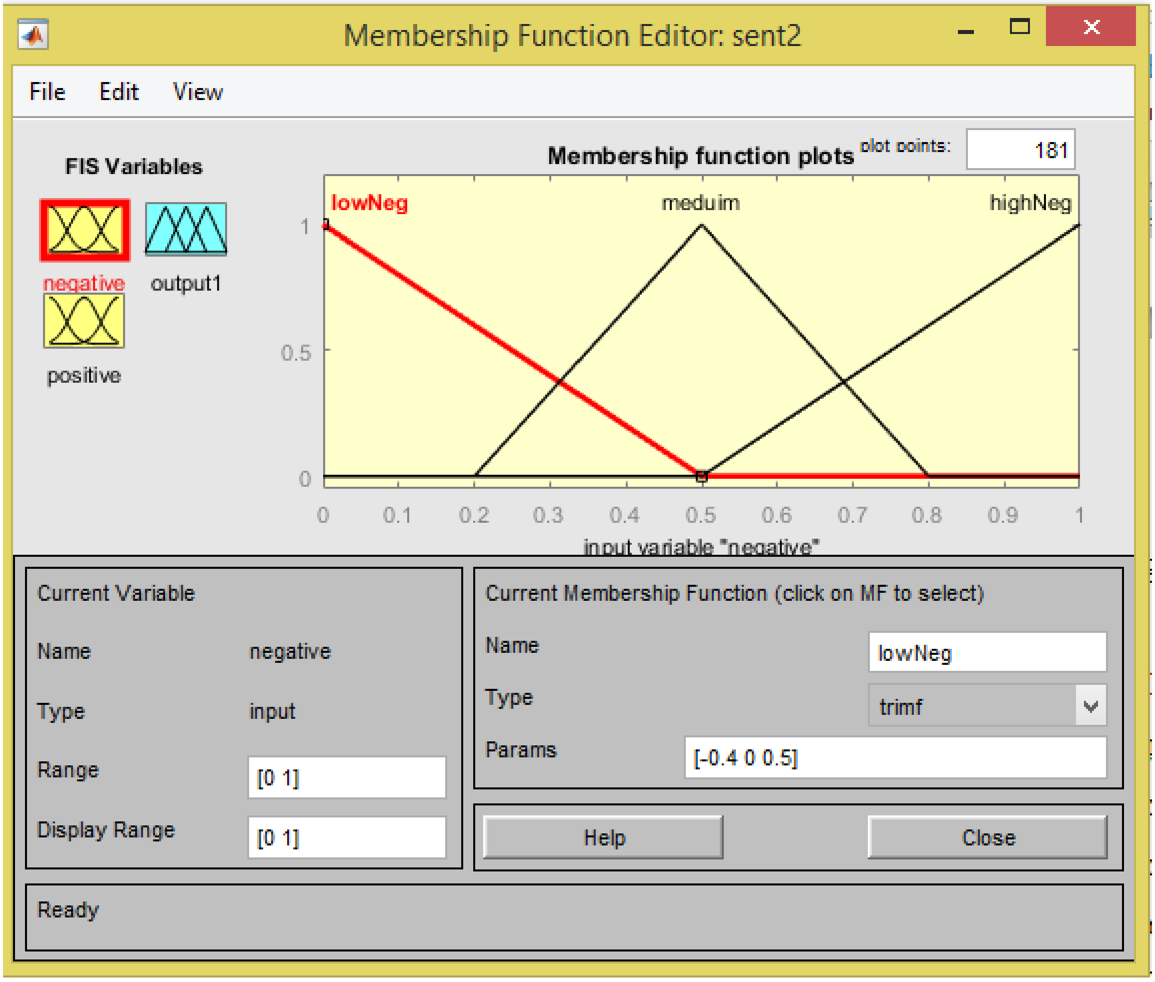
System "sent": 2 inputs, 1 output, and 9 rules

FIGURE 3.3: IMPLEMENTATION OF FIS.

3.7.1 Fuzzify Inputs

The first step is to take the inputs and determine the degree to which they belong to each of the appropriate fuzzy sets via membership functions, the input is always a crisp numerical value. The output is a fuzzy degree of membership in the qualifying linguistic set[45].

Our implementation in fuzzy inference system is built on two inputs which are: positive and negative (as shown in figure 3.4); nine rules and each of the rules depends on resolving the inputs into several different fuzzy linguistic sets: positive score is low, positive score is medium, or positive score is high, and the same for negative input. For output we have five fuzzy linguistic sets: Verry_Happy, Happy, Neutral, Mad, Verry_Mad . Before the rules can be evaluated, the inputs and ouput must be fuzzified according to each of these linguistic sets.



Membership Function Editor: sent2

plot points: 181

Current Variable

Name negative

Type input

Range [0 1]

Display Range [0 1]

Current Membership Function (click on MF to select)

Name lowNeg

Type trimf

Params [-0.4 0 0.5]

Ready

FIGURE 3.4: MEMBERSHIP FUNCTION EDITOR.

The rules was defined as follows

1. If (negative is LowNeg)and(positive is LowPos)then(output is Very Mad)
2. If (negative is LowNeg)and(positive is MediumPos)then(output is Happy)
3. If (negative is LowNeg)and(positive is HighPos)then(output is Very Happy)
4. If (negative is MediumNeg)and(positive is LowPos)then(output is Mad)
5. If (negative is MediumNeg)and(positive is MediumPos)then(output is Neutral)
6. If (negative is MediumNeg)and(positive is HighPos)then(output is Happy)

7. If (negative is HighNeg)and(positive is LowPos)then(output is Very Mad)
8. If (negative is HighNeg)and(positive is MediumPos)then(output is Mad)
9. If (negative is HighNeg)and(positive is HighPos)then(output is Neutral)

3.7.2 Apply Fuzzy Operator

Any number of well-defined methods can fill in for the AND operation or the OR operation. In the toolbox you can create your own methods for AND and OR by writing any function and setting that to be your method of choice.

The following figure shows the And operator *min* at work, evaluating the antecedent of the rule 4 for the output calculation. The two different pieces of the antecedent (negative score is meduim and positive score is low) yielded the fuzzy membership values 0.5 and 0.3 respectively. The fuzzy And operator simply selects the minimum of the two values, 0.3, and the fuzzy operation for rule 4 is complete. The probabilistic And method would still result in 0.3.

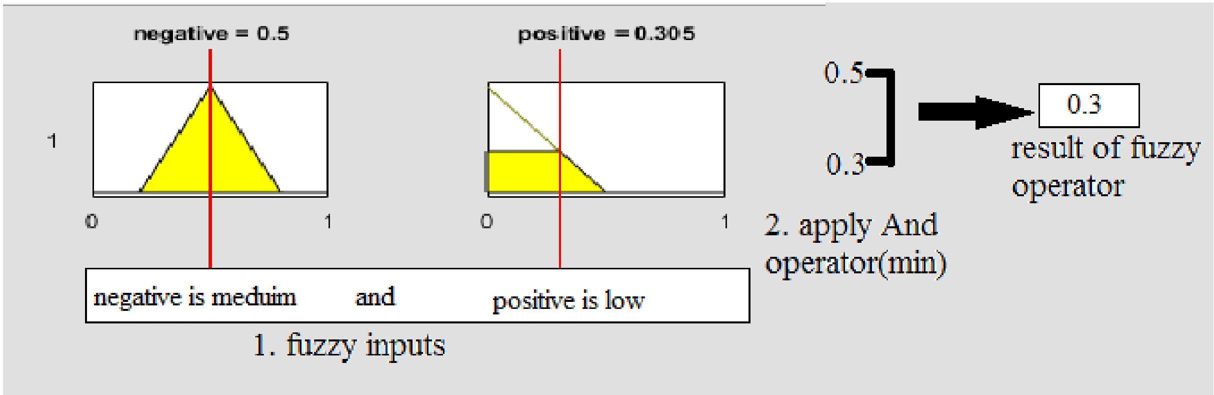


FIGURE 3.5: FUZZY OPERATOR.

3.7.3 Apply implication method

Before applying the implication method, you must determine the rule weight. Every rule has a *weight* (a number from 0 through 1), which is applied to the number given by the antecedent. Generally, this weight is 1 and thus has no effect on the implication process. However, you can decrease the effect of one rule relative to the others by changing its weight value to something other than 1 [45].

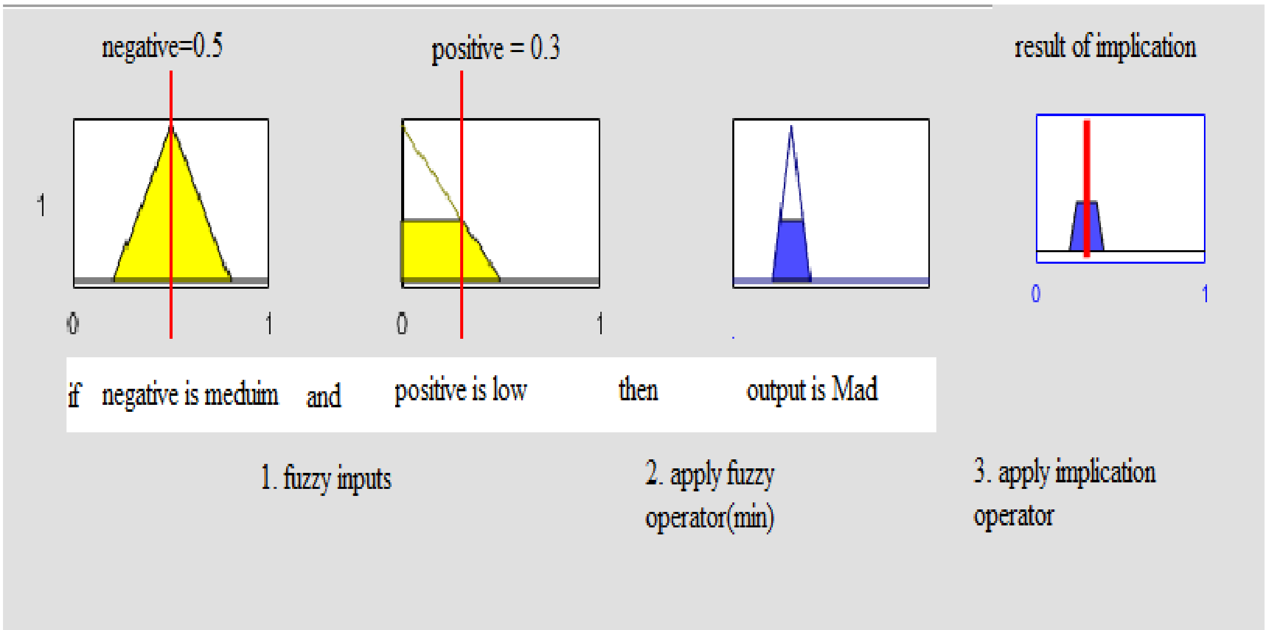

negative=0.5
positive = 0.3
result of implication
1
0
1
0
1
0
1
if negative is meduim and positive is low then output is Mad
1. fuzzy inputs
2. apply fuzzy operator(min)
3. apply implication operator

FIGURE 3.6: IMPLICATION METHOD[45].

3.7.4 Defuzzification

Defuzzification is the process of obtaining a single number from the output of the aggregated fuzzy set. It is used to transfer fuzzy inference results into a crisp output [45].

There are five built-in defuzzification methods supported: centroid, bisector, middle of maximum (the average of the maximum value of the output set), largest of maximum, and smallest of maximum. Perhaps the most popular defuzzification method is the centroid calculation, which returns the center of area under the curve, as shown in the following:

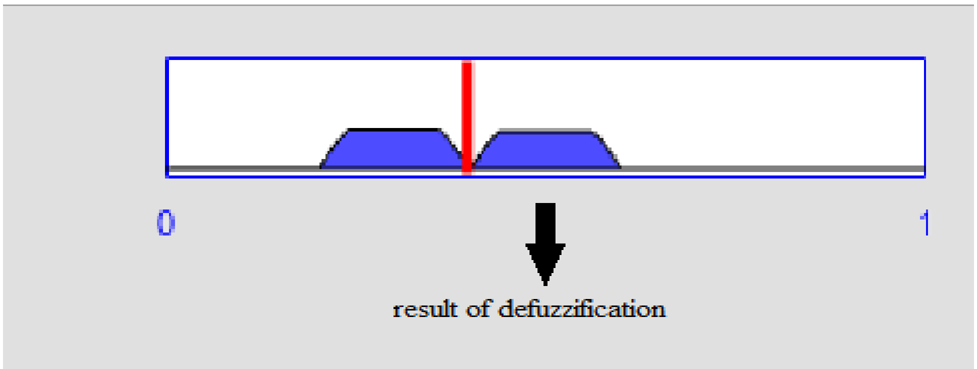
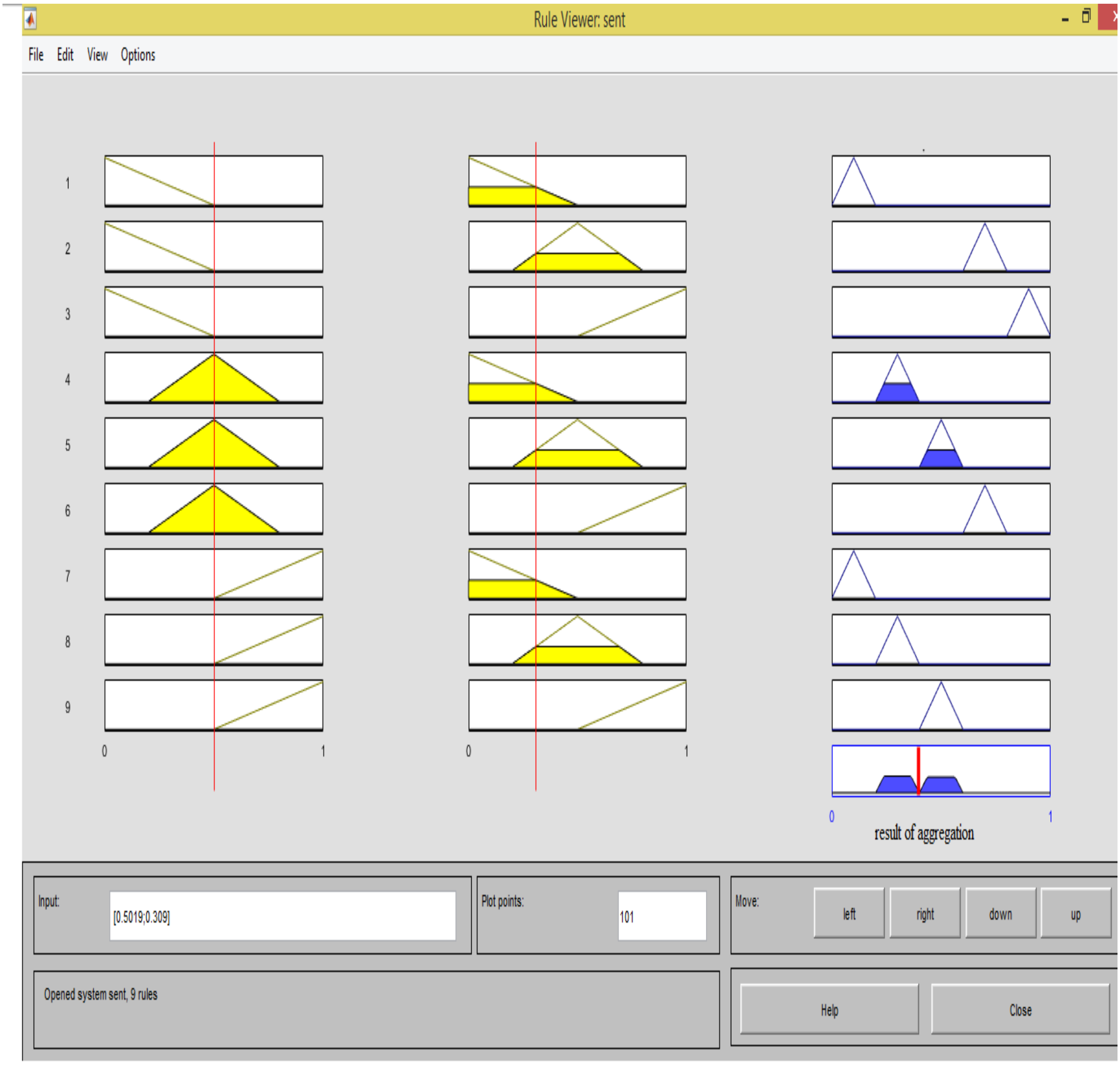


FIGURE 3.7: DEFUZZIFICATION.

3.7.5 Fuzzy Inference Diagram

The fuzzy inference diagram is the composite of all the smaller diagrams presented so far in this section. It simultaneously displays all parts of the fuzzy inference process you have examined. Information flows through the fuzzy inference diagram as shown in the following:



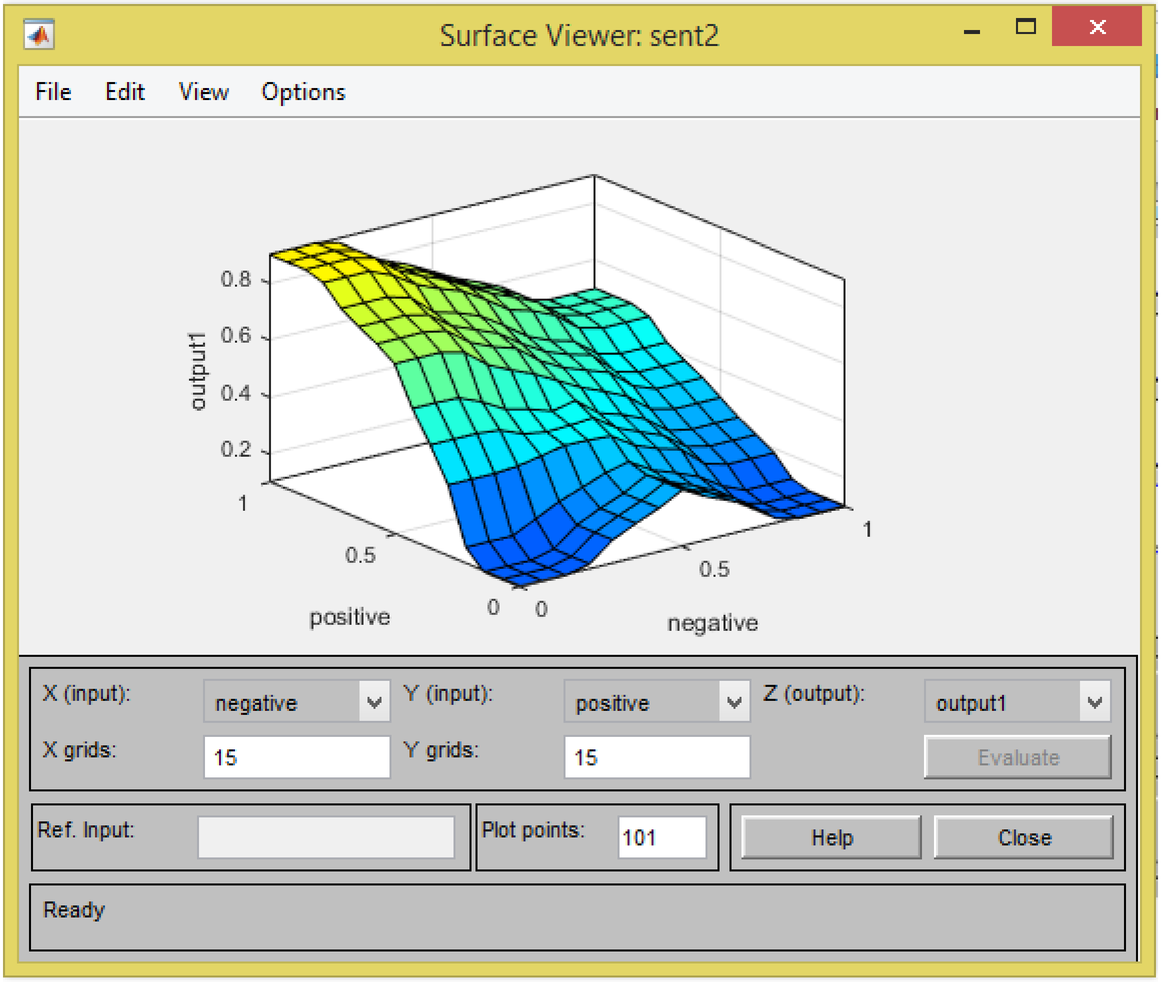
Input: [0.5019;0.309]

Plot points: 101

Opened system sent, 9 rules

FIGURE 3.8: RULES VIEWER EDITOR.

In the following figure shown the surface viewer which represent the dependency of output and the inputs.



X (input): negative
Y (input): positive
Z (output): output1
X grids: 15
Y grids: 15
Ref. Input:
Plot points: 101
Ready

FIGURE 3.9: SURFACE VIEWER EDITOR.

3.8 Conclusion

In this chapter we described a fuzzy inference system which an important part of fuzzy logic. In most practical applications such systems perform crisp nonlinear mapping, which is specified in the form of fuzzy rules encoding expert or common-sense knowledge about the problem at hand.

In the next chapter we present the different analysis and their corresponding result.

Chapter IV:

Results and analysis

4.1 Background:

The growth in number of fuzzy logic applications led to the need of finding efficient ways to implement them. In the previous chapter we implement the fuzzy inference system, and determinate the inputs, output, and creating rules.

In this chapter we present the different analyses scenarios and their corresponding results. At the end, we present the analyses and comparison of the different results in order to extract the best one.

4.2 Data analysis results

In order to analysis the result of our FIS model, we create four rules groups with four FISs we change only number of rules where the first one contains nine rules, the second contains four rules, and the two others contain three rules, as shown in the following :

Rule Group 1:

1. If (negative is LowNeg)and(positive is LowPos)then(output is Very Mad)
2. If (negative is LowNeg)and(positive is MediumPos)then(output is Happy)
3. If (negative is LowNeg)and(positive is HighPos)then(output is Very Happy)
4. If (negative is MediumNeg)and(positive is LowPos)then(output is Mad)
5. If (negative is MediumNeg)and(positive is MediumPos)then(output is Neutral)
6. If (negative is MediumNeg)and(positive is HighPos)then(output is Happy)
7. If (negative is HighNeg)and(positive is LowPos)then(output is Very Mad)
8. If (negative is HighNeg)and(positive is MediumPos)then(output is Mad)
9. If (negative is HighNeg)and(positive is HighPos)then(output is Neutral)

Rule Group 2:

1. If (negative is HighNeg)and(positive is LowPos)then(output is Very Mad)
2. If (negative is MediumNeg)and(positive is LowPos)then(output is Mad)
3. If (negative is MediumNeg)and(positive is MediumPos)then(output is Neutral)
4. If (negative is LowNeg)and(positive is MediumPos)then(output is Happy)

Rule Group 3:

1. If (negative is HighNeg)and(positive is MediumPos)then(output is Mad)
2. If (negative is HighNeg)and(positive is HighPos)then(output is Neutral)
3. If (negative is MediumNeg)and(positive is HighPos)then(output is Happy)

<u>Rule Group 4:</u>

1. If (negative is MediumNeg) and (positive is LowPos) then (output is Mad)
2. If (negative is LowNeg) and (positive is LowPos) then (output is Neutral)
3. If (negative is LowNeg) and (positive is HighPos) then (output is Very Happy)

4.2.1 Analysis related to the first rule group:

After analyzing the 23 posts with the rule group one, we present the results in a relative circles as the shown the following figure:

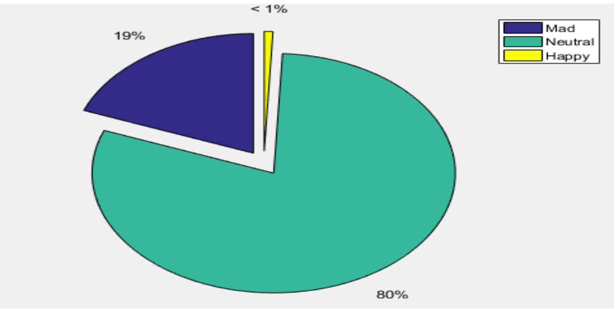


FIGURE 4.1: RESULT ANALYSIS FOR POST1

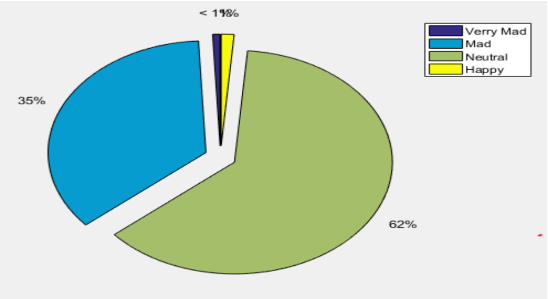


FIGURE 4.2: RESULT ANALYSIS FOR POST 2

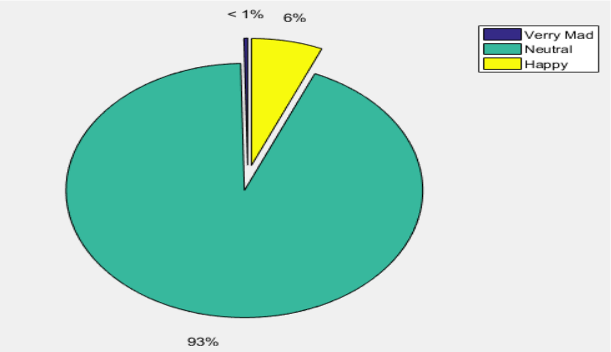


FIGURE 3.3: ANALYZE RESULT FOR POST 3

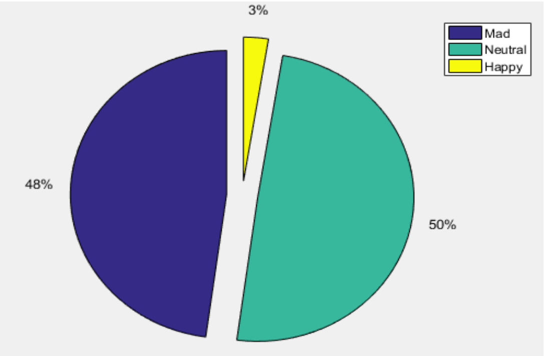


FIGURE 4.4: ANALYZE RESULT FOR POST 4

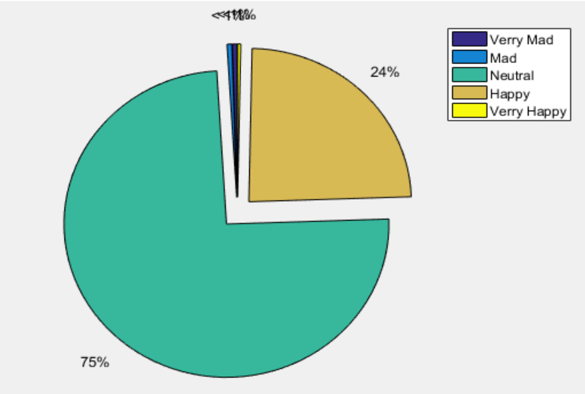


FIGURE 4.5: RESULT ANALYSIS FOR POST 5

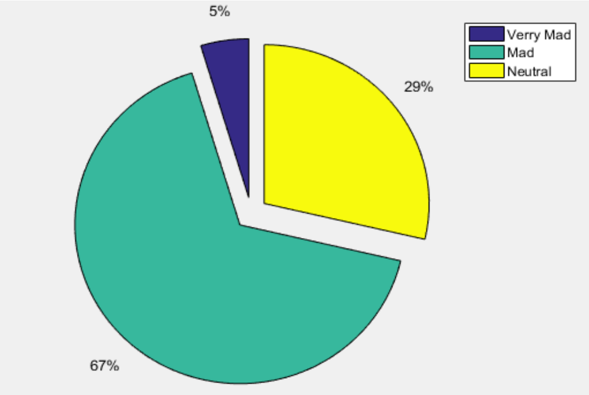


FIGURE 4.6: RESULT ANALYSIS FOR POST 6

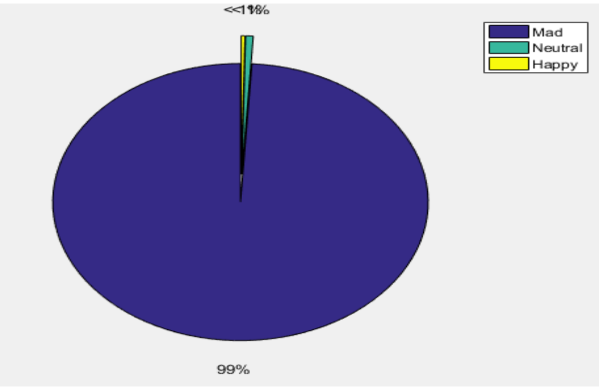


FIGURE 4.7: RESULT ANALYSIS FOR POST 7

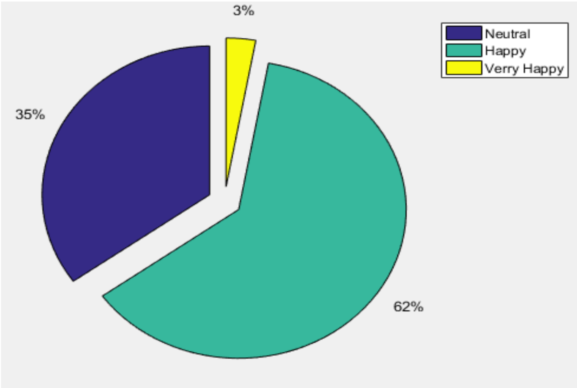


FIGURE 4.8: RESULT ANALYSIS FOR POST 8

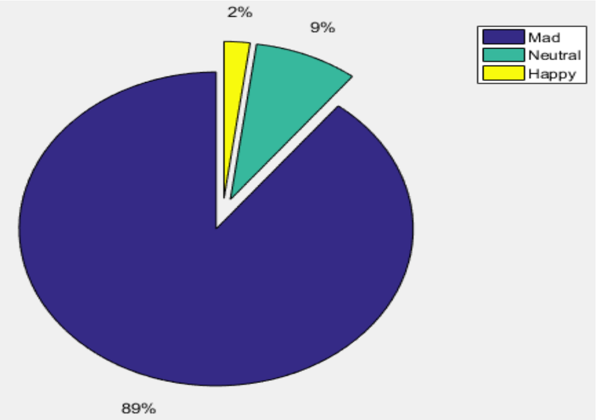


FIGURE 4.9: RESULT ANALYSIS FOR POST 9

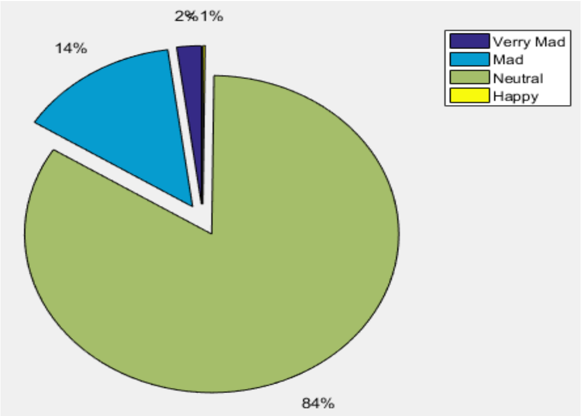


FIGURE 4.10: RESULT ANALYSIS FOR POST 10

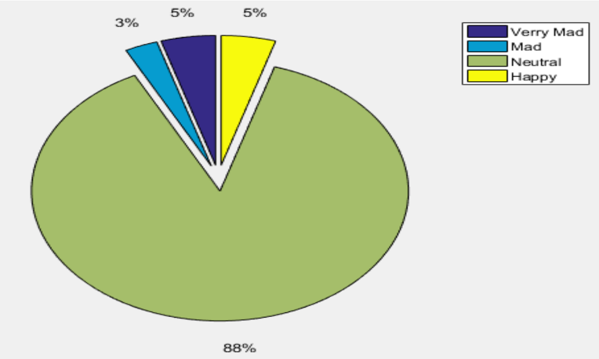


FIGURE 4.11: RESULT ANALYSIS FOR POST 11

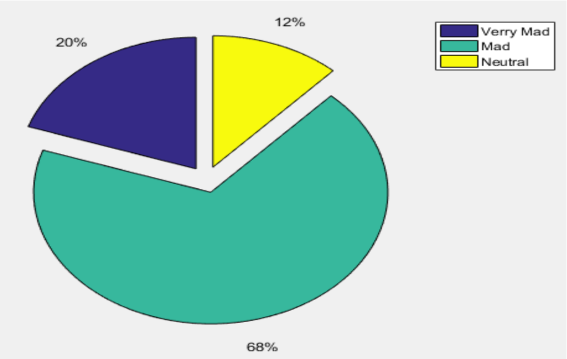


FIGURE4. 12: RESULT ANALYSIS FOR POST 12

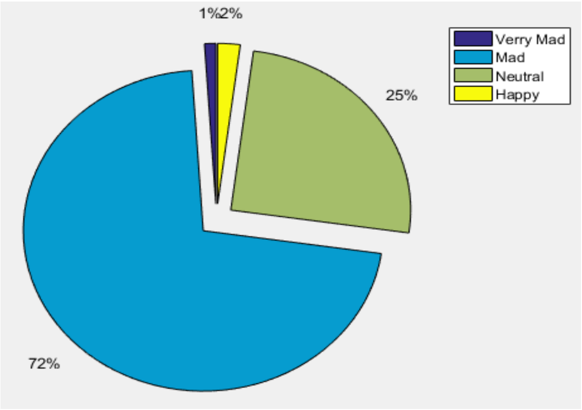


FIGURE 4.13: RESULT ANALYSIS FOR POST 13

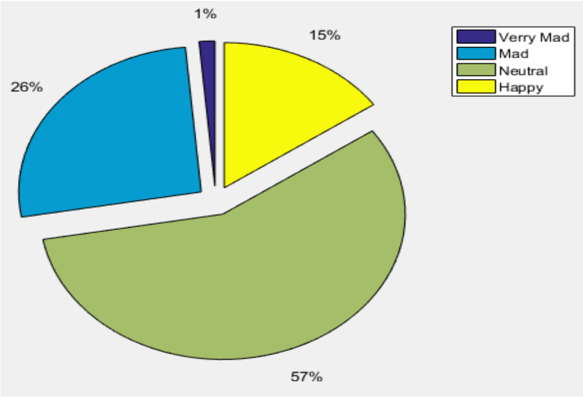


FIGURE4. 14: RESULT ANALYSIS FOR POST 14

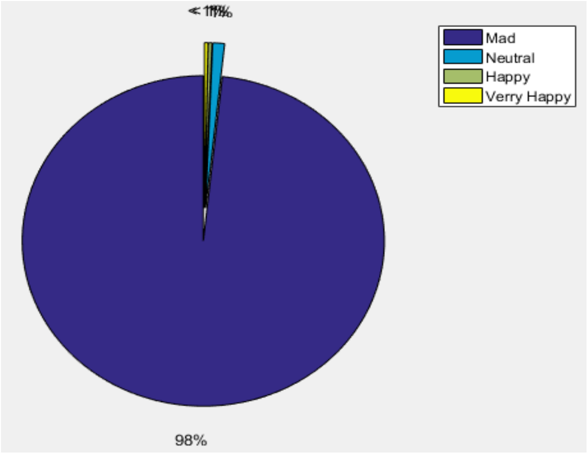


FIGURE 4.15: RESULT ANALYSIS FOR POST 15

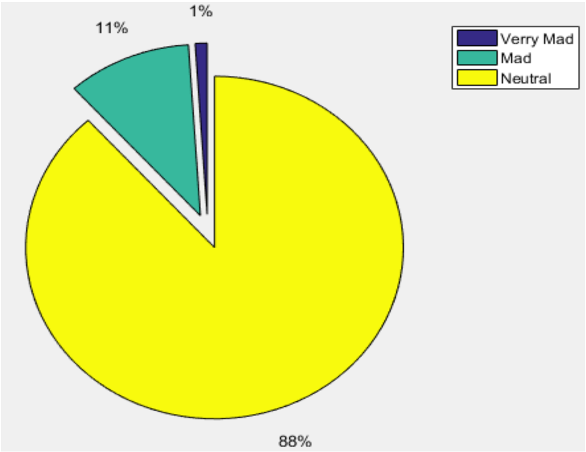


FIGURE 4.16: RESULT ANALYSIS FOR POST 16

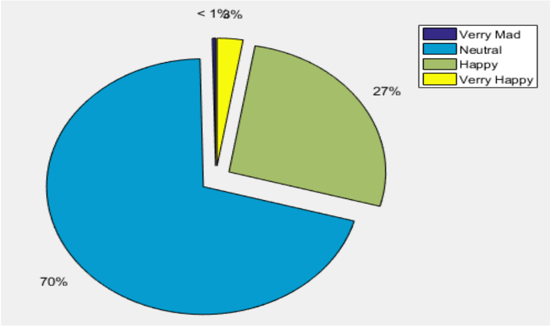


FIGURE 4.17: RESULT ANALYSIS FOR POST 17

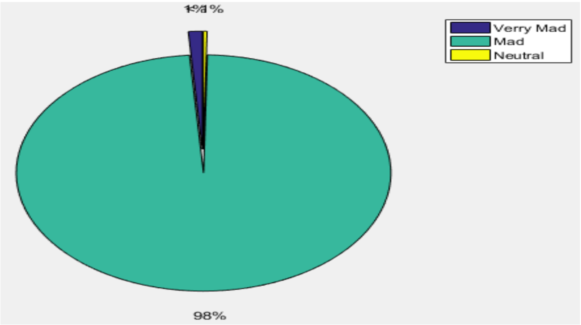


FIGURE 4.18: RESULT ANALYSIS FOR POST 18

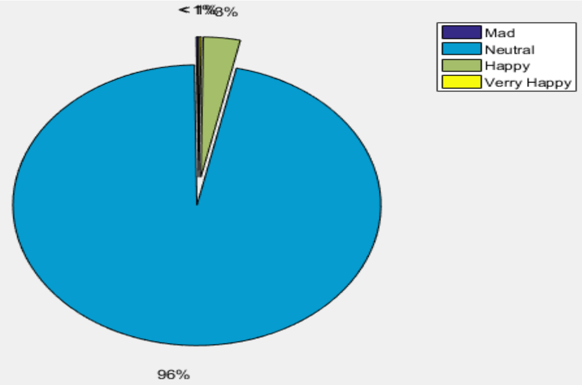


Figure 4.19: Result analysis for post 19

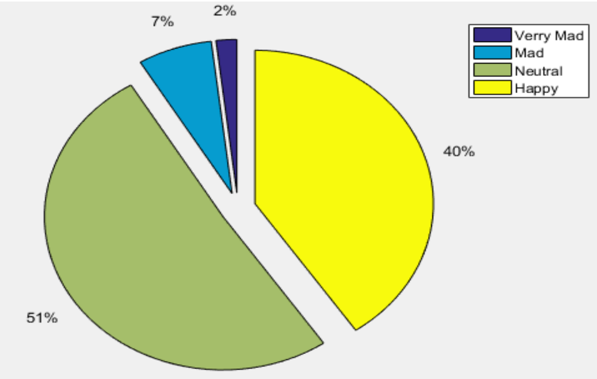


Figure 4.20: Result analysis for post 20

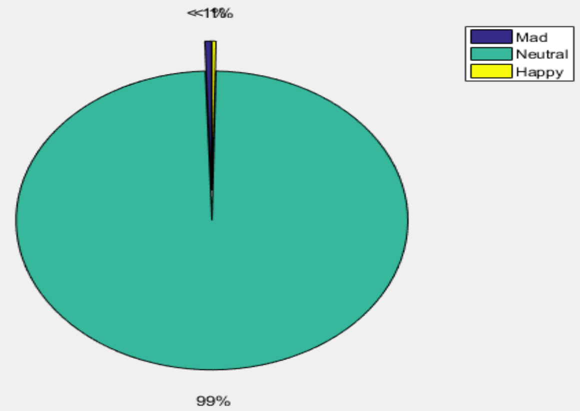


Figure 4.21: Result analysis for post 21

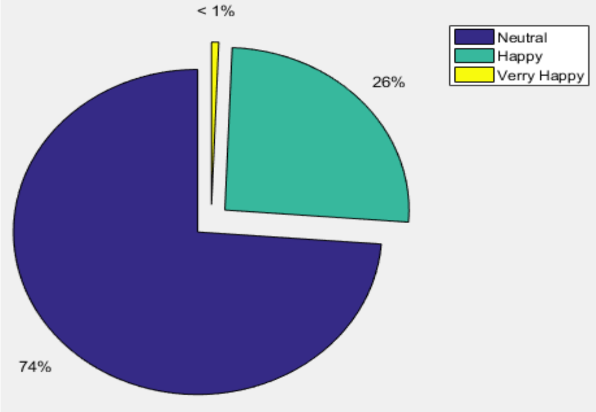


Figure 4.22: Result analysis for post 22

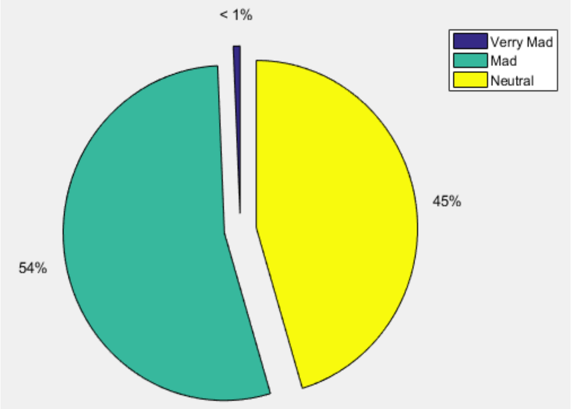


Figure 4.23: Result analysis for post 23

4.2.2 Comparison the results detected by FL with the Results detected by expert:

In other side, we had divided the 23 Fb posts into 6 post groups:

- Post Group 1: has four posts.
- Post Group 2: has four posts.
- Post Group 3: has four posts.

- Post Group 4: has four posts.
- Post Group 5: has four posts.
- Post Group 6: has three posts

We gave each post group to five experts in order to get their opinions. Then, we gathered the opinions and compared them with our results detected by fuzzy logic.

A model of evaluation that we had given to the experts in order to get opinions is shown as follows:

<table>
<tr><td colspan="30">National poll: Sanders and Clinton neck-and-neck
استطلاع وطني: ساندرز وكلينتون</td></tr>
<tr>
<td colspan="3">Hillary will always win anything close.
ستفوز هيلاري دائمًا بأي شيء قريب.</td>
<td colspan="3">Go Bernie, Anybody But Clinton!
اذهب بيرني ، أي شخص ما عدا كلينتون!</td>
<td colspan="3">Their necks look a lot alike, too
تبدو رقابهم متشابهين كثيرًا أيضًا</td>
<td colspan="3">That means they are both doing a good job.
هذا يعني أن كلاهما يقوم بعمل جيد.</td>
<td colspan="3">National Poll results: "is this all that we have to offer America?"</td>
<td colspan="3">I do not believe it.</td>
<td colspan="3">The best comedy team of the season...</td>
<td colspan="3">VOTE FOR MR TRUMP if you love this country.</td>
<td colspan="3">Neither of them will win.</td>
<td colspan="3">HILLARY GO HOME AND WASH DISHES</td>
</tr>
<tr><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td><td>H</td><td>M</td><td>N</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

Table 4.8 a model of expert evaluation

In each post group, we presented each post with 10 comments chosen randomly. After reading the post and their comments, the expert had to decide if the comment is Happy, Mad or neutral by putting a cross under his choice.

After that, we calculated the percentage of each choice and we gathered them in tables as shown below:

1. Post group1

Experts	Posts											
	P1			P2			P3			P4		
	H	N	M	H	N	M	H	N	M	H	N	M
Expert 1	60%	10%	30%	30%	0%	70%	0%	50%	50%	40%	30%	30%

Expert 2	40%	0%	60%	20%	0%	80%	20%	20%	60%	20%	30%	50%
Expert 3	30%	30%	40%	40%	20%	40%	20%	40%	40%	40%	20%	40%
Expert 4	40%	10%	50%	30%	0%	70%	10%	40%	50%	30%	30%	40%
Expert 5	60%	10%	30%	20%	0%	80%	20%	50%	30%	40%	30%	30%

Table 4.2: Opinion of experts for post group 1.

2. Post group2

Experts	Posts											
	P5			P6			P7			P8		
	H	N	M	H	N	M	H	N	M	H	N	M
Expert 1	20%	20%	60%	20%	30%	50%	30%	40%	30%	40%	30%	30%
Expert 2	10%	10%	80%	20%	30%	40%	50%	20%	30%	20%	10%	70%
Expert 3	30%	20%	50%	20%	30%	50%	20%	20%	60%	20%	0%	80%
Expert 4	70%	10%	20%	60%	20%	20%	60%	20%	10%	30%	30%	40%
Expert 5	0%	60%	40%	10%	10%	80%	30%	10%	60%	60%	10%	30%

Table 4.3: Opinion of experts for post group 2.

3. Post group 3

Experts	Posts											
	P9			P10			P11			P12		
	H	N	M	H	N	M	H	N	M	H	N	M
Expert 1	50%	40%	10%	50%	30%	20%	40%	40%	20%	40%	10%	50%
Expert 2	30%	10%	60%	20%	40%	40%	40%	20%	40%	50%	30%	20%

Expert 3	20%	30%	50%	0%	70%	30%	10%	50%	40%	10%	40%	50%
Expert 4	10%	30%	60%	10%	70%	20%	0%	80%	20%	10%	50%	40%
Expert 5	10%	20%	70%	0%	60%	40%	10%	70%	20%	0%	30%	70%

Table 4.4: Opinion of experts for post group 3.

4. **Post group 4**

Experts	Posts											
	P13			P14			P15			P16		
	H	N	M	H	N	M	H	N	M	H	N	M
Expert 1	0%	40%	60%	20%	30%	50%	40%	30%	30%	10%	50%	40%
Expert 2	40%	10%	50%	50%	10%	40%	40%	20%	40%	40%	40%	20%
Expert 3	10%	20%	70%	10%	60%	30%	10%	20%	70%	30%	60%	10%
Expert 4	10%	10%	80%	10%	60%	30%	20%	10%	70%	20%	70%	10%
Expert 5	30%	20%	50%	30%	10%	60%	50%	30%	20%	20%	60%	20%

Table 4.5: Opinion of experts for post group 4.

5. **Post group 5**

Experts	Posts											
	P17			P18			P19			P20		
	H	N	M	H	N	M	H	N	M	H	N	M
Expert 1	10%	50%	40%	0%	20%	80%	0%	70%	30%	10%	60%	30%
Expert 2	50%	20%	30%	30%	10%	60%	30%	30%	40%	60%	10%	30%
Expert 3	10%	50%	40%	20%	10%	70%	20%	50%	30%	10%	60%	30%

Expert 4	20%	50%	30%	20%	30%	50%	30%	50%	20%	20%	60%	20%
Expert 5	20%	60%	20%	20%	20%	60%	20%	60%	20%	30%	50%	20%

Table 4.6: Opinion of experts for post group 5.

6. Post group 6

Experts	Posts								
	P21			P22			P23		
	H	N	M	H	N	M	H	N	M
Expert 1	50%	20%	30%	40%	20%	40%	40%	50%	10%
Expert 2	10%	40%	50%	10%	70%	20%	0%	40%	60%
Expert 3	20%	60%	20%	30%	20%	50%	20%	20%	60%
Expert 4	20%	50%	30%	10%	50%	40%	20%	10%	70%
Expert 5	30%	20%	50%	30%	60%	10%	20%	10%	70%

Table 4.7: Opinion of experts for post group 6.

We created a new dataset based on the old one using OpenRefine, composed of the 23 posts and 230 comments chosen in the 6 post groups. Then, we analyzed them with the FIS that we had created it. The comparison between the results detected by FL and those detected by experts is shown in the following:

Posts		Results detected by FL	Results detected by experts
P1	H	2%	46%
	N	94%	12%
	M	4%	42%
P2	H	<1%	28%
	N	91%	4%
	M	9%	68%
P3	H	3%	14%
	N	97%	40%
	M	<1%	46%

Posts		Results detected by FL	Results detected by experts
P4	H	2%	34%
	N	97%	28%
	M	<1%	38%
P5	H	16%	26%
	N	84%	24%
	M	0%	50%
P6	H	0%	26%
	N	55%	26%
	M	45%	48%
P7	H	<1%	38%
	N	2%	22%
	M	88%	38%
P8	H	33%	34%
	N	67%	16%
	M	0%	50%
P9	H	0%	24%
	N	12%	26%
	M	88%	50%
P10	H	0%	16%
	N	95%	54%
	M	5%	30%
P11	H	3%	20%
	N	97%	52%
	M	0%	28%
P12	H	0%	22%
	N	50%	32%
	M	50%	46%
P13	H	2%	18%
	N	70%	20%
	M	28%	62%
P14	H	12%	24%
	N	88%	34%

	Posts	Results detected by FL	Results detected by experts
	M	0%	42%
P15	H	0%	32%
	N	2%	22%
	M	98%	42%
P16	H	0%	24%
	N	95%	56%
	M	5%	20%
P17	H	21%	22%
	N	97%	46%
	M	0%	32%
P18	H	0%	18%
	N	<1%	18%
	M	99%	64%
P19	H	<1%	20%
	N	99%	52%
	M	0%	28%
P20	H	0%	26%
	N	95%	48%
	M	5%	26%
P21	H	<1%	26%
	N	99%	38%
	M	<1%	36%
P22	H	5%	24%
	N	95%	44%
	M	0%	32%
P23	H	0%	20%
	N	73%	26%
	M	27%	54%

Table 4.8: results comparison

Evaluating a machine-learning model is an essential part of any project. In the coming part, we will cover the type of evaluation metrics that we had used to measure the performance of our model.

4.2.3 Evaluation metrics:

Evaluation metrics is a way to quantify performance of a machine-learning model. It is used to examine how good a model is.

d) Confusion Matrix

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

We had assume that our problem is a binary classification problem. We have some samples belonging to two classes: YES or NO. In addition, we have our own classifier, which predicts a class for a given input sample.

We had two outputs:

- An output that detects by an expert or not.
- An output that detects by Fuzzy Logic or not.

There are four important terms:

- **True Positive**: An output has detected by the expert and our model detected it.
- **False Negative**: An output has detected by the expert and our model did not detect it.
- **False Positive**: An output does not has detected by the expert and our model detected it.
- **True Negative**: An output does not has detected by the expert and our model did not detect it.

	Detected by FL	Not detected by FL
Detected by Expert	True Positive	False Negative
Not detected by Expert	False Positive	True Negative

Table 4.9 : Confusion matrix

- **True Positive Rate (Sensitivity):** True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$True\ positive\ Rate = \frac{True\ positive}{False\ negative + True\ positive}$$

- **True Negative Rate(Specificity):**

$$True\ negative\ rate = \frac{True\ negative}{True\ negative + False\ positive}$$

- **False Positive Rate:** False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

$$False\ positive\ rate = \frac{False\ positive}{False\ positive + True\ negative}$$

- **False Negative Rate:**

$$False\ negative\ rate = \frac{False\ negative}{False\ negative + True\ positive}$$

e) **Recall**: is the ability of a model to find all the relevant cases within a dataset. The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives [49].

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives}$$

f) **Precision**: is the ability of a classification model to identify only the relevant data points. Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives [49].

$$Precision = \frac{True\ positives}{True\ positives + False\ negatives}$$

While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant. As with most concepts in data science, there is a trade-off in the metrics we choose to maximize. In the case of recall, when we increase the recall, we decrease the precision [49].

We analyzed the 23 posts by our model using three rule groups, and we calculate the confusion matrix for the three decisions: Happy, Mad and Neutral, and the 4 metrics: TP, TN, FP and FN. Also we calculate the recall and the precision, then we draw the curves of FP and FN in terms of number of rules for the three decisions (Happy, mad and Neutral).

The matrixes, the metrics, recall, and precision are represented in the following tables:

	Mad		Neutral		Happy	
	Detected by Expert	Detected by FL	Detected by Expert	Detected by FL	Detected by Expert	Detected by FL
Post 1	No	Yes	Yes	No	Yes	No
Post 2	No	Yes	Yes	No	No	No
Post 3	No	Yes	Yes	No	No	No
Post 4	No	Yes	No	No	No	No
Post 5	No	Yes	No	No	No	No
Post 6	No	Yes	Yes	No	No	No
Post 7	Yes	No	Yes	No	No	No
Post 8	No	Yes	Yes	No	No	No
Post 9	Yes	Yes	No	No	No	No
Post 10	No	No	Yes	Yes	No	No
Post 11	No	No	Yes	Yes	No	No
Post 12	No	Yes	No	No	No	No
Post 13	No	Yes	Yes	No	No	No
Post 14	No	Yes	Yes	No	No	No
Post 15	Yes	Yes	Yes	No	No	No
Post 16	No	No	Yes	Yes	No	No
Post 17	No	No	Yes	Yes	No	No
Post 18	Yes	Yes	No	No	No	No
Post 19	No	No	Yes	Yes	No	No
Post 20	No	No	Yes	Yes	No	No
Post 21	No	No	Yes	Yes	No	No
Post 22	No	No	Yes	Yes	No	No
Post 23	No	Yes	Yes	No	No	No

Table 4.10: Confusion matrix for the three decision (Mad, Neutral, Happy).

a. Happy

Summary table

	Detected by FL	Not detected by FL
Detected by Expert	0	1
Not detected by Expert	0	22

Table 4.11: Reduced Confusion matrix for "Happy" decision

TPR=0%, TNR=1%, FPR=0%, FNR=0%

$$\text{Recall}=\frac{Tps}{TPs+Fns}=0$$

$$\text{Precision}=\frac{Tps}{Tps+Fps}=0$$

b. Mad

Summary table

	Detected by FL	Not detected by FL
Detected by Expert	3	1
Not detected by Expert	10	9

Table 4.12: Reduced Confusion matrix for "Mad" decision

TPR=0,23%, TNR=0,47%, FPR=0,53%, FNR=0,25%

$$\text{Recall}=\frac{Tps}{TPs+Fns}=\frac{27}{27+6}=0{,}81$$

$$\text{Precision}=\frac{Tps}{Tps+Fps}=\frac{27}{27+22}=0{,}55$$

c. Neutral

Summary table

	Detected by FL	Not detected by FL
Detected by Expert	8	10
Not detected by Expert	0	5

Table 4.13: Reduced Confusion matrix for "Neutral" decision

TPR=1%, TNR=1%, FPR=0%, FNR=0,55%

$$\text{Recall}=\frac{Tps}{TPs+Fns}=\frac{22}{22+22}=0,5$$

$$\text{Precision}=\frac{Tps}{Tps+Fps}=\frac{22}{22+2}=0,91$$

Curves: After calculating the four terms of confusion matrix we draw the curves of FN(False Negative rate) and FP(False Positive rate) for the three decisions as we show in figures

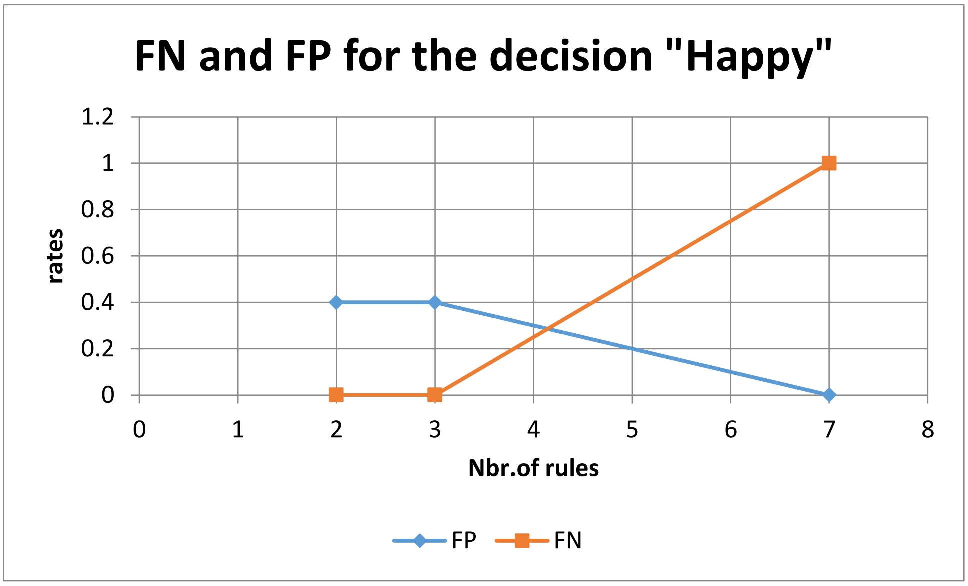


*FIGURE 4.24: **FL AND FN FOR THE DECISION "HAPPY".***

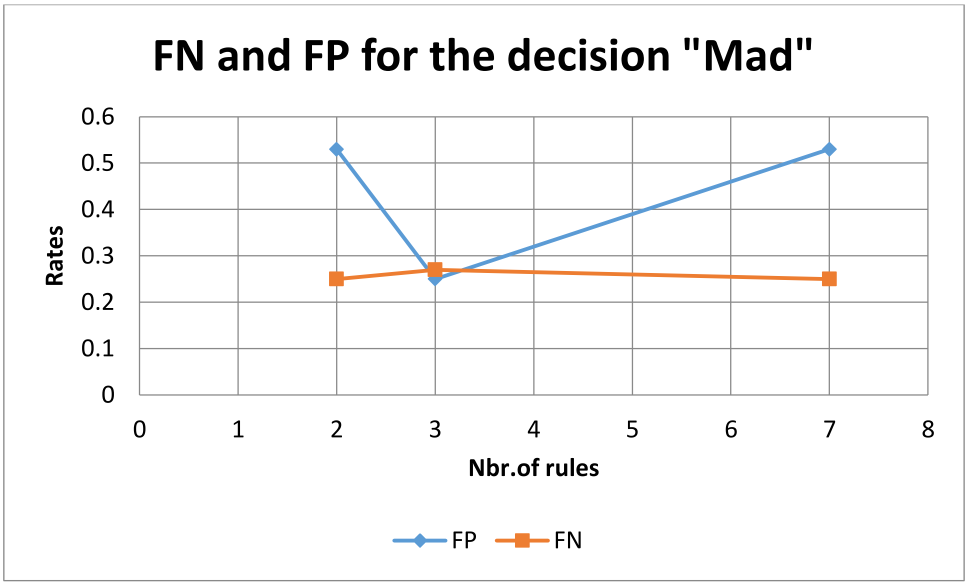


*FIGURE 4.25: **FL AND FN FOR THE DECISION "MAD".***

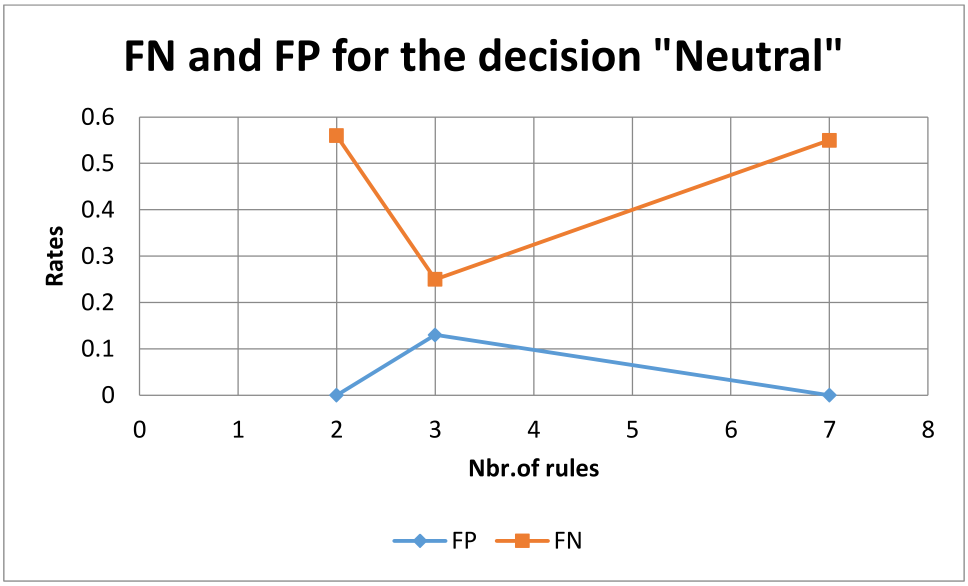


*FIGURE 4.26: **FL AND FN FOR THE DECISION "NEUTRAL".***

4.2.4 Analysis and comparison with all the groups of rules:

We analyzed the 23 posts using fuzzy logic method, but this time we apply the 4 rule groups. After that we gathered the results and compared them. The following table shows the results:

Posts		Rule Group 1	Rule Group 2	Rule Group 3	Rule Group 4
Post 1	Happy	1%	0%	0%	0%
	Mad	19%	19%	99%	99%
	Neutral	80%	81%	1%	1%
	Very Happy	0%	0%	0%	0%
	Very Mad	0%	0%	0%	0%
Post 2	Happy	2%	0%	0%	0%
	Mad	35%	35%	1%	1%
	Neutral	62%	64%	2%	2%
	Very Happy	0%	0%	0%	0%
	Very Mad	1%	<1%	97%	97%

Post 3	Happy	6%	0%	0%	0%
	Mad	0%	0%	0%	0%
	Neutral	93%	100%	21%	21%
	Very Happy	0%	0%	0%	0%
	Very Mad	1%	0%	79%	79%
Post 4	Happy	2%	0%	0%	0%
	Mad	48%	48%	89%	89%
	Neutral	50%	52%	11%	11%
	Very Happy	0%	0%	0%	0%
	Very Mad	0%	0%	0%	0%
Post 5	Happy	24%	0%	0%	0%
	Mad	<1%	<1%	<1%	<1%
	Neutral	75%	99%	89%	90%
	Very Happy	<1%	0%	0%	0%
	Very Mad	<1%	<1%	10%	<1%
Post 6	Happy	0%	0%	0%	0%
	Mad	67%	67%	3%	3%
	Neutral	29%	29%	0%	0%
	Very Happy	0%	0%	0%	0%
	Very Mad	5%	5%	97%	97%
Post 7	Happy	<1%	0%	0%	0%
	Mad	99%	99%	99%	99%
	Neutral	<1%	<1%	<1%	<1%
	Very Happy	0%	0%	0%	0%

	Very Mad	0%	0%	0%	0%
Post 8	Happy	62%	3%	97%	0%
	Mad	0%	0%	0%	0%
	Neutral	35%	97%	3%	100%
	Very Happy	3%	0%	0%	0%
	Very Mad	0%	0%	0%	0%
Post 9	Happy	2%	0%	0%	0%
	Mad	89%	89%	96%	96%
	Neutral	9%	11%	4%	4%
	Very Happy	0%	0%	0%	0%
	Very Mad	0%	0%	0%	0%
Post 10	Happy	1%	0%	0%	0%
	Mad	14%	14%	2%	2%
	Neutral	84%	84%	14%	14%
	Very Happy	0%	0%	0%	0%
	Very Mad	2%	2%	84%	84%
Post 11	Happy	5%	3%	0%	0%
	Mad	3%	0%	5%	5%
	Neutral	88%	93%	47%	47%
	Very Happy	0%	0%	0%	0%
	Very Mad	5%	5%	49%	49%
Post 12	Happy	0%	0%	0%	0%
	Mad	68%	68%	21%	21%
	Neutral	12%	12%	0%	0%

	Very Happy	0%	0%	0%	0%
	Very Mad	20%	20%	79%	79%
Post 13	Happy	2%	0%	0%	0%
	Mad	72%	72%	2%	2%
	Neutral	25%	27%	0%	0%
	Very Happy	0%	0%	0%	0%
	Very Mad	1%	1%	98%	98%
Post 14	Happy	15%	0%	0%	0%
	Mad	26%	26%	1%	1%
	Neutral	57%	72%	28%	28%
	Very Happy	0%	0%	0%	0%
	Very Mad	1%	1%	71%	71%
Post 15	Happy	<1%	0%	<1%	<1%
	Mad	98%	98%	99%	99%
	Neutral	<1%	2%	<1%	<1%
	Very Happy	<1%	0%	0%	0%
	Very Mad	0%	0%	0%	0%
Post 16	Happy	0%	0%	0%	0%
	Mad	11%	11%	1%	1%
	Neutral	88%	88%	14%	14%
	Very Happy	0%	0%	0%	0%
	Very Mad	1%	1%	85%	85%
	Happy	27%	3%	74%	76%

Post 17	Mad	0%	0%	0%	0%
	Neutral	70%	96%	3%	<1%
	Very Happy	3%	0%	0%	0%
	Very Mad	<1%	<1%	23%	23%
Post 18	Happy	0%	0%	0%	0%
	Mad	98%	98%	<1%	<1%
	Neutral	1%	<1%	0%	0%
	Very Happy	0%	0%	0%	0%
	Very Mad	1%	<1%	99%	99%
Post 19	Happy	3%	0%	85%	85%
	Mad	<1%	0%	15%	15%
	Neutral	96%	100%	<1%	51%
	Very Happy	<1%	0%	0%	0%
	Very Mad	0%	0%	0%	0%
Post 20	Happy	40%	0%	0%	0%
	Mad	7%	7%	1%	1%
	Neutral	51%	91%	69%	69%
	Very Happy	0%	0%	0%	0%
	Very Mad	2%	2%	29%	29%
Post 21	Happy	<1%	0%	0%	0%
	Mad	<1%	<1%	4%	4%
	Neutral	99%	99%	96%	96%
	Very Happy	0%	0%	0%	0%
	Very	0%	<0%	0%	0%

	Mad				
Post 22	Happy	26%	0%	<1%	74%
	Mad	0%	0%	0%	0%
	Neutral	74%	100%	26%	26%
	Very Happy	<1%	0%	74%	0%
	Very Mad	0%	0%	0%	0%
Post 23	Happy	0%	0%	0%	0%
	Mad	54%	54%	<1%	<1%
	Neutral	45%	45%	<1%	<1%
	Very Happy	0%	0%	0%	0%
	Very Mad	<1%	<1%	99%	99%

Table 4.14: Fuzzy sets result comparison of rule groups

By passing to the crisp sets, we can summarize the results as the following:

	Rule group 1	Rule group 2	Rule group 3	Rule group 4
Post 1	Neutral	Neutral	Mad	Mad
Post 2	Neutral	Neutral	Very Mad	Very Mad
Post 3	Neutral	Neutral	Very Mad	Very Mad
Post 4	Mad	Neutral	Mad	Mad
Post 5	Neutral	Neutral	Neutral	Neutral
Post 6	Mad	Mad	Very Mad	Very Mad
Post 7	Mad	Mad	Mad	Mad
Post 8	Happy	Neutral	Happy	Neutral
Post 9	Mad	Mad	Mad	Mad
Post 10	Neutral	Neutral	Very Mad	Very Mad
Post 11	Neutral	Neutral	Very Mad	Very Mad
Post 12	Mad	Mad	Very Mad	Very Mad
Post 13	Mad	Mad	Very Mad	Very Mad
Post 14	Neutral	Neutral	Very Mad	Very Mad

Post 15	Mad	Mad	Mad	Mad
Post 16	Neutral	Neutral	Mad	Mad
Post 17	Neutral	Neutral	Happy	Happy
Post 18	Mad	Mad	Very Mad	Very Mad
Post 19	Neutral	Neutral	Happy	Happy
Post 20	Neutral	Neutral	Neutral	Neutral
Post 21	Neutral	Neutral	Neutral	Neutral
Post 22	Neutral	Neutral	Very Happy	Very Happy
Post 23	Mad	Mad	Very Mad	Very Mad

Table 4.15: crisp sets result comparison of rule groups

4.2.5 Analysis of the results:

From the two tables above, we observed the following:

- For the posts 5, 7, 9, 15, 20 and 21, the rule groups 1, 2, 3 and 4 give the same crisp set ("Neutral" for posts 5, 20 and21, and "Mad" for posts 7, 9 and15) with a slight difference in the fuzzy set rates.
- For the posts 1, 2, 3, 4, 11, 14 and19, the rule groups 1 and 2 give the same crisp set ("Neutral"). However, for the rule group 3 and 4 give the same crisp and fuzzy set.
- For the post 8; the rule groups 1 and 3 give the same crisp set ("Happy") with a difference of 35% in fuzzy set rate, and the rule groups 2 and 4 give the same crisp set ("Neutral") with 3% of difference in fuzzy set rate.
- For the posts 6, 12, 13, 16, 18 and 23, the rule groups 1 and 2 give the same crisp and fuzzy set ("Mad" for 6, 12, 13, 18 and 23 and "Neutral" for16) with the same rate. The same thing for the rule groups 3 and 4 ("Very Mad" for 6, 12, 13, 16, 18 and 23)
- For the post 10, the rule groups 1 and 2 give the same crisp and fuzzy set ("Neutral"), and the rule groups 3 and 4 give the same crisp and fuzzy set ("Very Mad") with the same rate, and the 4 rule groups give the same rate (84%),
- For the post 22, the rule groups 1 and 2 give the same crisp set ("Neutral") with a difference of 26% in the fuzzy set rate. The rule group 3 give "Very Happy" with a rate of 74%, and "Happy" by the rule group 4 with the same rate.

4.3 Conclusion:

At the end of chapter, we conclude that there is an inverse relationship between the metrics FP and FN and the number of rules, which means that whenever the number of rules increase; the metrics FN and FP decrease, and this decrease indicates the accuracy of results founded. In other word we can say; "less rules less accurate results; more rules more accurate results".

General Conclusion

Big data is a collection of vast amount of data that is difficult to handle with existing computer memory. The abundant amount of information generated from the various social networking sites, especially Facebook alone logs produced 25 terabytes (TB) of data per day.

Handling of such big data with existing resources is the major challenge. The challenges include categorization, searching, sharing, and visualization and analyze of big data with limited resources. Many algorithms have been proposed to perform analyses of big data but, only few of them address the fuzzy logic methods.

Sentiment analysis is nothing but special field of text analysis. In short, focus and analyze the extracted opinions (sentiments or emotional contents) from the posted comments. Our project goal is to analyze the sentiments on politic posts which are extracted from Facebook. We have implemented a fuzzy logic model constructed by fuzzy Inference system. And after analyzing, we evaluate our model by comparing the obtained results by other results obtained from experts' evaluations. We conclude that the number of rules used in the fuzzy system has an important effect on the accuracy of results.

As a future work, we suggest investigating more the quality of rules to be added to the fuzzy inference system. Moreover, hybridizing the fuzzy logic with neural nets could lead to better rule selection and thus better analyses for Big Data.

.

References

[1] Miroslav Vozábal, Tools and Methods for Big Data Analysis, Master Thesis, Pilsen, 2016

[2] Zaharia, Matei; Chowdhury, Mosharaf; Das, Tahagata; Dave, Ankur; Ma, Justin; McCauley, Murphy; Franklin, Michael; Shenker, Scott; Stoica, Ion;. Resilient Distributed Datasets: A Fault-Tolerant. Berkeley: University of California at Berkeley, Electrical Engineering and Computer Sciences, 2011.

[3] Manyika, James and Chui, Michael. *Big data: The next frontier for innovation, competition, and productivity.* s.l. : McKinsey Global Institute, 2011. 978-0983179696.

[4] Iafrate, Fernando and Front, Matter. *From Big Data to Smart Data.* Chap : John Wiley & Sons, 2015.

[5] http://www.greenplum.com, January, 2019.

[6] Mahesh G Huddar, A Survey on Big Data Analytical Tools, Hirasugar Institute of Technology, January 2013.

[7] Dr Hemlata Chahal, Big Data Analytics, Maharshi Dayanand University, February 2016

[8] https://www.sas.com/en_us/insights/big-data/what-is-big-data.html, Big Data What it is and why it matters, January, 2019.

[9] https://datafloq.com/read/how-social-media-companies-use-big-data/1957, How Social Media Companies Use Big Data, January, 2019..

[10] https://searchbusinessanalytics.techtarget.com/definition/bigdataanalytics, data analytics resources and information, January, 2019.

[11] https://www.youtube.com/watch?v=9s-vSeWej1U, what is Hadoop, January, 2019.

[12] Atima, Han Zhuang, Ishita Vedvyas, Rishikesh Dole, Tutorial: OpenRefine

[13] Karen Howells, Ahmet Ertugan, Applying fuzzy logic for sentiment analysis of social media network data in marketing, aNear East University, Faculty of Economics and Administrative Sciences, 22-23 August 2017

[14] M. AL-maimani ,N. salim, A. M. Al-naamany, semantic and fuzzy aspects of opinion mining, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

[15] https://www.guru99.com/what-is-fuzzy-logic.html, Fuzzy Logic Tutorial: What is, Application & Example, February, 2019.

[16] R. Vatrapu, R. R. Mukkamala, A. Hussain, A. B. Flesch, “Social Set Analysis: A Set Theoretical Approach to Big Data Analytics”, April 28, 2015.

[17]R. Vatrapu, A. Hussain, D. Hardt, and Z. Jaffari, “Social data analytics tool: A demonstrative case study of methodology and software”, in *Ana-lyzing Social Media Data and Web Networks,* M. Cantijoch, R. Gibson, S. Ward, Eds. Basingstoke, U.K.: Palgrave Macmillan, 2014.

[18]A. Hussain and R. Vatrapu, “Social data analytics tool (SODATO)”, in *Advancing the Impact of Design Science: Moving from Theory to Practice*, 2014.

[19] C. C. Ragin, “*Fuzzy-Set Social Science”*, USA: Univ. Chicago Press, 2000.

[20] M. J. Smithson and J. Verkuilen, “*Fuzzy Set Theory: Applications in the Social Sciences* (Quantitative Applications in the Social Sciences)”. New York, Feb. 2006.

[21] A. Kechris, “*Classical Descriptive Set Theory”*, , NY, 2012.

[22] M. Kryszkiewicz, “Rough set approach to incomplete information systems”, 1998.

[23] N. M. Tichy, M. L. Tushman, and C. Fombrun, “Social network analysis for organizations”, Oct. 1979.

[24] D. Krackhardt, “Cognitive social structures”, Jun. 1987.

[25] J. Zhan and X. Fang, “Social computing: The state of the art”, 2011.

[26] J. Karikoski and M. Nelimarkka, “Measuring social relations with multiple datasets”, 2011.

[27] J. Sabater and C. Sierra, “Reputation and social network analysis in multiagent systems”, 2002.

[28] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and A.Wallace, “Finding overlapping communities in social networks”, Aug. 2010.

[29] O. Macindoe and W. Richards, “Comparing networks using their _ne structure”, 2011.

[30] B. Pang and L. Lee, “Opinion mining and sentiment analysis”, 2008.

[31] C. R. Fink, D. S. Chou, J. J. Kopecky, and A. J. Llorens, “Coarse-and _negrained sentiment analysis of social media text”, 2011.

[32] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis”, 2009.

[33] H. Chen, P. De, Y. Hu, and B.-H. Hwang, “Sentiment revealed in social media and its effect on the stock market”, Jun. 2011.

[34] D. Hardt and J. Wulff, “What is the meaning of 5 _0s? An investigation of the expression and rating of sentiment”, Sep. 2012.

[35] S. P. Robertson, “Changes in referents and emotions over time in electionrelated social networking dialog”, Jan. 2011.

[36] M. Salathé and S. Khandelwal, “Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control”, 2011.

[37] T. Menezes, C. Roth, and J.-P. Cointet, “Finding the semantic-level precursorson a blog network”, 2011.

[38] R. R. Mukkamala, A. Hussain, and R. Vatrapu, “Towards a set theoretical approach to big data analytics”, Jun./Jul. 2014.

[39] R. R. Mukkamala, A. Hussain, and R. Vatrapu, “Towards a formal model of social data”, IT Univ. Copenhagen, Copenhagen, Denmark, Nov. 2013.

[40] MathWorks, “MATLAB Getting Started Guide”, 2011.

[41]https://raw.githubusercontent.com/saodem74/SentimentAnalysis/master/Data/comment_data.csv, April, 2019.

[42] https://github.com/saodem74/Sentiment-Analysis-facebook-comments,(Sentiment-Analysis-facebook-comments) , April, 2019.

[43] https://www.mathworks.com/help/fuzzy/fuzzy-inference-systemmodeling.html,(Fuzzy Logic Toolbox) , April, 2019.

[44] Fuzzy Inference Systems. pdf

[45] **https://www.mathworks.com/help/fuzzy/fuzzy-inference-process.html, (Fuzzy Inference Process), April, 2019.**

[46]https://www.google.com/search?q=Big+Data+characteristic&source, (Big Data characteristic and resources.) , March, 2019.

[47] Puneet Singh Duggal , Sanchita Paul, “Big Data Analysis: Challenges and Solutions”, 2013.

[48] Jasmine Zakir, “Big Data Analytics”, 2015.

[49]https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c, (Beyond Accuracy: Precision and Recall) , March, 2019.

[50] https://www.sciencedirect.com/topics/engineering/fuzzy-inference, (fuzzy inference system) , March, 2019.

[51] H.-J. Zimmermann, “Fuzzy set theory”.

[52] https://developers.google.com/prediction, (L'API Cloud Prediction est obsolète) , March, 2019.

[53] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis”, 2009.

[54]https://www.facebook.com/McDonalds/photos/a.10150151878945584.414818.10150097174480584/10156340092750584/?type=3, McDonalds. *McDonalds Facebook Post*, June, 2019.

[55] https://edoras.sdsu.edu/doc/matlab/toolbox/fuzzy/fuzzyt10.html, June, 2019.