VISVESVARAYA TECHNOLOGICAL UNIVERSITY JNANA SANGAMA, BELAGAVI_590018

AN INTERNSHIP REPORT

On

"Stockport | Predictive Sentiment Analysis"

Submitted in partial fulfillment of the requirement for the award of degree of

BACHELOR OF ENGINEERING

In

COMPUTER SCIENCE AND ENGINEERING

Submitted By

HARSHITHA KG (4MO21CS013)

Under the guidance of

Junior engineer officer

VARCONS TECHNOLOGIES PVT LTD

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING 2022-2023 MYCEM COLLEGE OF ENGINEERING AND MANAGEMENT

(Affiliated to VTU and approved by AICTE, Recognized by Govt. Of Karnataka)

(An ISO 9001-2015 Certified Institution)

#1072, T. Narasipura Road, Near Big Banyan Tree, Chikkahalli, Mysuru-570028, Karnataka, India.

MYSORE COLLEGE OF ENGINEERING AND MANAGEMENT

#1072, T. Narasipura Road, Near Big Banyan Tree, Chikkahalli, Mysuru.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to Certify that the Internship report entitled **"Stockport | Predictive Sentiment Analysis"** submitted by **Ms. HARSHITHA K G(4MO21CS013)** a bona-fide student Of **Mysore College Of Engineering and Management Mysuru, Karnataka 570028,** in partial fulfillment for the award of **Bachelor of Engineering in Computecience And Engineering** at the **Visvesvaraya Technological University, Belgaum-590018,** During year 2022-2023. It is certified all correction/suggestion indicated have been incorporated in the report.

Signature of the Guide: **Signature of the HOD:**

SIGNATURE OF PRINCIPAL

MYCEM,

Mysore

MYSORE COLLEGE OF ENGINEERING AND MANAGEMENT

#1072, T. Narasipura Road, Near Big Banyan Tree, Chikkahalli, Mysuru

DECLARATION

I **Harshitha KG (4MO21CS013)**, student of 2nd semester, studying at Mysore College of Engineering and Management, Mysore, hereby declare that the training report on **"Stockport | Predictive Sentiment Analysis"** submitted to Visvesvaraya Technological University, Belagavi, in partial fulfillment of degree of Bachelor of Engineering is the original work conducted by me.

The information and data in the report is authentic to the best of my knowledge. This internship training report is not being published or submitted to any other university for award of any other degree, diploma or other similar titles.

Date :17/11/22 **HARSHITHA KG**

Place: Mysuru. **4MO21CS013**

Internship ,Offer Letter

Varcons Technologies Pvt Ltd

Date: **14th October, 2022**

Name: **Harshitha Kg**
USN: **4MO21CS013**

Dear Student,

We would like to congratulate you on being selected for the **Machine Learning With-Python(Research Based)** Internship position with **Varcons Technologies Pvt Ltd**, effective Start Date **14th October, 2022**, All of us are excited about this opportunity provided to you!

This internship is viewed as being an educational opportunity for you, rather than a part-time job. As such, your internship will include training/orientation and focus primarily on learning and developing new skills and gaining a deeper understanding of concepts of **Machine Learning With Python(Research Based)** through hands-on application of the knowledge you learn while you train with the senior developers. You will be bound to follow the rules and regulations of the company during your internship duration.

Again, congratulations and we look forward to working with you!

Sincerely,

Spoorthi H C
Director
Varcons Technologies Pvt Ltd
213, 2nd Floor,
18 M G Road, Ulsoor,
Bangalore-560001

ACKNOWLEDGEMENT

First and foremost, I would like to express my gratitude to my Organizational study Mr. Ningaraju KL managing director VARCONS TECHNOLOGIES PVT LTD, BANGALORE.I am also in debited to the management of VTU college for provide me with a good study environment and excellent library facilities. I thankful to their help and support rendered by the teaching and non-teaching staff of department of business administration.

Lastly, I take this opportunity to offer my regard to all those who have support me directly and indirectly in complementing this organizational study

Finally, it's my foremost duty to thankful my parents for encouragement and supportive during entire course.

Place: MYSURU

Harshitha kg

USN: 4MO21EC016

Date: 17/11/22

Table of Contents

CHAPTER 1
INTRODUCTION ABOUT ORGANISATION AND INDUSTRY

Introduction :

Information Technology in India is an industry consisting of two major components: IT services and Digital_Marketing. The IT industry accounted for 8% of India's GDP in 2020. The IT–BPM sector overall employs 4.5 million people as of March 2021. The Indian IT–BPM industry has the highest employee attrition rate. As IT–BPM sector evolves, many are concerned that artificial intelligence (AI) will drive significant automation and destroy jobs in the coming years. The United States accounts for two-thirds of India's IT services exports.

Industry Background: -

The technology industry is comprised of companies that design, manufacture, or distribute electronic devices such as computers, computer-related equipment, computer services and software, scientific instruments, and electronic components and products. Technology enables consumers and businesses to thrive in a digital world. The composition of this industry is very different than that of most others; due to the brisk pace of innovation there is an unusually extensive investment in research and development required. As a result, the industry's workforce consists of a much larger proportion of engineers and other highly-skilled technical workers, relative to other industries, especially since product creation requires creativity, expertise, and precision. The technology industry also employs a relatively large workgroup engaged in sales and promotion, as the success of a new or improved product depends heavily upon consumers being aware of, and interested in, the item. While most of the sales for this industry occur in developed countries, most of the

production of computer hardware takes place in emerging countries where manufacturing and assembly costs are lower.

History of the Industry: -

India's IT Services industry was born in Mumbai in 1967 with the creation of Tata Consultancy Services[11] who in 1977 partnered with Burroughs which began India's export of IT services. The first software export zone, SEEPZ – the precursor to the modern-day IT park – was established in Mumbai in 1973. More than 90 percent of the country's software exports were from SEEPZ in the 1980s.

Within 90 days of its establishment, the Task Force produced an extensive background report on the state of technology in India and an IT Action Plan with 108 recommendations. The Task Force could act quickly because it built upon the experience and frustrations of state governments, central government agencies, universities, and the software industry. Much of what it proposed was also consistent with the thinking and recommendations of international bodies like the World Trade Organization (WTO), International Telecommunication Union (ITU), and World Bank. In addition, the Task Force incorporated the experiences of Singapore and other nations, which implemented similar programs. It was less a task of invention than of sparking action on a consensus that had already evolved within the networking community and government.Regulated VSAT links became visible in 1994. Desai (2006) describes the steps taken to relax regulations on linking in 1991:

In 1991 the Department of Electronics broke this impasse, creating a corporation called Software Technology Parks of India (STPI) that, being owned by the government, could provide VSAT communications without breaching its monopoly. STPI set up software technology parks in different cities, each of which provided satellite links to be used by firms; the local link was a wireless

radio link. In 1993 the government began to allow individual companies their own dedicated links, which allowed work done in India to be transmitted abroad directly. Indian firms soon convinced their American customers that a satellite link was as reliable as a team of programmers working in the clients' office.
A joint EU-India group of scholars was formed on 23 November 2001 to further promote joint research and development. On 25 NOV 2002, India and the European Union agreed to bilateral cooperation in the field of science and technology. From 2017, India holds a Associate Member State status at CERN, while a joint India-EU Software Education and Development Center will be located in Bangalore.

The technology industry is relatively young; its origins can be traced to the 1904 invention of the two-element electron tube. Developments such as the transistor followed, as well as integrated circuits in the 1950s, and analog devices in 1960. Many of these inventions were a result of military research. The 1970s brought the invention of the integrated circuit board and the microprocessors that soon followed, made home and personal computers a possibility in subsequent years. However, until the internet was available to common consumers, computers were not very popular. In the 1990s, when the internet was available to all, there was an explosion in the use of personal computers. That explosion made household names out of entrepreneurs such as Bill Gates and Steve Jobs. The early part of 2000 saw a drastically reduced demand for computers, but in 2003 and 2004 the market experienced a turnaround as consumers sought multi-tasking computers that could handle a myriad of photo, video, and audio applications. Since then the technology has spread to mobile phones and tablets that offer all the services of a normal stationary computer. Cloud computing has revolutionized data storage and as a result devices are continually getting smaller and more advanced.

Varcons Technologies Pvt Ltd Industry Overview: -

Information and communication technology (ICT) has been used in schools

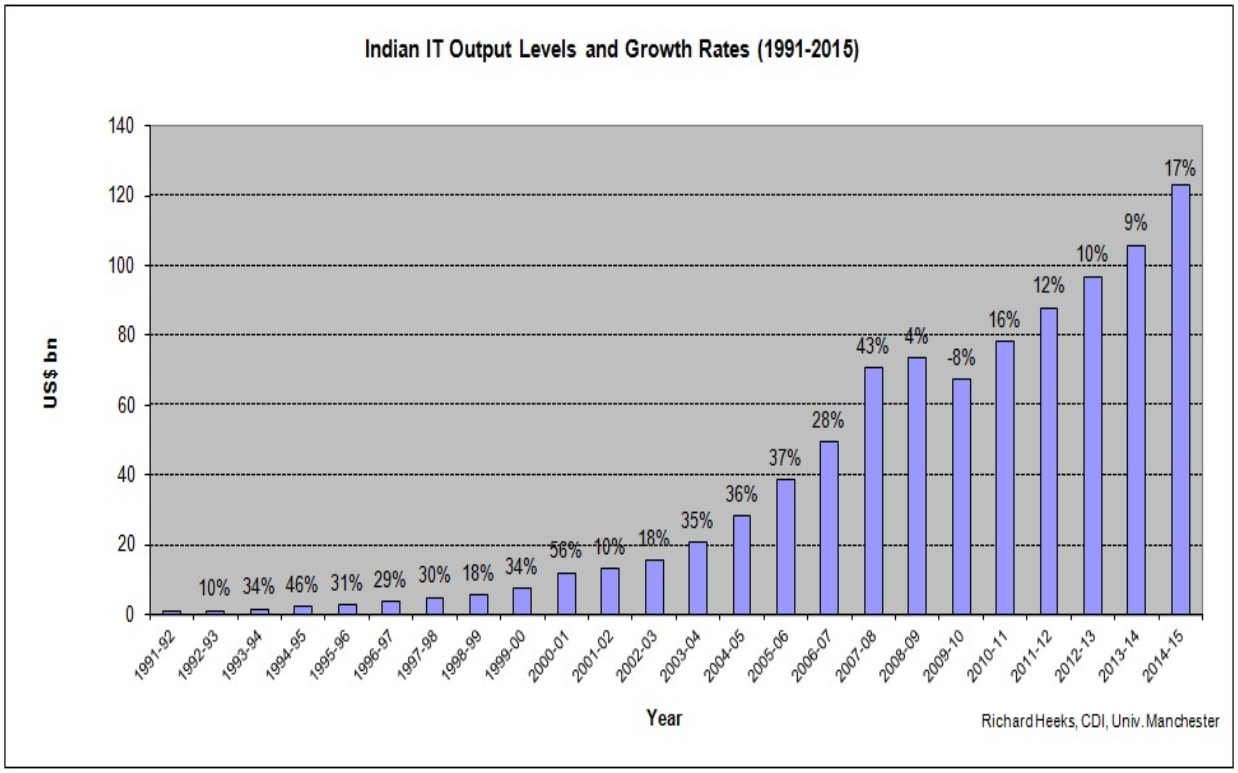
since the 1980s, but the advent of the World Wide Web, along with increases in computing power in low cost and portable forms, has made use of ICT much more prominent for learning. ICT is being used to support teaching, learning, and assessment. Current technology trends with the potential to further change learning practices include the increasing availability of open-source course content on the Internet; the rise of collaborative, user-generated content; and immersive, multi-user games with highly realistic dynamic graphics. Coupled with the rise in young people's access to technology outside of school, these trends are likely to make ICT an increasingly important factor in learning.

Contribution of Industry to National Economy: -

- The share of service sector in the GDP (Gross Domestic Product) has been stagnant at 17% over the last two decades. The total contribution of industry to the GDP is 27% out of which 10% comes from mining, quarrying, electricity and gas.
- The growth of the service sector had been 7% in the last decade. Since 2003, the growth rate has been 9 to 10% per annum. The desired growth rate over the next decade is 12%.
- The National Manufacturing Competitiveness Council (NMCC) has been set with the objectives of improving productivity through proper policy interventions by the government and renewed efforts by the industry.

Industry Growth :-

The IT industry accounted for **8% of India's GDP in 2020**. Exports from the Indian IT industry are expected to increase by 1.9% to reach US$ 220 billion in FY21. In 2020, the IT industry recorded 152,000 new hires.



Market Size

The IT & BPM industry's revenue is estimated at ~US$ 198 billion in FY21, an increase of 2.3% YoY. The domestic revenue of the IT industry is estimated at US$ 45 billion and export revenue is estimated at US$ 150 billion in FY21. According to Gartner estimates, IT spending in India is estimated to reach US$ 93 billion in 2021 (7.3% YoY growth) and further increase to US$ 108.5 billion in 2022. The BPM sector in India currently employs >1.4 million people, while IT and BPM together have >4.5 million workers, as of FY21.

Indian software product industry is expected to reach US$ 100 billion by 2025. Indian companies are focusing to invest internationally to expand global footprint and enhance their global delivery centers. In line with this, in February 2021, Tata Consultancy Services announced to recruit ~1,600 technology employees across the UK over the next year. The development would build capabilities for TCS to deliver efficiently to the UK customers.

The data annotation market in India stood at ~ US$ 260 million in FY20, of which the US market contributed ~ 60% to the overall value. The market is expected to reach ~ US$ 7 billion by 2030 due to accelerated domestic demand

for AI.

Investments/ Developments

Indian IT's core competencies and strengths have attracted significant investment from major countries. The computer software and hardware sector in India attracted cumulative foreign direct investment (FDI) inflows worth US$ 79.05 billion between April 2000 and March 2021. The sector ranked 2nd in FDI inflows as per the data released by Department for Promotion of Industry and Internal Trade (DPIIT). In FY21, computer software and hardware topped FDI investments, accounting for 54% share of the total FDI inflows of US$ 91.72 billion.

Leading Indian IT firms like Infosys, Wipro, TCS and Tech Mahindra are diversifying their offerings and showcasing leading ideas in blockchain and artificial intelligence to clients using innovation hubs and research and development centers to create differentiated offerings.

Some of the major developments in the Indian IT and ITeS sector are as follows:

- In August 2021, Tata Consultancy Services was adjudged a leader in the Nelson Hall NEAT for CX Services in Banking, Financial Services and Insurance (BFSI).
- In August 2021, SAP India and Microsoft announced the introduction of Tech Saksham, a collaborative skilling initiative aimed at enabling young women (from underprivileged regions) to pursue careers in technology. 68,000 women students will be trained in artificial intelligence (AI), cloud computing, web design and digital marketing as a result of this collaboration.
- In July 2021, Wipro announced plans to invest US$ 1 billion over the next three years to expand its cloud technology capabilities through acquisitions and collaborations.

- In July 2021, Infosys announced that it has set up an Automotive Digital Technology and Innovation Centre in Stuttgart, Germany. Automotive IT infrastructure professionals stationed in Germany will transfer from Daimler AG to the new Digital Technology and Innovation Centre as part of Infosys' relationship with Daimler.
- In July 2021, TCS expanded its strategic partnership with Royal London, the largest mutual life insurance, pensions and investment company in the UK, to help the latter transform its pension platform estate and deliver market-leading services to members and customers.
- In July 2021, Tata Technologies partnered with Strategy's, a 3D printing technology company, to provide advanced additive manufacturing technologies to the Indian manufacturing ecosystem.
- In July 2021, Tech Mahindra Foundation and Wipro GE Healthcare have joined forced to offer skilling and up skilling courses to students and healthcare technicians.
- In July 2021, HCL announced a multi-year agreement with Fiscals Group, consisting of a family of lifestyle brands including Fiscals, Gerber, Royal Copenhagen, Waterford and Wedgwood for digital transformation.
- In July 2021, TCS launched Jile 5.0, a key release of its Enterprise Agile, on-the-cloud services, planning and delivery tool that enables enterprises to meet the large-scale development needs of multiple distributed teams.

Government Initiatives

Some of the major initiatives taken by the Government to promote IT and ITeS sector in India are as follows:

- In August 2021, the India Internet Governance Forum (IIGF) – 2021 was launched at Electronics Nike tan in New Delhi by the National Internet Exchange of India (NIXI), the Ministry of Electronics and Information Technology (Meaty) and the Chairman of the Coordination Committee of the IIGF-2021. The event will take place over three days beginning October 20, 2021. The meeting's topic this year is Inclusive Internet for Digital India.

- On July 2, 2021, the Ministry of Heavy Industries and Public Enterprises launched six technology innovation platforms to develop technologies for globally competitive manufacturing in India. The six technology platforms have been developed by IIT Madras, Central Manufacturing Technology Institute (CMTI), International Centre for Automotive Technology(iCAT), Automotive Research Association of India(ARAI), BHEL and HMT in association with IISc Banglore.
- In July 2021, the Arun Jaitley National Institute of Financial Management (AJNIFM) and Microsoft have formed a strategic partnership to build AI and emerging technologies Centre of Excellence.
- In NOV 2021, the Indian government announced plans to launch Biotech-PRIDE (Promotion of Research and Innovation through Data Exchange) to deposit biological data in the country's national repository.
- In May 2021, My Gov, the citizen engagement platform of the Government of India, in partnership with the Department of Higher Education launched an innovation challenge to create an Indian language learning app.
- In order to establish an enabling environment for the IT industry, in April 2021, the Development of Advanced Computing (C-DAC) launched three innovative technologies Automatic Parallelizing Compiler (CAPC), Cyber Security Operation Centre (CSoC) as a Service, and C-DAC's indigenous High-performance Computing software solutions—Parallel Development Environment (Parade).
- In Budget 2021, the government has allocated Rs. 53,108 crores (US$ 7.31 billion) to the IT and telecom sector.
- Department of Telecom, Government of India and Ministry of Communications, Government of Japan signed a MoU to enhance cooperation in areas of 5G technologies, telecom security and submarine optical fibro cable system.
- In 2020, the government released "Simplified Other Service Provider" (OSP) guidelines to improve the ease of doing business in the IT Industry, Business Process Outsourcing (BPO) and IT-enabled Services.

CHAPTER 2

ORGANIZATION PROFILE

Varcons Technologies Pvt Ltd is into providing of IT Services and Training for over 15 years now. They cater to customers from different Industries/Verticals in India and Worldwide.

Motto: Our inspiration and motivation is to produce creative, efficient and cost effective solutions for our clients in a desired & apt manner. We not only listen to our customer's ideas and requirements, but also try to add to the value by contributing some of our own ideas innovatively.

Overview & Credentials of Varcons Technologies Pvt Ltd

Belongs to one of the Largest Business IT giants in India, VARCONS TECHNOLOGIES PVT LTD Group Incorporated in 2005, 15+ years of experience and excellence Excellent client base, spread across India and globally Certified for ISO 9001: 2008 for Quality Management Best Project management, QC, and Delivery processes Timeliness, Reliability, Quality all at affordable prices Our clients belong to various verticals including, **Technology, eLearning, Examinations, Yoga , ERP , Apps**

E-Commerce, Database Management…etc.

The company has been formed by a group of professionals having vivid experience and wide exposure in Information Technology. People involved here are young qualified business graduates and qualified engineers from the renowned universities across the Karnataka.

The resource personnel working in the company have been consistently providing reliable support services and consultancy to a wide variety of corporate houses either in the capacity of executive or as business partner or consultant. Bottom line of the company philosophy is building a long-term business partnership with its clients where interpersonal relationship, reliability, assured quality and target oriented modern technology are the major building blocks.

It is a company where professionals from both technical and functional field group together with an objective of providing appropriate business solutions. It realizes the importance of functional knowledge and its impact in developing business solutions. We constantly strive to be a leading technology firm with profound business and functional knowledge. The key to the company's success is the maintenance of a close working relationship with the clients through ensuring the best possible solutions to their needs; to establish and maintain a thorough knowledge and understanding of client's objective and help them maximize the benefits. We want to establish ourselves as the best choice in Computing and Information Technology Services, Consultancy and Development by offering the full spectrum of services.

Nature and business:

The Company is a system integrator of Information Technology and telecommunication services provider, by designing a computer system and a computer communication system, which are integrating and working efficiently and effectively as the needs of the customers. A system with services of computer hardware, system software, application software, and computer network equipment. The Company has a comprehensive sales style or Turn Key, starting from understands the needs of customers, consulting, project planning, system design, installation operation, maintenance, as well as training for the clear understanding in operation.

Area of operation:

IT operations is the overarching term for the processes and services administered by an organization's information technology (IT) department. As such, IT operations include administrative processes with support for hardware and software. Important roles of the IT operations team include tech management, quality assurance, infrastructure management and confirmation that finished products meet all the customer's needs and expectations. IT operations support both internal and external clients. Every organization that uses computers has at least loosely defined IT operations, based on how the organization tends to solve internal and client needs. Elements of IT operations are chosen to deliver effective services at the required quality and cost -- usually considered to be separate from IT applications. In the example of a software development company, IT operations include all IT functions other than software development and management. There is, however, always some overlap between departments.

Vision of the Organization:

Our Vision is to achieve 100% customer satisfaction by delivering quality products and services at an affordable cost. Our forward vision is to strive to become an entity in technology based corporate solutions, capable of

demanding unconditional response from the targeted niche. We also believe that for our scope of improvisation – sky is the limit and we are always ready to take our achievements to the next level. We are growing and would always like to remain on the growing streak.

Mission of the organization:

Our Mission is to achieve the reputation of a quality, high standard & reliable solution & service Provider Company in the IT industry.

Our keys for development:

- Desire for Excellence.
- Trust and confidence build-up.
- Innovation.
- Transparency.
- Teamwork.

We believe in

- Motivation.
- Collective responsibility and leadership.
- Professionalism and ethics.
- Adding values to our client needs.

Quality policies of Varcons Technologies Pvt Ltd:

We believe quality is defined by our customers. The direct measure of how well we are delivering on our quality commitment is the degree to which we meet our customer's requirements and exceed their expectations. Our customer's success is the most important factor in our long-term success."Impeccable Software's is committed to satisfying our customers by meeting or exceeding agreed upon requirements and expectations."We have set forth before us the following objectives, which are measurable and are consistent with the 'Quality Policy'.

WORK FLOW MODEL

The work flow model of the Varcons Technologies Pvt Ltd as shows given

below:

Figure representing work flow model of Varcons Technologies Pvt Ltd.

Workflow:

Workflows are designed to help you optimize business processes, streamlining them for efficiency and consistent stellar results. However, you won't be doing much good by simply tacking on a workflow into your business processes. **Creating an efficient workflow model** is key to reaping these benefits.

Basic components of a workflow:

Predefined

Steps are a predefined framework of tasks in the workflow. They provide clarity on what happens at each stage of the workflow until the endpoint. Steps may be manual or automated based on different workflows.

Stakeholders

Stakeholders are the people who are responsible for carrying out specific tasks in the workflow. Stakeholders are assigned to either each step of the workflow, a group of steps, or to the whole workflow. In some workflows, the steps are

completely automated, making stakeholders less active. Stakeholders only step into the workflow during specific conditions or when problems arise.

Conditions

Conditions are rules for the workflow. They determine when a particular step is completed and what the next step should be. Conditions are most useful for approval-type workflows where some steps are skipped based on the information.

Conclusion:

Workflow management is a win-win for businesses.

It provides both the employees and business owners with a better, more structured, and efficient way to communicate and collaborate.

SERVICES NEXT GENERATION SPEED: -

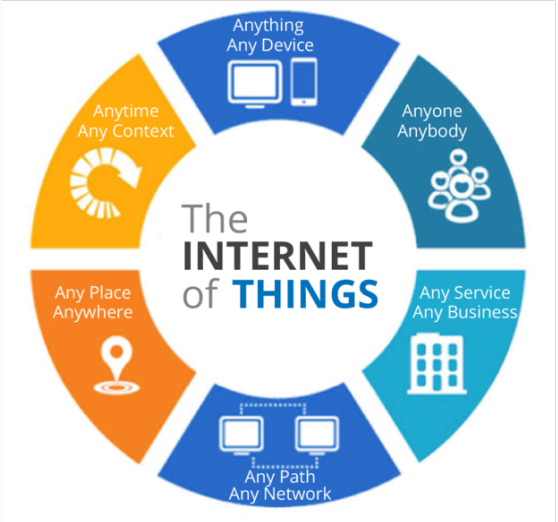
SOFTWARE DEVELOPMENT

We focus on understanding business requirements and effective application

development for all sizes and complexities of businesses and organizations. Our constant Endeavor is to use latest technologies and following industry best practices for building robust applications for our clients

INTERNET OF THINGS

The Internet of Things (IOT) is the inter-networking of physical devices, vehicles (also referred to as "connected devices" and "smart devices"), buildings, and other items embedded with electronics, software, sensors, actuators, and network connectivity which enable these objects to collect and exchange data.



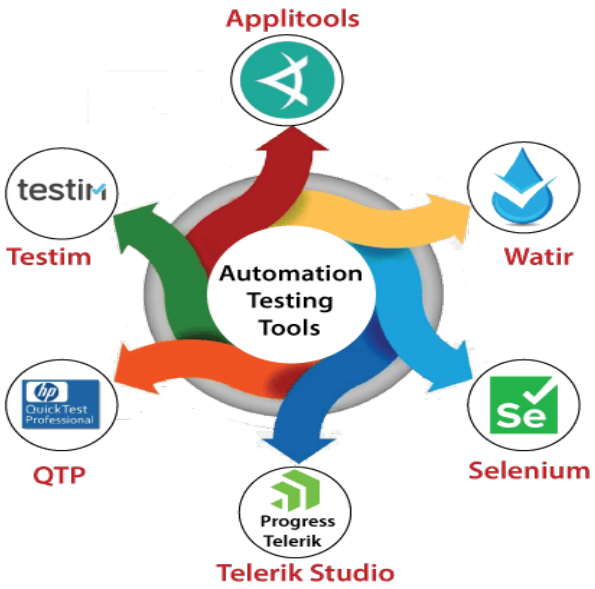
MOBILE APPLICATION

Our team extends feature-rich mobile apps for IOS, Android and Windows Phone for information on the go. We develop apps for business, games, location tracking, social media and various other requirements. Our team of

designers, developers and testers can help you from concept to final launch of your mobile application.

TESTING SERVICES

Our team of software application quality assurance experts and testers provides software testing as an independent service. Software testing is an essential part of our delivery process. We identify all the flaws, bugs and errors and debug them to deliver robust applications to our clients.



DATA ANALYTICS

Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.

team working on your software maintenance and support for the entire contract period.

PRODUCT MAINTENANCE AND SUPPORT

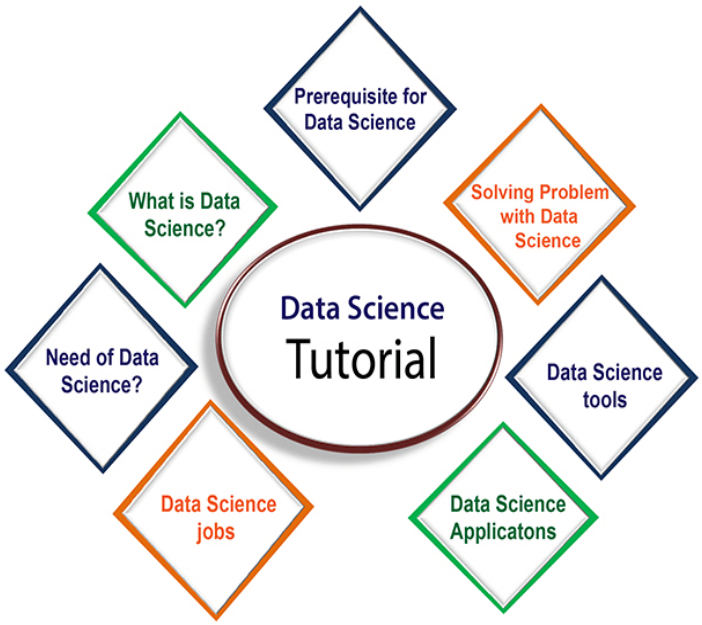
Our support and maintenance service is flexible just like our other service offerings, as per the complexities and needs of the project. From a need as small as allocating certain fixed number of hours for product or software maintenance per month, to a dedicated.

OUR PRODUCT

We Believe in our product which takes us to the higher level.

Data Science Tutorial for Beginners

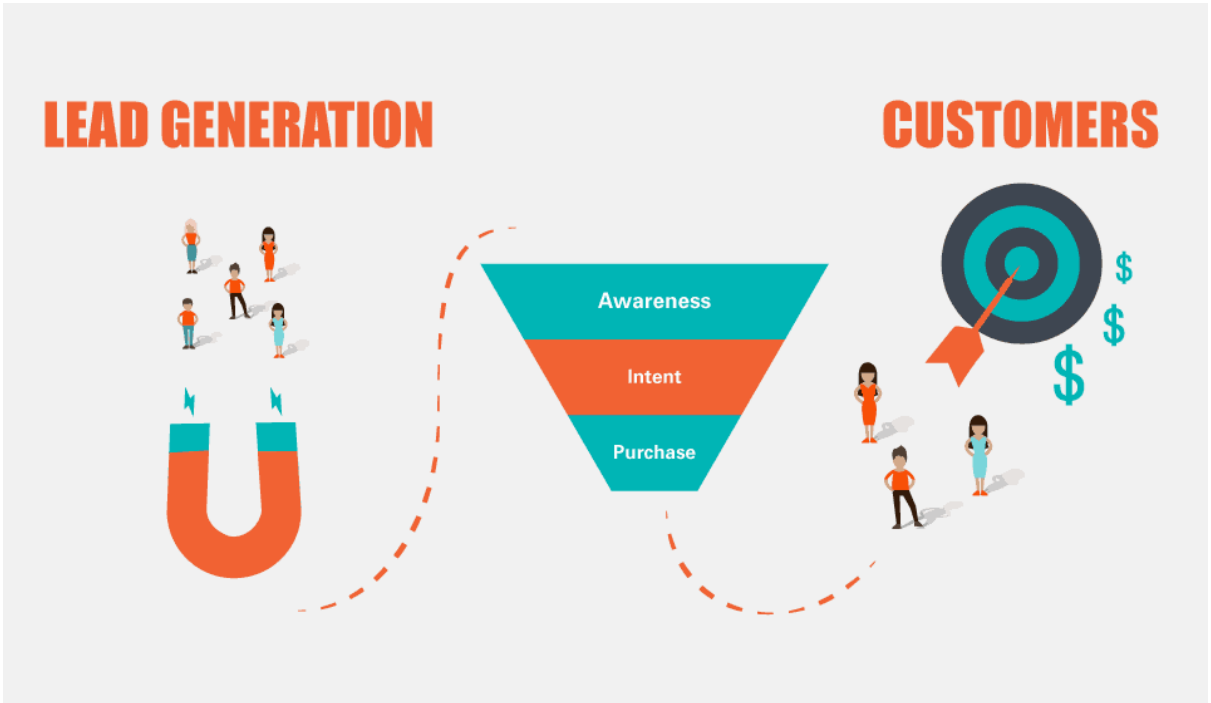
Data Science has become the most demanding job of the 21st century. Every organization is looking for candidates with knowledge of data science. In this tutorial, we are giving an introduction to data science, with data science Job roles, tools for data science, components of data science, application, etc.



The following article provides an outline for SEO Tools. SEO means Search Engine Optimization. The practice is to optimize your websites to position them highly in Google's or other search engine search results. Organic SEO is focused on improving the ranking of our websites.

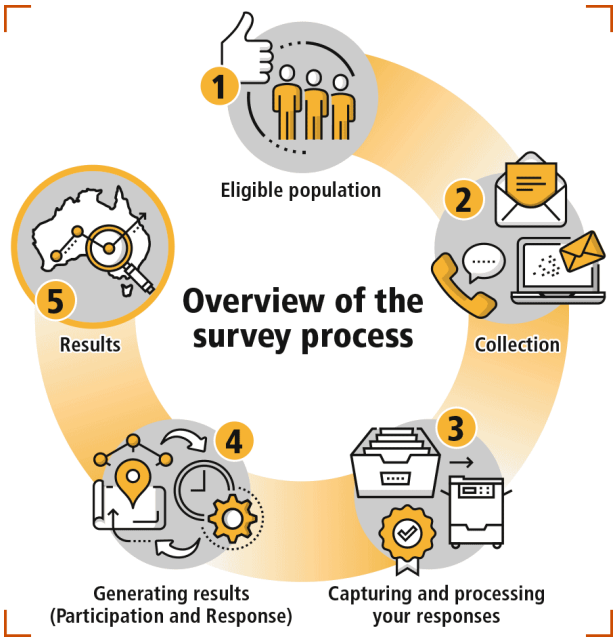
Lead Generation?

Lead generation refers to the process of generating interest among consumers for a product or service with the end goal of turning that interest into a sale. In the world of online marketing, lead generation often involves collecting a site visitor's contact information (the definition of a "lead"), usually through a web form or survey.



VARCONS TECHNOLOGIES PVT LTD Survey Tool (VARCONS TECHNOLOGIES PVT LTD)

Survey responses could also be provided via an online form and a telephony service.

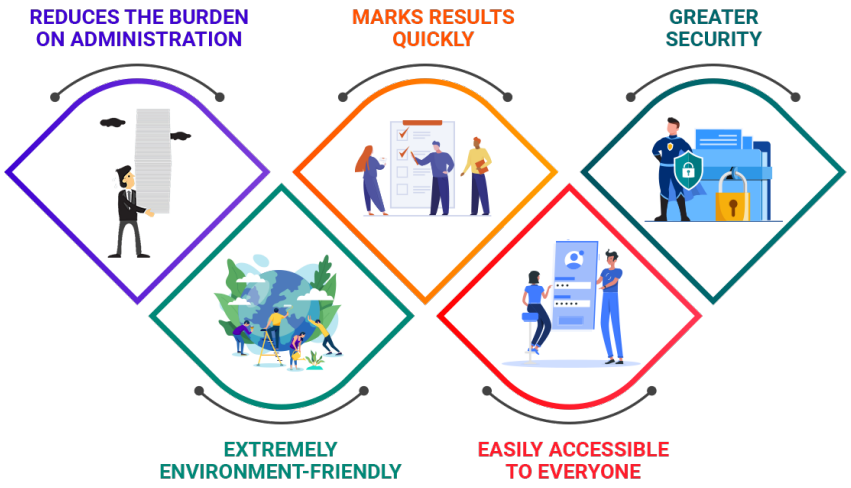


Human Resource Management (HRMS)

Human Resource Management more commonly known as HR has been defined in many different ways but at the heart of every definition is the human capital - the employee. HRM is defined as the effective management of people in an organization or company so that they may contribute to the company's or organization's business success.. HRM systems and policies are designed to maximize employee performance to achieve strategic objectives.

VARCONS TECHNOLOGIES PVT LTD Assessment Tool

Assessment Tool: the instrument that is used to collect data for each outcome. The actual product that is handed out to students for the purpose of assessing whether they have achieved a particular learning outcome(s).



OWNERSHIP PATTERN IN VARCONS TECHNOLOGIES PVT LTD:

Ownership pattern followed in Varcons Technologies Pvt Ltd corporation.

Table no 2.2 table representing ownership pattern

CHAIRMAN	
MANAGINING DIRECTOR	

JOINT MANAGING DIRECTOR	
DEPUTY MANAGING DIRECTOR	

OUR INVOLVEMENT:

Hruthvik HR Solutions is the Pioneer of consulting Business providing a holistic approach. Our structured process enables us to bring a unique approach. We work with clients to design their structures, roles, and responsibilities. We help them hire the right people and advise them on how to reward, develop and motivate their workforce & drive superior performance.

- Strategy planning
- Assessment
- Procurement
- Re-engineering solutions
- Planning, audits, best practices etc.

SOFTWARE DEVELOPMENT

Beside the hardware and network solutions; with design and development expertise in diverse platforms, best-of-breed tools and techniques, combined with industry best practices, Varcons Technologies Pvt Ltd offers scalable end-to-end application development and management solutions from requirement analysis for deployment and rollout. We are developing software, related to garments- production management, commercial jobs, buying, accounting software for trading, manufacturing house and conglomerates. We hope to come to you with desired software at a reasonable cost. Varcons Technologies Pvt Ltd services span the following application lifecycle stages:

APPLICATION DEVELOPMENT– Providing end-to-end development from

requirement analysis for deployment and rollout.

APPLICATION MAINTENANCE– Changing or enhancing software to meet changing or increasing business demands in the post-rollout phase of an application

APPLICATION SUPPORT– Providing first, second, third line support and on-call support. On-call support further includes Gold (24x7), Silver and Bronze support.

APPLICATION INTEGRATION/MIGRATION/TRANSFORMATION– Replacing, migrating and integrating legacy or bespoke systems with COTS products.

APPLICATION MANAGEMENT– The application management layer cuts across all software engineering activities listed above. APSIS takes complete ownership of the outsourced suite of applications as per the agreed scope and manages the support. This typically involves transition management, project management, proactive risk and scope change management, quality management, SLA management etc.

NETWORK & INFRASTRUCTURE SERVICES

Growth in the Solution Integration (SI) services market is fuelled by the need for seamless business processes across an organization's complete value chain of customers, partners, suppliers, and employees. Varcons Technologies Pvt Ltd services enable clients to identify, develop, and implement the best-fit solutions which are equipped to meet their changing business requirements.

Varcons Technologies Pvt Ltd Integration services offer to:

- Leverage IT investments
- Minimize risks
- Maximize compatibility
- Maximize interoperability

Varcons Technologies Pvt Ltd provides total project management, right from architectural design, integration, system and interface development to migration backed by world-class methodologies, well-defined solution frameworks and extensive integration experience with tier-1 service providers.

HARDWARE SALES & SUPPORT SERVICES

We offer servers, computers, computer accessories and services by sourcing from local market and from international market as well. Our team of experts is ready to serve you when you are worried due to lack of confidence in "commitment of service". You are hereby requested to call us for any kind of requirement of computers, computer parts and services whatever and whenever you need.

Varcons Technologies Pvt Ltd also offer Supply, Delivery & Installation of:

- Data Centre , Structure Cabling Systems.
- CISCO Switch, Firewall, Wireless & Routers.
- **KASPERSKY** antivirus and **UTM.**
- PC, Server & Storage.
- Server & Switch Equipment maintenance.

MANAGED SERVICES :

Varcons Technologies Pvt Ltd has the expertise and experience to manage an enabling infrastructure and applications and run outsourced operations for large telecom operators smoothly. Varcons Technologies Pvt Ltd Managed Services offerings cover the entire array of IT outsourcing services including networks, IT infrastructure, applications and business processes. This provides our customers the best of both worlds - control and flexibility over their information systems without either the pain or cost of running them.

AREARS OF EXPERTISE:

Main Strength of Varcons Technologies Pvt Ltd lies in the blend of professionals, specialized and highly focused operation. Increasing customer's awareness is the strength where it excels over its competitors. Our strength lies in our ability to blend current management practice and IT expertise into cost-effective Computer Aided Management Solutions, Products and Services. Varcons Technologies Pvt Ltd understands the need for skill transfer to client personnel. Our offers cover the following major areas:

• System analysis. • Business process re-engineering. • Process development localization. • Customized and target oriented Workflow design. • Specialization in Client / Server and Internet / Intranet application and technologies. • Automation of Financial Institutions, Telco-Operators, Electronic Medias & Business Institutions with the latest development. • Customized software development as ancillary product for deployed international software. • National software deployment - product development localization. • Network Monitoring/ Network Management Support. • Network, Security & Threat Management Solution. • Enterprise Server & Storage Solution.	• Hardware at competitive price. • Desktop maintenance and support. • Server maintenance and support. • Production operations. • Free Lancing. • Power Solution. • Outsourcing the employers. • Infrastructure Management Solution. • Data Centre Operations and Service Delivery. • Systems Integration. • Project Management. • Change/request implementation. • Project support. • Standby support. • Web based support & solution development. • E-business solutions. • Data conversion & data entry. • Overall automation consultancy. • Any Type of VDO editing Solutions.

ACHIEVEMENT /AWARDS IF ANY:

Varcons Technologies Pvt Ltd has participation in one of top 11 companies listed in the “2021 Best Software Development Agencies in Boston” by Expertise ranking agency.

INNOVATION FAIR 2020-21)

FUTURE GROWTH & PROSPECTS:

In the year 2021, IT industry in India about 15 million people are employed directly and indirectly. Direct employment in IT sector is generated by computer hardware and software industries. The indirect employment generation takes place through the adoption of information technology in other industries.

India and other South Asian countries areas preferred off shoring destination for many IT companies. In the same year, IT industry contributed 8% of the GDP. Information Technology is growing faster than ever in recent years not just in India, but throughout the world. There is a noticeable growth in the revenue fetched by E-commerce in the IT sector in India and the growth of IT will continue to mature in lines with the earlier trends'.

Yet again IT services are predicted to grow more in the year 2021 as the pandemic has fuelled digital spending. The Indian IT services sector will see a higher demand for digital transformation after it saw a plateau in 2020. Now that the customers focus more on digital services, online services and online work platforms there will be accelerated spends.With this kind of growth projection there will be a better hiring intent for employers in the first quarter of 2021. After pandemic, India’s hiring activities are getting better when compared to other economies and will continue to rise.

FUTURE OF IT INDUSTRY IN INDIA

Figure 2.3: Future of IT Industry in India

The global IT professional services market size was valued at USD 777.28 billion in 2021 and is expected to register a compound annual growth rate (CAGR) of 11.2% from 2022 to 2030. The rise of automation to eliminate mundane tasks and radical shifts in customer demand such as customized pricing and enhanced customer experience are pushing enterprises to implement IT services across the globe. Additionally, the COVID-19 pandemic tested the professional services industry by forcing them to implement remote working at a large scale and adjust their business strategies to the rapidly changing market conditions. Additionally, the COVID-19 outbreak also accelerated several technological changes across industries, where firms survived the pandemic with the help of technology by focusing on resource management and talent acquisition.

Final Words

India, as compared to other western countries, in the past has been considered as slow when it comes to IT and ITES development but with the Government now taking stringent steps to fast forward the implementation of new technologies. Joining hands with the Indian Government, Aerologic, is now one of the most chief companies in the country that works to strategise, implement and execute the next generation tech like Open Source Technologies, Machine

Learning, Artificial Intelligence, Smart Cities, Block chain, Internet of Things, SAP, Mobile APPs IOS, Android and Digital Marketing tools & more.

FRONT END

HTML – HYPER TEXT MARKUP LANGUAGE

Hypertext Markup Language (HTML) is the standard markup language for documents designed to be displayed in a **web browser**. It can be assisted by technologies such as Cascading Style Sheets (CSS) and **scripting languages** such as **JavaScript**.

Web browsers receive HTML documents from a **web server** or from local storage and **render** the documents into multimedia web pages. HTML describes the structure of a web page **semantically** and originally included cues for the appearance of the document.

HTML elements are the building blocks of HTML pages. With HTML constructs, **images** and other objects such as **interactive forms** may be embedded into the rendered page. HTML provides a means to create **structured documents** by denoting structural **semantics** for text such as headings, paragraphs, lists, **links**, quotes and other items. HTML elements are delineated by tags, written using **angle brackets**.

Tags such as `<img />` and `<input />` directly introduce content into the page. Other tags such as

`<p>` surround and provide information about document text and may include other tags as sub-elements.

Browsers do not display the HTML tags, but use them to interpret the content of the page.

5.1 HTML TAGS, STYLES, ATTRIBUTES

In this section we will discuss on components of HTML language such as,

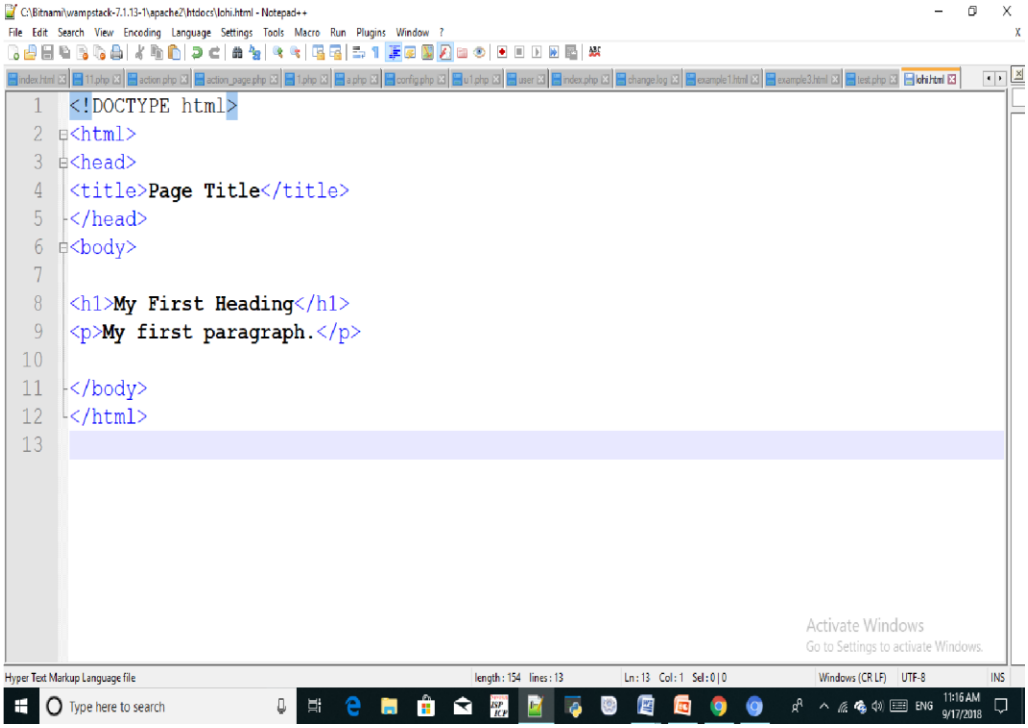
I. HTML Tags
II. HTML Styles
III. HTML attributes

5.1.1 HTML TAGS

HTML tags are element names surrounded by angle brackets: <tag name>content goes here...</tag name>

- HTML tags normally come **in pairs** like ** ** and ** **
- The first tag in a pair is the **start tag,** the second tag is the **end tag**
- The end tag is written like the start tag, but with a **forward slash** inserted before the tag name. Basic structure of HTML program is explained by following example,
- The **<!DOCTYPE html>** declaration defines this document to be HTML5
- The **<html>** element is the root element of an HTML page
- The **<head>** element contains meta information about the document
- The **<title>** element specifies a title for the document
- The **<body>** element contains the visible page content
- The **<h1>** element defines a large heading
- The ** ** element defines a paragraph
- The start tag is also called the **opening tag**, and the end tag the **closing tag.**
- The purpose of a web browser (Chrome, IE, Firefox, and Safari) is to read HTML documents and display them. The browser does not

display the HTML tags, but uses them to determine how to display the document.



Step1: write the code in notepad and save it in **.html**

Fig. 5.1: Screenshot for a sample HTML program

Step 2: Run the code in any browser like chrome, safari, and fire fox exc.

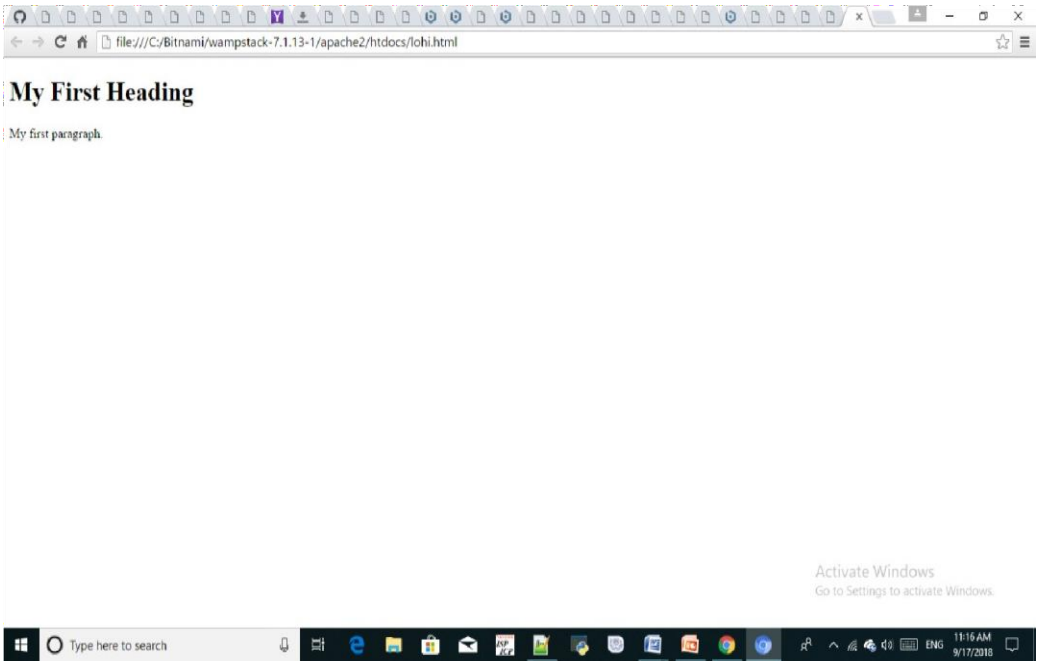


Fig. 5.2: Screenshot of sample HTML program output on browser

5.1.2 HTML ATTRIBUTES

- All HTML elements can have attributes
- Attributes provide additional information about an element
- Attributes are always specified in the start tag
- Attributes usually come in name/value pairs like: name="value"

Following sections we shall consider execution of simple HTML code and its output to illustrate attribute concepts.

a. PROGRAM TO ALLIGN A DATA

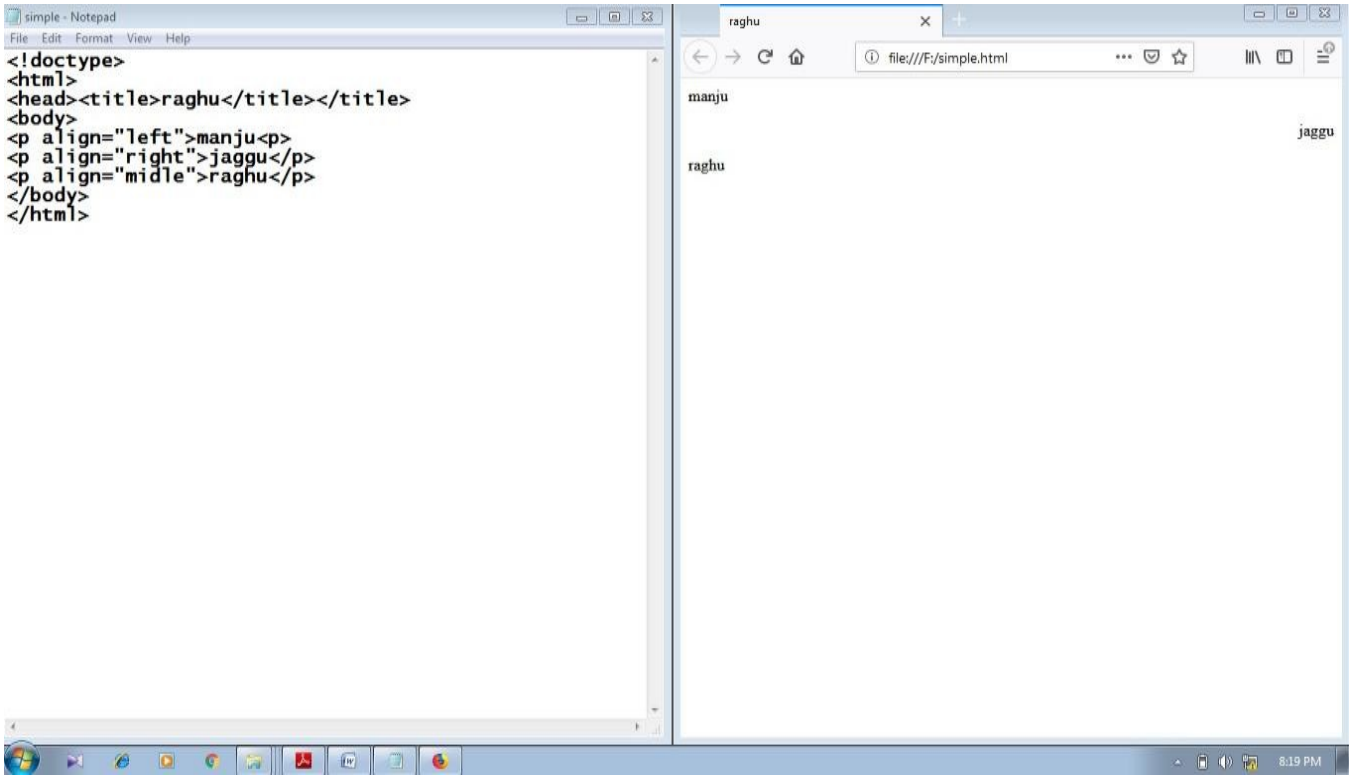


Fig. 5.3: Screenshot for a sample HTML program to align data

b. PROGRAM TO CREATE A FORM

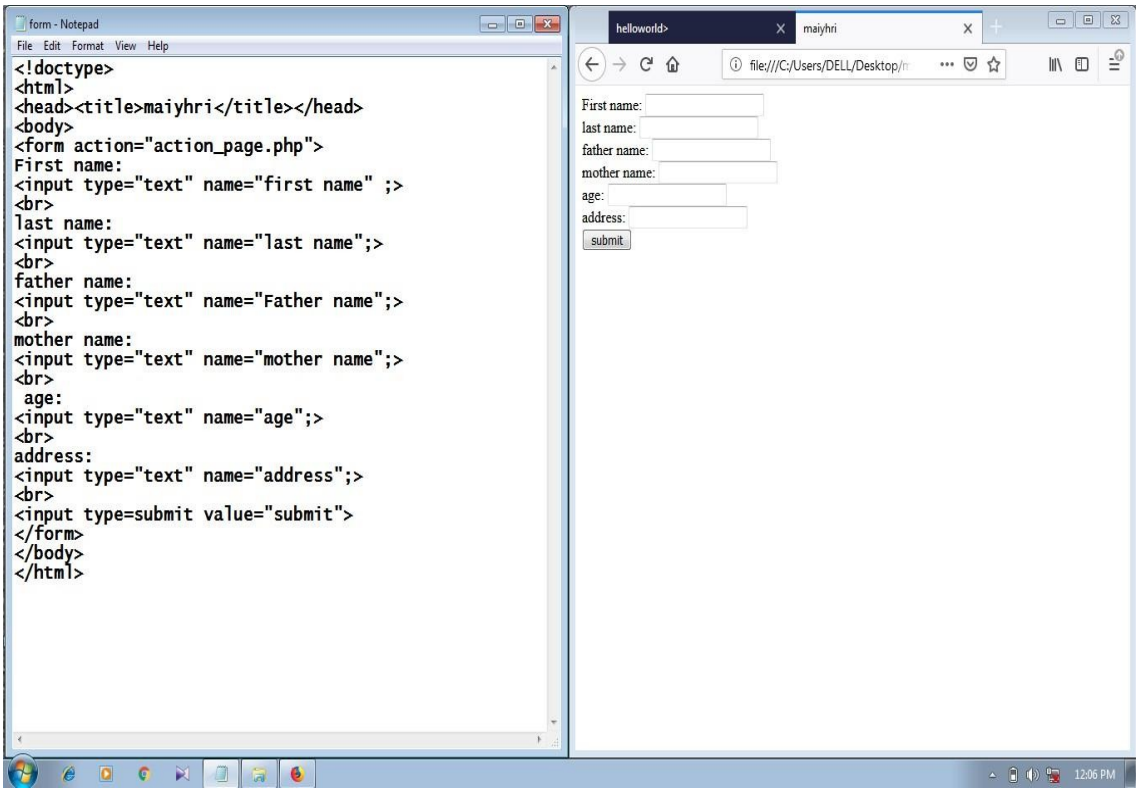
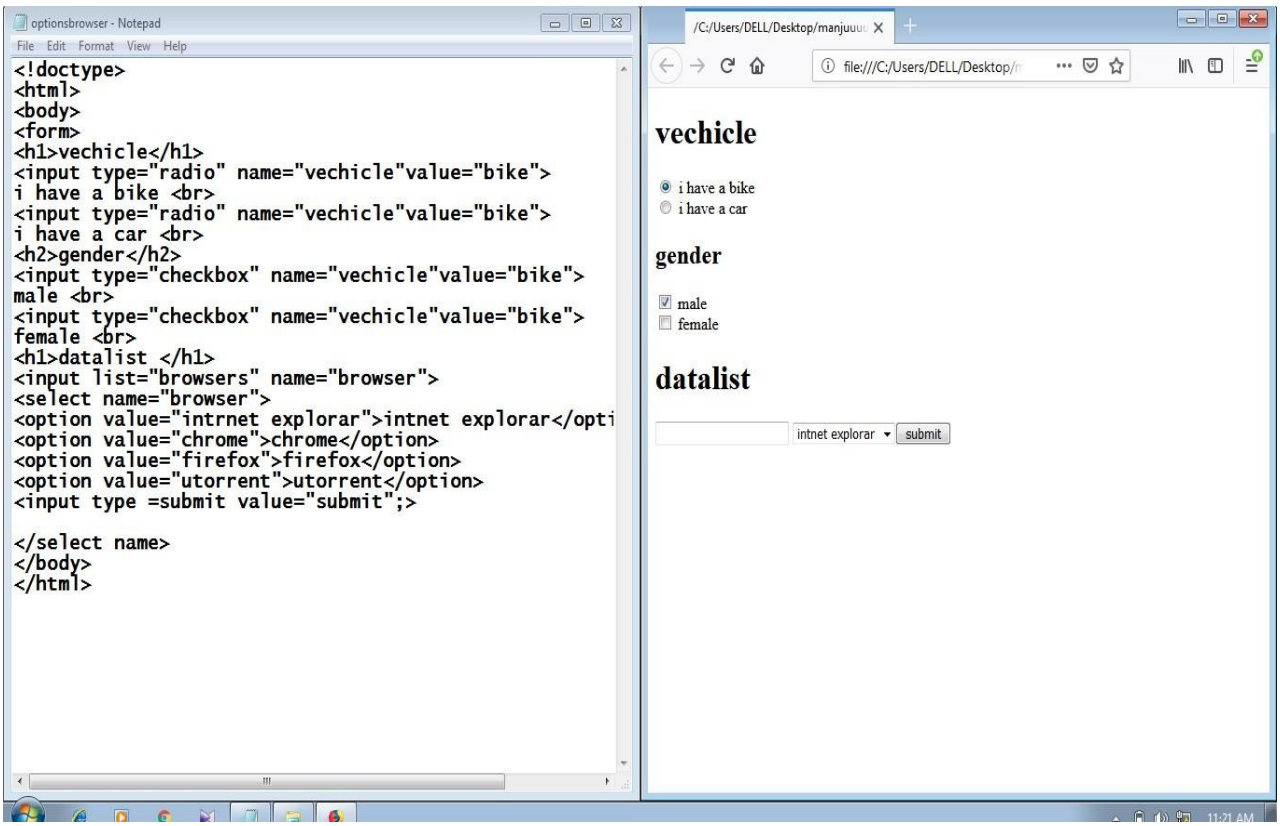


Fig. 5.4: Screenshot for a sample HTML program to create a form



c. PROGRAM TO SELECT OPTIONS USING RADIO AND CHECK BUTTONS

Fig. 5.5: Screenshot for a sample HTML program to check buttons

5.2 CSS TUTORIAL

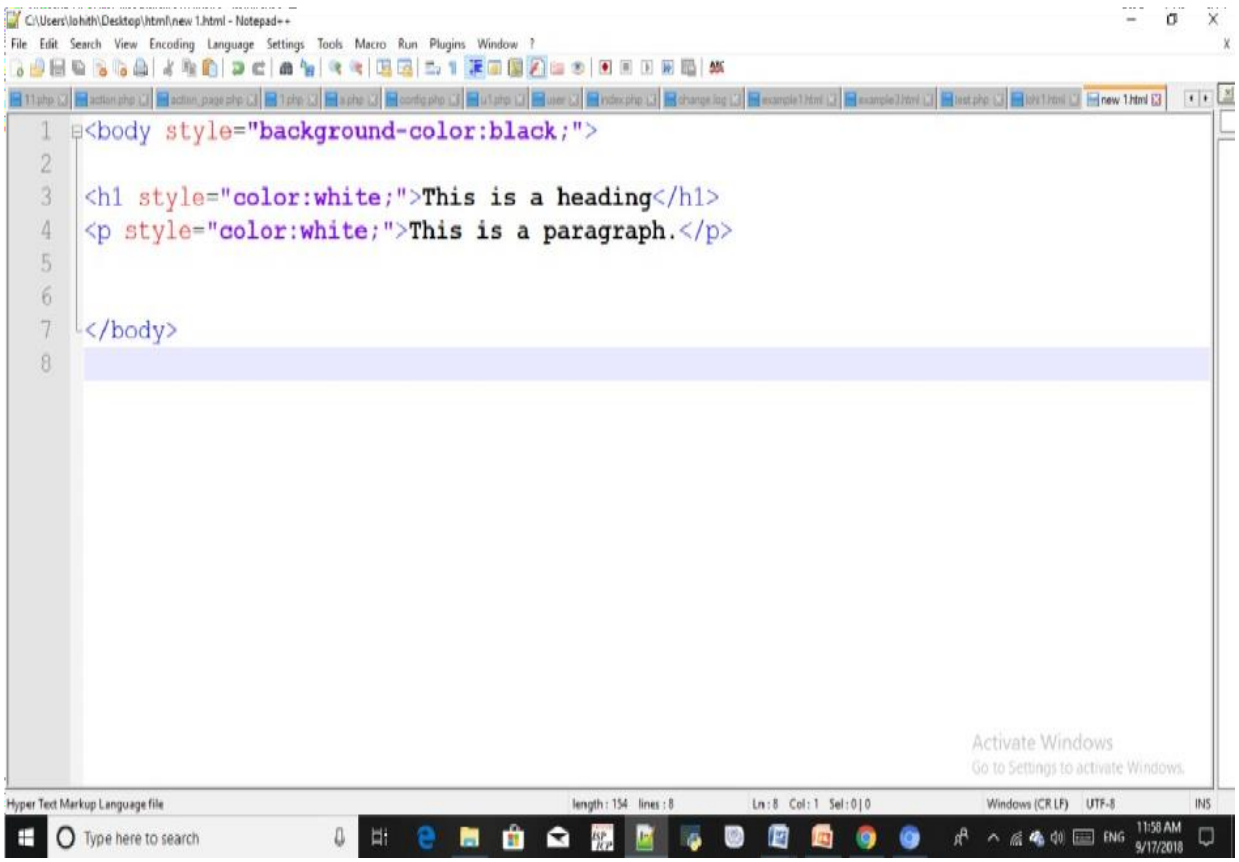
CSS is a language that describes the style of an HTML document. CSS describes how HTML element should be displayed.

5.2.1 HTML BACKGROUND COLOR

The background-color property defines the background color for an HTML element. This example sets the background color for a page to powder blue:

Example:

```
<body style="background-color: powder blue;">
<h1 style="color:white;">This is a heading</h1>
<p style="color: white;">This is a paragraph. </p> </body>
```



Step 1: write a simple code using HTML style attributes

Fig. 5.6: Screenshot for a sample HTML program for style attributes

Step 2: output for this code using HTML style attributes we get as follow.

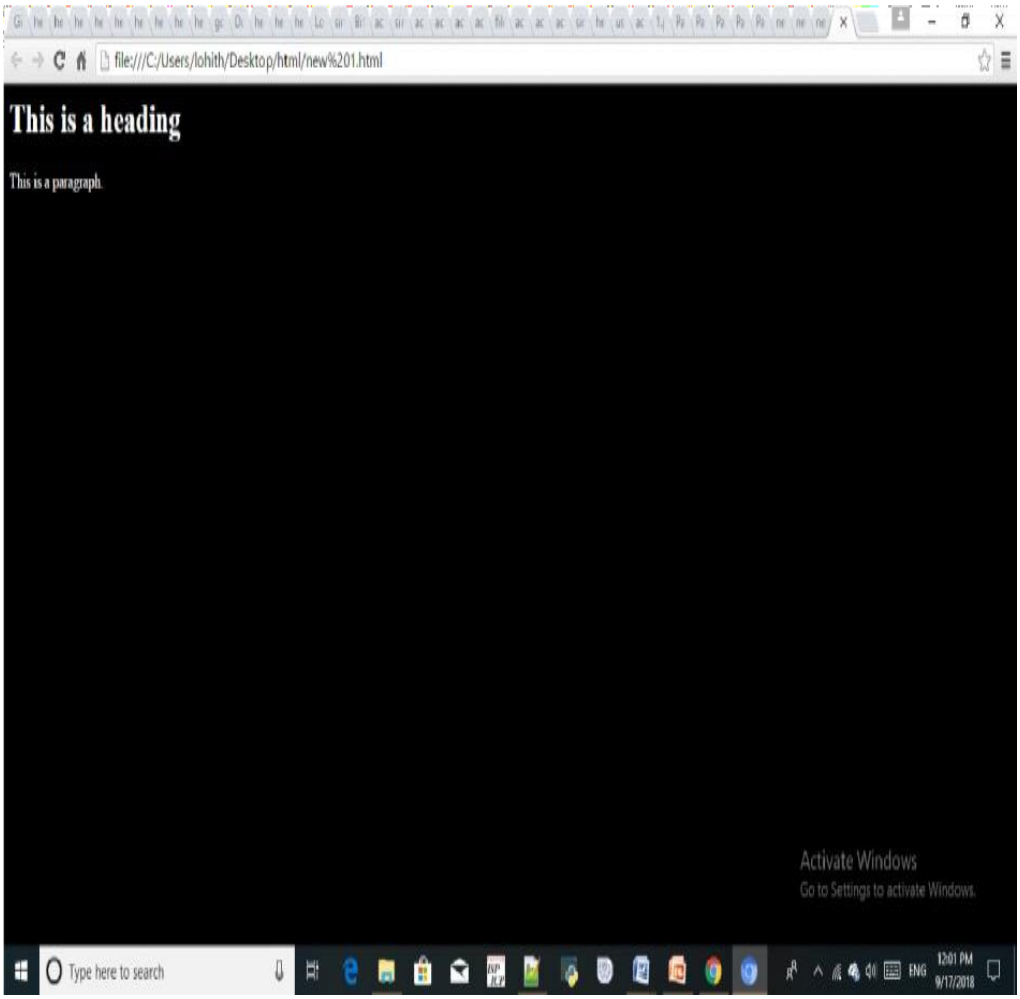


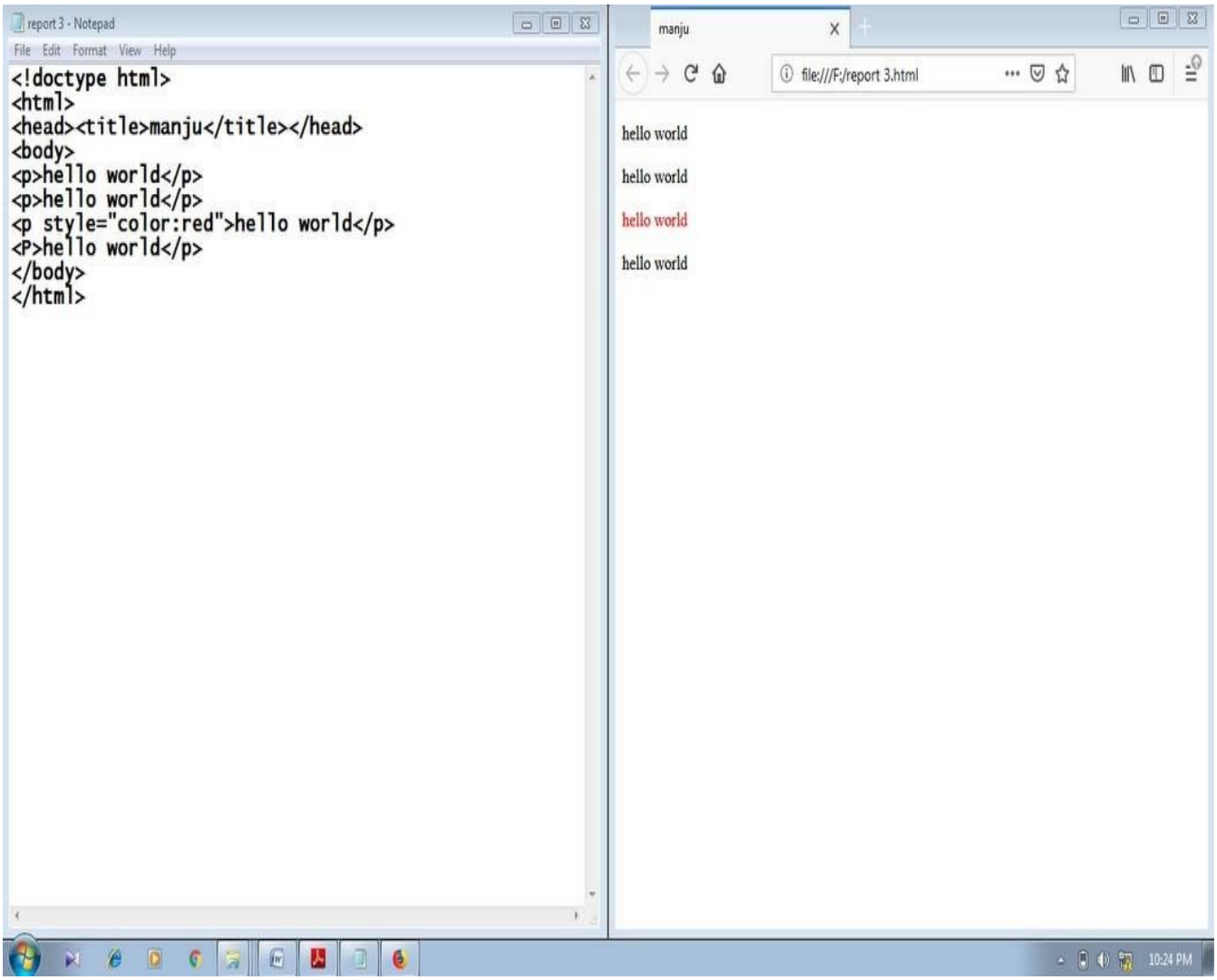
Fig. 5.7: Screenshot for a sample HTML program for style attributes

5.3 TYPES IN CSS

5.3.1 INLINE CSS o It contains the CSS property in the body section attached with element is known as inline CSS. o This kind of style is specified within a HTML tag using style attribute. o This can be used when a single HTML document must be

styled uniquely.

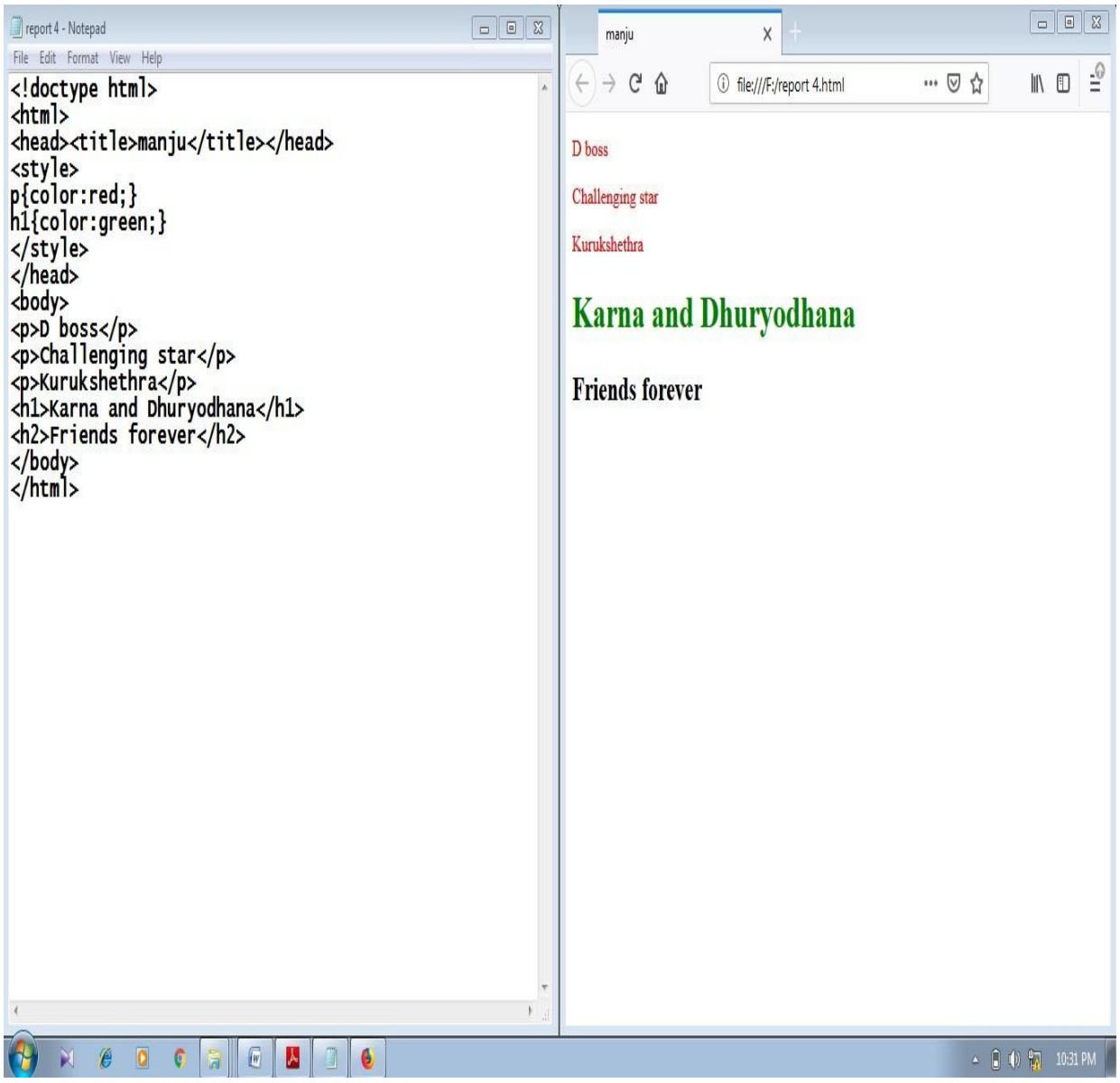
Fig. 5.8: Screenshot for a sample HTML program for inline



CSS 5.3.2 INTERNAL CSS

o Internal style sheet is used to unique style for a single document. o It is defined in <head> section of the HTML page inside the <style> tag.

Fig. 5.9: Screenshot for sample HTML program for internal CSS



5.3.3 EXTERNAL CSS

External CSS is a file that contains only CSS code and is saved with a ".CSS" file extension. This CSS file is then referenced in our HTML using the <link> instead of <style>. In External CSS we create a .css file and use it in our HTML page as per our requirements. Generally external Cascading Style Sheets are used whenever we have many HTML attributes and we can use them as required; there is no need to rewrite the CSS style again and again in a complete body of HTML that inherits the property of the CSS file. There are two ways to create a CSS file. The first is to write the CSS code in Notepad and save it as a .css file, the second one is to directly add the style sheet in our Solution Explorer and direct Visual Studio to use it on our HTML page.

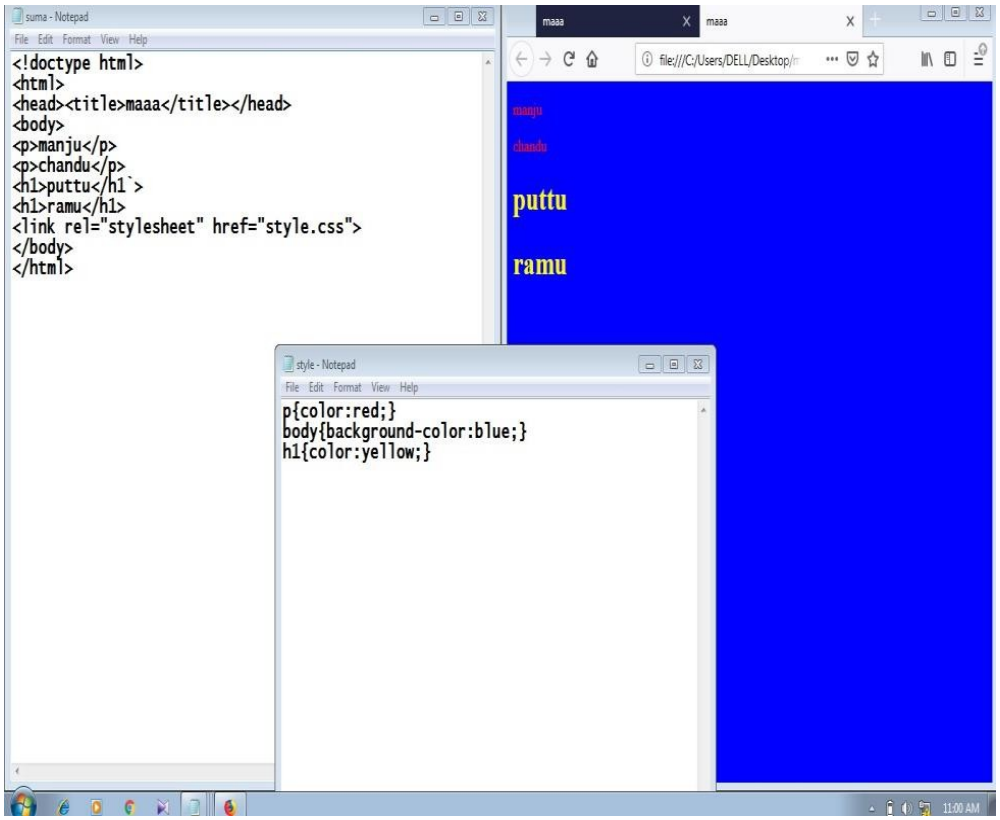


Fig. 5.10: Screenshot for a sample HTML program for external CSS

CHAPTER-3

1.1 Motivations

Being extremely interested in everything having a relation with the Machine Learning, the independent project was a great occasion to give me the time to learn and confirm my interest for this field. The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities. We can use Machine Learning in Finance, Medicine, almost everywhere. That's why I decided to conduct my project around the Machine Learning.

1.2 Idea

This project was motivated by my desire to investigate the sentiment analysis field of machine learning since it allows to approach natural language processing which is a very hot topic actually. Following my previous experience where it was about classifying short music according to their emotion, I applied the same idea with tweets and try to figure out which is positive or negative.

1.3 Sources

Because I truly think that sharing sources and knowledge's allow to help others but also ourselves, the sources of the project are available at the following link: https://github.com/marclamberti/TwitterEmotionAnalysis
Feel free to give me your point of view or ideas for anything you want. I used ipython notebook which is very useful to understand the entire process of my project since you can follow each step with the

corresponding code. Here is the direct link to it: http://nbviewer.ipython.org/github/marclamberti/TwitterEmotionAnalysis/blob/master/TwitterSenti mentAnalysis.ipynb

2. The Project

Sentiment analysis, also refers as opinion mining, is a sub machine learning task where we want to determine which is the general sentiment of a given document. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinion about a selling objects, or predict stock markets for a given company like, if most people think positive about it, possibly its stock markets will increase, and so on. Sentiment analysis is actually far from to be solved since the language is very complex (objectivity/subjectivity, negation, vocabulary, grammar,...) but it is also why it is very interesting to working on.

In this project I choose to try to classify tweets from Twitter into “positive” or “negative” sentiment by building a model based on probabilities. Twitter is a microblogging website where people can share their feelings quickly and spontaneously by sending a tweets limited by 140 characters. You can directly address a tweet to someone by adding the target sign “@” or participate to a topic by adding an hastag “#” to your tweet. Because of the usage of Twitter, it is a perfect source of data to determine the current overall opinion about anything.

2.1 Data

To gather the data many options are possible. In some previous paper researches, they built a program to collect automatically a corpus of tweets based on two classes, “positive” and “negative”, by querying

Twitter with two type of emoticons:

- Happy emoticons, such as “:)”, “:P”, “:-)” etc.
- Sad emoticons, such as “:(“, “:’(”, “=(“.

Others make their own dataset of tweets my collecting and annotating them manually which very long and fastidious.

Additionally to find a way of getting a corpus of tweets, we need to take of having a balanced data set, meaning we should have an equal number of positive and negative tweets, but it needs also to be large enough. Indeed, more the data we have, more we can train our classifier and more the accuracy will be.

After many researches, I found a dataset of 1578612 tweets in english coming from two sources: Kaggle and Sentiment140. It is composed of four columns that are *ItemID*, *Sentiment*, *SentimentSource* and *SentimentText*. We are only interested by the *Sentiment* column corresponding to our label class taking a binary value, 0 if the tweet is negative, 1 if the tweet is positive and the *SentimentText* columns containing the tweets in a raw format.

	ItemID	Sentiment	SentimentSource	SentimentText
0	1	0	Sentiment140	is so sad for my APL friend.....
1	2	0	Sentiment140	I missed the New Moon trailer...
2	3	1	Sentiment140	omg its already 7:30 :O
3	4	0	Sentiment140	.. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...
4	5	0	Sentiment140	i think mi bf is cheating on me!!! T_T
5	6	0	Sentiment140	or i just worry too much?
6	7	1	Sentiment140	Juuuuuuuuuuuuuuuussssst Chillin!!
7	8	0	Sentiment140	Sunny Again Work Tomorrow :-| TV Tonight
8	9	1	Sentiment140	handed in my uniform today . i miss you already
9	10	1	Sentiment140	hmmmm.... i wonder how she my number @-)

Table 2.1.1: Example of twitter posts annotated with their corresponding sentiment, 0 if it is negative, 1 if it is positive.

In the *Table 2.1.1* showing the first ten twitter posts we can already notice some particularities and difficulties that we are going to encounter during the preprocessing steps.

- The presence of **acronyms** "bf" or more complicated "APL". Does it means apple ? Apple (the company) ? In this context we have "friend" after so we could think that he refers to his smartphone and so Apple, but what about if the word "friend" was not here ?

- The presence of **sequences of repeated characters** such as "Juuuuuuuuuuuuuuuussssst", "hmmmm". In general when we repeat several characters in a word, it is to emphasize it, to increase its impact.

- The presence of **emoticons**, ":O", "T_T", ":- |" and much more, give insights about user's moods.

- **Spelling mistakes** and "**urban grammar**" like "im gunna" or "mi".

- The presence of **nouns** such as "TV", "New Moon".

Furthermore, we can also add,

- People also indicate their moods, emotions, states, between two *such as,* *\cries*,
  hummin, *sigh*.
- The negation, “can't”, “cannot”, “don't”, “haven't” that we need to handle like: “I don’t like chocolate”, “like” in this case is negative.

We could also be interested by the grammar structure of the tweets, or if a tweet is subjective/objective and so on. As you can see, it is **extremely complex** to deal with languages and even more when we want to analyse text typed by users on the Internet because people don’t take care of making sentences that are grammatically correct and use a ton of acronyms and words that are more or less english in our case.

We can visualize a bit more the dataset by making a chart of how many positive and negative tweets does it contains,

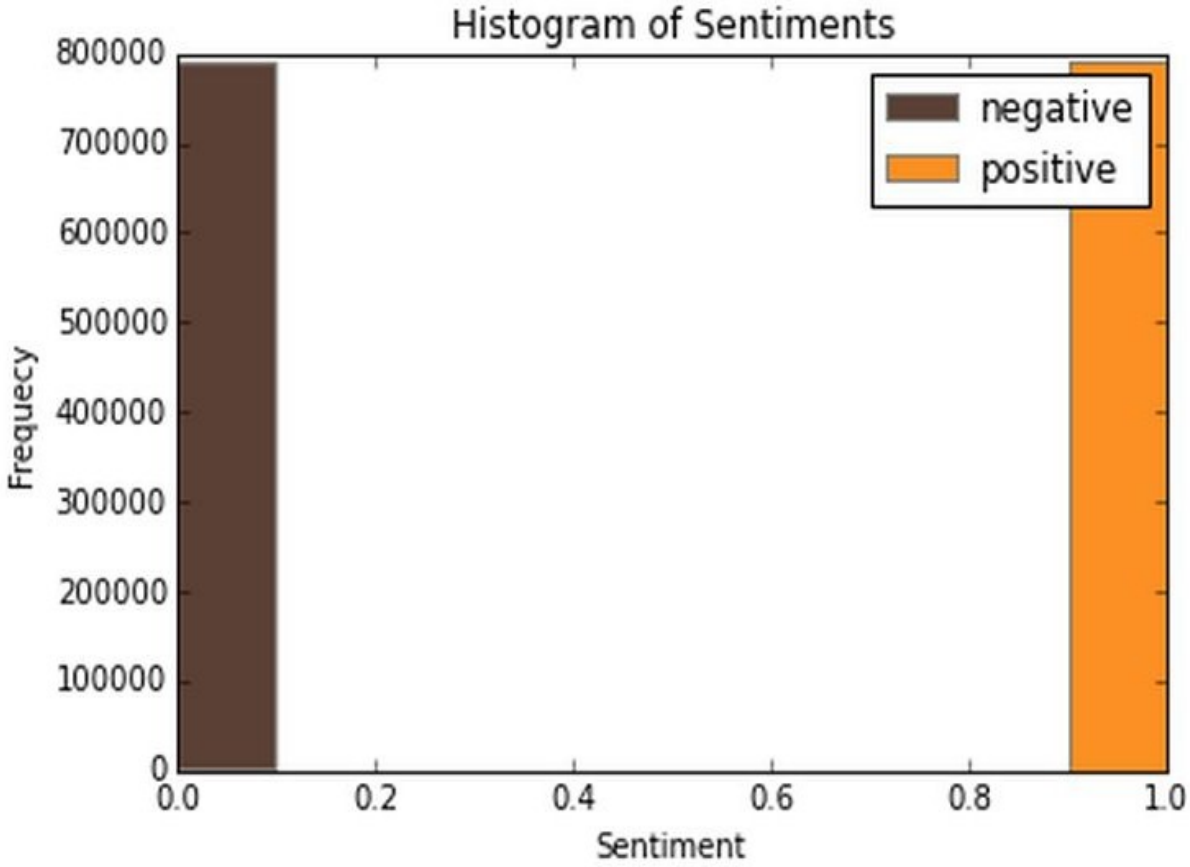


Figure 2.1.1: Histogram of the tweets according to their sentiment

We have exactly 790177 positive tweets and 788435 negative tweets which signify that the dataset is well- balanced. There is also no duplicates.

Finally, let's recall the Twitter terminology since we are going to have to deal with in the tweets:

- Hashtag: A hashtag is any word or phrase immediately preceded by the # symbol. When you click on a hashtag, you'll see other Tweets containing the same keyword or topic.
- @username: A username is how you're identified on Twitter, and is always preceded immediately by the @ symbol. For instance, Katy Perry is @katyperry.
- MT: Similar to RT (Retweet), an abbreviation for "Modified Tweet." Placed before the Retweeted text when users manually retweet a message with modifications, for example shortening a Tweet.
- Retweet: RT, A Tweet that you forward to your followers is known as a Retweet. Often used to pass along news or other valuable discoveries on Twitter, Retweets always retain original attribution.
- Emoticons: Composed using punctuation and letters, they are used to express emotions concisely, ";) :) ...".

Now we have the corpus of tweets, we need to use other resources to make easier the pre- processing step.

2.2 Resources

In order to facilitate the pre- processing part of the data, we introduce five resources which are,

- An **emoticon dictionary** regrouping 132 of the most used emoticons in western with their sentiment, negative or positive.
- An **acronym dictionary** of 5465 acronyms with their translation.
- A **stop word dictionary** corresponding to words which are filtered out before or after processing of natural language data because they are not useful in our case.
- A **positive and negative word dictionaries** given the polarity

(sentiment out- of- context) of words.

- A **negative contractions and auxiliaries dictionary** which will be used to detect negation in a given tweet such as “don’t”, “can’t”, “cannot”, etc.

The introduction of these resources will allow to uniform tweets and remove some of their complexities with the acronym dictionary for instance because a lot of acronyms are used in tweets. The positive and negative word dictionaries could be useful to increase (or not) the accuracy score of the classifier. The emoticon dictionary has been built from wikipedia with each emoticon annotated manually. The stop word dictionary contains 635 words such as “the”, “of”, “without”. Normally they should not be useful for classifying tweets according to their sentiment but it is possible that they are.

Also we use Python 2.7 (https://www.python.org/) which is a programming language widely used in data science and scikit- learn (http://scikit- learn.org/) a very complete and useful library for machine learning containing every technique, methods we need and the website is also full of tutorials well explained. With Python, the libraries, Numpy (http://www.numpy.org/) and Panda (http://pandas.pydata.org/) for manipulating data easily and intuitively are just essential.

2.3 Pre-processing

Now that we have the corpus of tweets and all the resources that could be useful, we can pre- process the tweets. It is a very important since all the modifications that we are going to during this process will directly impact the classifier's performance. The pre- processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The result of pre- processing will be consistent and uniform data that are workable to maximize the classifier's performance.

All of the tweets are pre- processed by passing through the following steps in the same order.

2.3.1 Emoticons

We replace all emoticons by their sentiment polarity **||pos||** and **||neg||** using the emoticon dictionary. To do the replacement, we pass through each tweet and by using a regex we find out if it contains emoticons, if yes they are replaced by their corresponding polarity.

	ItemID	Sentiment	SentimentSource	SentimentText
0	1	0	Sentiment140	is so sad for my APL friend.....
1	2	0	Sentiment140	I missed the New Moon trailer...
2	3	1	Sentiment140	omg its already 7:30 :O
3	4	0	Sentiment140	.. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...
4	5	0	Sentiment140	i think mi bf is cheating on me!!! T_T
5	6	0	Sentiment140	or i just worry too much?
6	7	1	Sentiment140	Juuuuuuuuuuuuuuuuussssst Chillin!!
7	8	0	Sentiment140	Sunny Again Work Tomorrow :-| TV Tonight
8	9	1	Sentiment140	handed in my uniform today . i miss you already
9	10	1	Sentiment140	hmmmm.... i wonder how she my number @-)

Table 2.3.1.1: Before processing emoticons, list of tweets where some of them contain emoticons.

	ItemID	Sentiment	SentimentSource	SentimentText
0	1	0	Sentiment140	is so sad for my APL friend.....
1	2	0	Sentiment140	I missed the New Moon trailer...
2	3	1	Sentiment140	omg its already 7:30 ||pos||
3	4	0	Sentiment140	.. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...
4	5	0	Sentiment140	i think mi bf is cheating on me!!! ||neg||
5	6	0	Sentiment140	or i just worry too much?
6	7	1	Sentiment140	Juuuuuuuuuuuuuuuuussssst Chillin!!
7	8	0	Sentiment140	Sunny Again Work Tomorrow ||neg|| TV Tonight
8	9	1	Sentiment140	handed in my uniform today . i miss you already
9	10	1	Sentiment140	hmmmm.... i wonder how she my number ||pos||

Table 2.3.1.2: After processing emoticons, they have been replaced by their corresponding tag

The data set contains 19469 positive emoticons and 11025 negative emotions.

2.3.2 URLs

We replace all URLs with the tag **||url||**. There is about 73824 urls in the data set and we proceed as the same way we did for the emoticons.

	ItemID	Sentiment	SentimentSource	SentimentText
50	51	0	Sentiment140	baddest day eveer.
51	52	1	Sentiment140	bathroom is clean..... now on to more enjoyable tasks.....
52	53	1	Sentiment140	boom boom pow
53	54	0	Sentiment140	but i'm proud.
54	55	0	Sentiment140	congrats to helio though
55	56	0	Sentiment140	David must be hospitalized for five days end of July (palatine tonsils). I will probably never see Katie in concert.
56	57	0	Sentiment140	friends are leaving me 'cause of this stupid love http://bit.ly/ZoxZC
57	58	1	Sentiment140	go give ur mom a hug right now. http://bit.ly/azFwv
58	59	1	Sentiment140	Going To See Harry Sunday Happiness
59	60	0	Sentiment140	Hand quilting it is then...

Table 2.3.2.1: Tweets before processing URLs.

	ItemID	Sentiment	SentimentSource	SentimentText
50	51	0	Sentiment140	baddest day eveer.
51	52	1	Sentiment140	bathroom is clean..... now on to more enjoyable tasks.....
52	53	1	Sentiment140	boom boom pow
53	54	0	Sentiment140	but i'm proud.
54	55	0	Sentiment140	congrats to helio though
55	56	0	Sentiment140	David must be hospitalized for five days end of July (palatine tonsils). I will probably never see Katie in concert.
56	57	0	Sentiment140	friends are leaving me 'cause of this stupid love ||url||
57	58	1	Sentiment140	go give ur mom a hug right now. ||url||
58	59	1	Sentiment140	Going To See Harry Sunday Happiness
59	60	0	Sentiment140	Hand quilting it is then...

Table 2.3.2.2: Tweets after processing URLs.

2.3.3 Unicode

For simplicity and because the ASCII table should be sufficient, we choose to remove any unicode character that could be misleading for the classifier.

	ItemID	Sentiment	SentimentSource	SentimentText
1578592	1578608	1	Sentiment140	'Zu SpÃ¤t' by Die Ã„rzte. One of the best bands ever
1578593	1578609	1	Sentiment140	Zuma bitch tomorrow. Have a wonderful night everyone goodnight.
1578594	1578610	0	Sentiment140	zummie's couch tour was amazing....to bad i had to leave early
1578595	1578611	0	Sentiment140	ZuneHD looks great! OLED screen @720p, HDMI, only issue is that I have an iPhone and 2 iPods . MAKE IT A PHONE and ill buy it @micro...
1578596	1578612	1	Sentiment140	zup there ! learning a new magic trick
1578597	1578613	1	Sentiment140	zyklonic showers *evil*
1578598	1578614	1	Sentiment140	ZZ Top â€“ I Thank You ...@hawaiibuzz .....Thanks for your music and for your ear(s) ...ALL !!!! Have a fab... â™« ||url||
1578599	1578615	0	Sentiment140	zzz time. Just wish my love could B nxt 2 me
1578600	1578616	1	Sentiment140	zzz twitter. good day today. got a lot accomplished. imstorm. got into it w yet another girl. dress shopping tmrw
1578601	1578617	1	Sentiment140	zzz's time, goodnight. ||url||

Table 2.3.3.1: Tweets before processing Unicode.

	ItemID	Sentiment	SentimentSource	SentimentText
1578592	1578608	1	Sentiment140	'Zu Spt' by Die rzte. One of the best bands ever
1578593	1578609	1	Sentiment140	Zuma bitch tomorrow. Have a wonderful night everyone goodnight.
1578594	1578610	0	Sentiment140	zummie's couch tour was amazing....to bad i had to leave early
1578595	1578611	0	Sentiment140	ZuneHD looks great! OLED screen @720p, HDMI, only issue is that I have an iPhone and 2 iPods . MAKE IT A PHONE and ill buy it @micro...
1578596	1578612	1	Sentiment140	zup there ! learning a new magic trick
1578597	1578613	1	Sentiment140	zyklonic showers *evil*
1578598	1578614	1	Sentiment140	ZZ Top I Thank You ...@hawaiibuzz .....Thanks for your music and for your ear(s) ...ALL !!!! Have a fab... ||url||
1578599	1578615	0	Sentiment140	zzz time. Just wish my love could B nxt 2 me
1578600	1578616	1	Sentiment140	zzz twitter. good day today. got a lot accomplished. imstorm. got into it w yet another girl. dress shopping tmrw
1578601	1578617	1	Sentiment140	zzz's time, goodnight. ||url||

Table 2.3.3.1: Tweets after processing Unicode.

2.3.4 HTML entities

HTML entities are characters reserved in HTML. We need to decode them in order to have characters entities to make them understandable.

```
'   Cannot get chatroom feature to work.   Updated Java to 10, checked ports, etc. I can see video, but in the &quot;chat,&quo
t; only a spinning circle.   '
```

Figure 2.3.4.1: A tweet before processing HTML entities.

```
u'   Cannot get chatroom feature to work.   Updated Java to 10, checked ports, etc. I can see video, but in the "chat," only a
spinning circle.   '
```

Figure 2.3.4.2: A tweet after processing HTML entities.

2.3.5 Case

The case is something that can appears useless but in fact it is really important for distinguish proper noun and other kind of words. Indeed: “General Motor” is the same thing that “general motor”, or “MSc” and “msc”. So reduce all letters to lowercase should be normally done wisely. In this

project, for simplicity we will not take care of that since we assume that it should not impact too much the classifier's performance.

	ItemID	Sentiment	SentimentSource	SentimentText
1578592	1578608	1	Sentiment140	'Zu Spt' by Die rzte. One of the best bands ever
1578593	1578609	1	Sentiment140	Zuma bitch tomorrow. Have a wonderful night everyone goodnight.
1578594	1578610	0	Sentiment140	zummie's couch tour was amazing....to bad i had to leave early
1578595	1578611	0	Sentiment140	ZuneHD looks great! OLED screen @720p, HDMI, only issue is that I have an iPhone and 2 iPods . MAKE IT A PHONE and ill buy it @micro...
1578596	1578612	1	Sentiment140	zup there ! learning a new magic trick
1578597	1578613	1	Sentiment140	zyklonic showers *evil*
1578598	1578614	1	Sentiment140	ZZ Top I Thank You ...@hawaiibuzz .....Thanks for your music and for your ear(s) ...ALL !!!! Have a fab... ||url||
1578599	1578615	0	Sentiment140	zzz time. Just wish my love could B nxt 2 me
1578600	1578616	1	Sentiment140	zzz twitter. good day today. got a lot accomplished. imstorm. got into it w yet another girl. dress shopping tmrw
1578601	1578617	1	Sentiment140	zzz's time, goodnight. ||url||

Table 2.3.5.1: Tweets before processing lowercase.

	ItemID	Sentiment	SentimentSource	SentimentText
0	1	0	Sentiment140	is so sad for my apl friend.....
1	2	0	Sentiment140	i missed the new moon trailer...
2	3	1	Sentiment140	omg its already 7:30 ||pos||
3	4	0	Sentiment140	.. omgaga. im sooo im gunna cry. i've been at this dentist since 11.. i was suposed 2 just get a crown put on (30mins)...
4	5	0	Sentiment140	i think mi bf is cheating on me!!! ||neg||
5	6	0	Sentiment140	or i just worry too much?
6	7	1	Sentiment140	juuuuuuuuuuuuuuuussssst chillin!!
7	8	0	Sentiment140	sunny again work tomorrow ||neg|| tv tonight
8	9	1	Sentiment140	handed in my uniform today . i miss you already
9	10	1	Sentiment140	hmmmm.... i wonder how she my number ||pos||

Table 2.3.5.2: Tweets after processing lowercase.

2.3.6 Targets

The target corresponds to usernames in twitter preceded by "@" symbol. It is used to address a tweet to someone or just grab the attention. We replace all usernames/targets by the tag **||target||**. Notice that in the data set we have 735757 targets.

	ItemID	Sentiment	SentimentSource	SentimentText
45	46	1	Sentiment140	@ginaaa <3 go to the show tonight
46	47	0	Sentiment140	@spiral_galaxy @ymptweet it really makes me sad when i look at muslims reality now
47	48	0	Sentiment140	- all time low shall be my motivation for the rest of the week.
48	49	0	Sentiment140	and the entertainment is over, someone complained properly.. @rupturerapture experimental you say? he should experiment with a me...
49	50	0	Sentiment140	another year of lakers .. that's neither magic nor fun ...
50	51	0	Sentiment140	baddest day eveer.
51	52	1	Sentiment140	bathroom is clean..... now on to more enjoyable tasks.....
52	53	1	Sentiment140	boom boom pow
53	54	0	Sentiment140	but i'm proud.
54	55	0	Sentiment140	congrats to helio though

Table 2.3.6.1: Tweets before processing targets.

	ItemID	Sentiment	SentimentSource	SentimentText
45	46	1	Sentiment140	||target|| <3 go to the show tonight
46	47	0	Sentiment140	||target|| ||target|| it really makes me sad when i look at muslims reality now
47	48	0	Sentiment140	- all time low shall be my motivation for the rest of the week.
48	49	0	Sentiment140	and the entertainment is over, someone complained properly.. ||target|| experimental you say? he should experiment with a melody...
49	50	0	Sentiment140	another year of lakers .. that's neither magic nor fun ...
50	51	0	Sentiment140	baddest day eveer.
51	52	1	Sentiment140	bathroom is clean..... now on to more enjoyable tasks.....
52	53	1	Sentiment140	boom boom pow
53	54	0	Sentiment140	but i'm proud.
54	55	0	Sentiment140	congrats to helio though

Table 2.3.6.2: Tweets after processing targets.

2.3.7 Acronyms

We replace all acronyms with their translation. An acronym is an abbreviation formed from the initial components in a phrase or a word. Usually these components are individual letters (as in NATO or laser) or parts of words or names (as in Benelux). Many acronyms are used in our data set of tweets as you can see in the following bar chart.

At this point, tweets are going to be tokenized by getting rid of the punctuation and using split in order to do the process really fast. We could use nltk.tokenizer but it is definitely much much slower (also much more accurate).

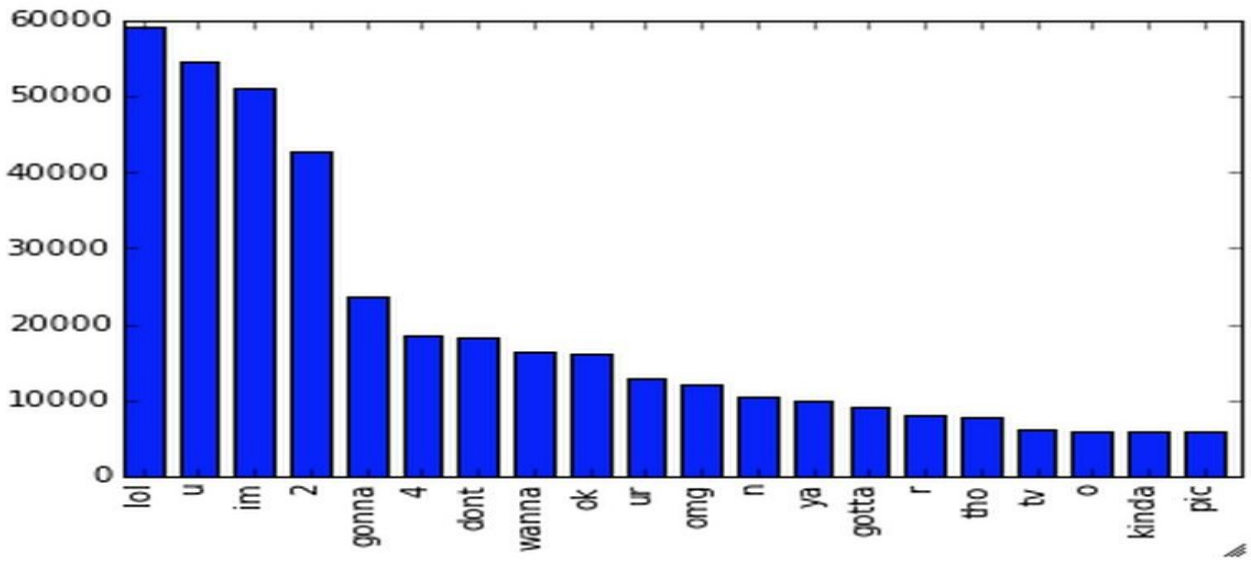


Figure 2.3.7.1: Top 20 of acronyms in the data set of tweets

As you can see, “lol”, “u”, “im”, “2” are really often used by users. The table below shows the top 20 acronyms with their translation and their count.

```
1) lol => laughing out loud : 59000
2) u => you : 54557
3) im => instant message : 51099
4) 2 => too : 42645
5) gonna => going to : 23716
6) 4 => for : 18610
7) dont => don't : 18363
8) wanna => want to : 16357
9) ok => okay : 16104
10) ur => your : 12960
11) omg => oh my god : 12178
12) n => and : 10415
13) ya => yeah : 9948
14) gotta => got to : 9243
15) r => are : 8132
16) tho => though : 7696
17) tv => television : 6246
18) o => oh : 6002
19) kinda => kind of : 5953
20) pic => picture : 5945
```

Table 2.3.7.1: Top 20 of acronyms in the data set of tweets with their translation and count

2.3.8 Negation

We replace all negation words such as “not”, “no”, “never” by the tag **||not||** using the negation dictionary in order to take more or less of sentences like "I don't like it". Here like should not be considered as positive because of the "don't" before. To do so we will replace "don't" by ||not|| and the word like will not be counted as positive. We should say that each time a negation is encountered, the words followed by the negation word contained in the positive and negative word dictionaries will be reversed, positive becomes negative, negative becomes positive, we will do this when we will try to find positive and negative words.

```
['i', "didn't", 'realize', 'it', 'was', 'that', 'deep', 'geez', 'give', 'a', 'girl', 'a', 'warning', 'atleast']
```

Figure 2.3.8.1: A tweet before processing negation words.

```
['i', '||not||', 'realize', 'it', 'was', 'that', 'deep', 'geez', 'give', 'a', 'girl', 'a', 'warning', 'atleast']
```

Figure 2.3.8.2: A tweet after processing negation words.

2.3.9 Sequence of repeated characters

Now, we replace all sequences of repeated characters by two characters (e.g: "helloooo"
= "helloo") to keep the emphasized usage of the word.

	ItemID	Sentiment	SentimentSource	SentimentText
1578604	1578620	1	Sentiment140	[zzzz, no, work, tomorrow, yayyy]
1578605	1578621	1	Sentiment140	[zzzzz, time, tomorrow, will, be, a, busy, day, for, serving, loving, people, love, you, all]
1578606	1578622	0	Sentiment140	[zzzzz, want, to, sleep, but, at, sister's, in, laws's, house]
1578607	1578623	1	Sentiment140	[zzzzzz, finally, night, tweeters]
1578608	1578624	1	Sentiment140	[zzzzzzz, sleep, well, people]
1578609	1578625	0	Sentiment140	[zzzzzzzzzz, wait, no, i, have, homework]
1578610	1578626	0	Sentiment140	[zzzzzzzzzzzz, whatever, what, am, i, doing, up, again]
1578611	1578627	0	Sentiment140	[zzzzzzzzzzzzzzzzzz, i, wish]

Table 2.3.9.1: Tweets before processing sequences of repeated characters.

	ItemID	Sentiment	SentimentSource	SentimentText
1578604	1578620	1	Sentiment140	[zz, no, work, tomorrow, yayy]
1578605	1578621	1	Sentiment140	[zz, time, tomorrow, will, be, a, busy, day, for, serving, loving, people, love, you, all]
1578606	1578622	0	Sentiment140	[zz, want, to, sleep, but, at, sister's, in, laws's, house]
1578607	1578623	1	Sentiment140	[zz, finally, night, tweeters]
1578608	1578624	1	Sentiment140	[zz, sleep, well, people]
1578609	1578625	0	Sentiment140	[zz, wait, no, i, have, homework]
1578610	1578626	0	Sentiment140	[zz, whatever, what, am, i, doing, up, again]
1578611	1578627	0	Sentiment140	[zz, i, wish]

Table 2.3.9.2: Tweets after processing sequences of repeated characters.

2.4 Machine Learning

Once we have applied the different steps of the preprocessing part, we can now focus on the machine learning part. There are three major models used in sentiment analysis to classify a sentence into positive or negative: SVM, Naive Bayes and Language Models (N- Gram). SVM is known to be the model giving the best results but in this project we focus only on probabilistic model that are Naive Bayes and Language Models that have been widely used in this field. Let's first introduce the Naive Bayes model which is well- known for its simplicity and efficiency for text classification.

2.4.1 Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive)independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.

Maximum likelihood training can be done by evaluating a closed form expression (mathematical expression that can be evaluated in a finite number of operations), which takes linear time.

It is based on the application of the Baye's rule given by the following formula:

$$P(C = c|D = d) = \frac{P(D = d|C = c)P(C = c)}{P(D = d)}$$

Formula 2.4.1.1: Baye's rule

where D denotes the document and C the category (label), d and c are instances of D and C

and $P(D = d) = \sum_{c \in C} P(D = d|C = c)P(C = c)$. We can simplify this expression by,

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Formula 2.4.1.2: Baye's rule simlifed

In our case, a tweet d is represented by a vector of K attributes such as $d = (w_1, w_2, ..., w_k)$.

Computing $P(d|c)$ is not trivial and that's why the Naive Bayes introduces the assumption that

all of the feature values w_j are independent given the category label c . That is, for $i =/ j$, w_i and

w_j are conditionally independent given the category label c . So the Baye's rule can be rewritten as,

$$P(c|d) = P(c) \times \frac{\prod_{j=1}^{K} P(w_i|c)}{P(d)}$$

Formula 2.4.1.3: Baye's rule rewritten

Based on this equation, maximum a posterior (MAP) classifier can be constructing by seeking the optimal category which maximizes the posterior $P(c|d)$:

$$c* = arg\max_{c \in C} P(c|d)$$

$$c* = arg\max_{c \in C} \left\{ P(c) \times \frac{\prod_{j=1}^{K} P(w_j|c)}{P(d)} \right\}$$

$$c* = arg\max_{c \in C} \left\{ P(c) \times \prod_{j=1}^{K} P(w_j|c) \right\}$$

Formula 2.4.1.4: Classifier maximizing the posterior probability P(c|d)

Note that $P(d)$ is removed since it is a constant for every category c . There are several variants of Naive Bayes classifiers that are:

- The **Multi- variate Bernoulli Model**: Also called binomial model, useful if our feature vectors are binary (e.g 0s and 1s). An application can be text classification with bag of words model where the 0s 1s are "word does not occur in the document" and "word occurs in the document" respectively.

- The **Multinomial Model**: Typically used for discrete counts. In text classification, we extend the Bernoulli model further by counting the number of times a word w_i appears over the number of words rather than saying 0 or 1 if word occurs or not.

- the **Gaussian Model**: We assume that features follow a normal distribution. Instead of discrete counts, we have continuous features.

For text classification, the most used considered as the best choice is the Multinomial Naive Bayes.

The prior distribution $P(c)$ can be used to incorporate additional assumptions about the relative frequencies of classes. It is computed by:

$$P(c) = \frac{N_i}{N}$$

Formula 2.4.1.5: Prior distribution $P(c)$ where N c .is the total number of training tweets and N_i is the number of training tweets in class

The likelihood $P(w_j|c)$ is usually computed using the formula:

$$P(w_j|c) = \frac{1 + count(w_j, c)}{|V| + N_i}$$

Formula 2.4.1.6: Likelihood $P(w_j|c)$

where $count(w_j, c)$ is the number of times that word w_j occurs within the training tweets of class

c , and $|V| = \sum w_j$ the size of the vocabulary. This estimation uses the simplest smoothing method to solve the **zero- probability problem**, that arises when our model encounters a word seen in the test set but not in the training set, **Laplace** or add- one since we use 1 as constant. We will see that Laplace smoothing method is not really effective compared to other smoothing methods used in language models.

2.4.2 Baseline

In every machine learning task, it is always good to have what we called a baseline. It often a “quick and dirty” implementation of a basic model for doing the first classification and based on its accuracy, try to improve it.

We use the Multinomial Naive Bayes as learning algorithm with the Laplace smoothing representing the classic way of doing text classification. Since we need to extract features from our data set of tweets, we use the **bag of words model** to represent it.

The bag of words model is a simplifying representation of a document where it is represented as a bag of its words without taking consideration of the grammar or word order. In text classification, the count (number of time) of each word appears is a document is used as a feature for training the classifier.

Firstly, we divide the data set into two parts, the training set and the test set. To do this, we first shuffle the data set to get rid of any order applied to the data, then we from the set of positive tweets and the set of negative tweets, we take 3/4 of tweets from each set and merge them

together to make the training set. The rest is used to make the test set. Finally the size of the training set is 1183958 tweets and the test set is 394654 tweets. Notice that they are balanced and follow the same distribution of the initial data set.

Once the training set and the test set are created we actually need a third set of data called the **validation set**. It is really useful because it will be used to **validate our model against unseen data and tune the possible parameters** of the learning algorithm to avoid underfitting and overfitting for example. We need this validation set because our test set should be used only to verify how well the model will **generalize**. If we use the test set rather than the validation set, our model could be **overly optimistic and twist the results**.

To make the validation set, there are two main options:

- Split the training set into two parts (60%, 20%) with a ratio 2:8 where each part contains an equal distribution of example types. We train the classifier with the largest part, and make prediction with the smaller one to validate the model. This technique works well but has the disadvantage of our classifier not getting trained and validated on all examples in the data set (without counting the test set).
- The **K- fold cross- validation**. We split the data set into k parts, hold out one, combine the others and train on them, then validate against the held- out portion. We repeat that process k times (each fold), holding out a different portion each time. Then we average the score measured for each fold to get a more accurate estimation of our model's performance.

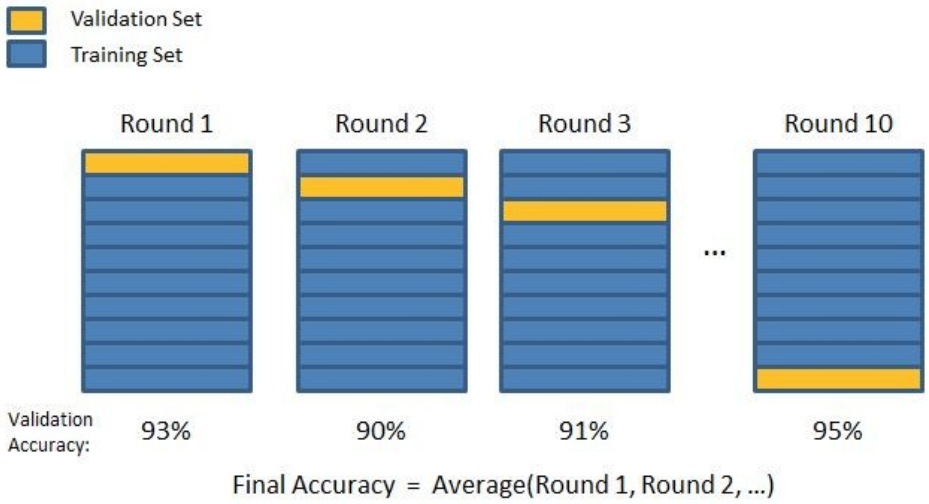


Figure 2.4.2.1: 10- fold cross validation

We split the training data into 10 folds and cross validate on them using scikit- learn as shown in the figure 2.4.2.1 above. The number of K- folds is arbitrary and usually set to 10 it is not a rule. In fact, determine the best K is still an unsolved problem but with lower K: computationally cheaper, less variance, more bias. With large K: computationally expensive, higher variance, lower bias.

We can now train the naive bayes classifier with the training set, validate it using the hold out part of data taken from the training set, the validation set, repeat this 10 times and average the results to get the final accuracy which is about **0.77** as shown in the screen results below,

```
Total tweets classified: 1183958
Score: 0.77653600187
Confusion matrix:
[[465021 126305]
 [136321 456311]]
```

Figure 2.4.2.2: Result of the naive bayes classifier with the score

representing the average of the results of each 10- fold cross- validation, and the overall confusion matrix.

Notice that to evaluate our classifier we two methods, the F1 score and a confusion matrix. The **F1 Score** can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. It a measure of a **classifier's accuracy**. The F1 score is given by the following formula,

$$F1 = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

Formula 2.4.2.1: F1 score

where the precision is the number of true positives (the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class,

$$Precision = \frac{TP}{TP + FP}$$

Formula 2.4.2.1: Precision

and the recall is the number of true positives divided by the total number of elements that actually belong to the positive class,

$$Recall = \frac{TP}{TP + FN}$$

Formula 2.4.2.1: Recall

A precision score of 1.0 means that every result retrieved was relevant (but says nothing about whether all relevant elements were retrieved) whereas a recall score of 1.0 means that all relevant documents were retrieved (but says nothing about how many irrelevant documents were also retrieved).

There is a **trade- off between precision and recall** where increasing one decrease the other and we usually use measures that combine precision and recall such as F- measure or MCC.

A **confusion matrix** helps to visualize how the model did during the classification and evaluate its accuracy. In our case we get about 156715 false positive tweets and 139132 false negative tweets. It is "about" because these numbers can vary depending on how we shuffle our data for example.

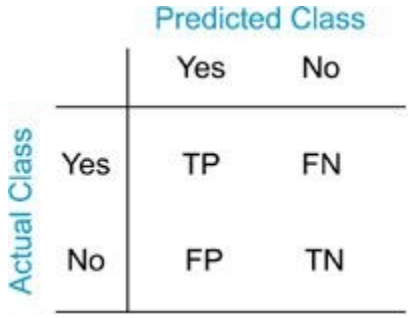


Figure 2.4.2.3: Example of confusion matrix

Notice that we still didn't use our test set, since we are going to tune our classifier for improving its results.

The confusion matrix of the naive bayes classifier can be expressed using a color map where dark colors represent high values and light colors represent lower values as shown in the corresponding color map of the naive bayes classifier below,

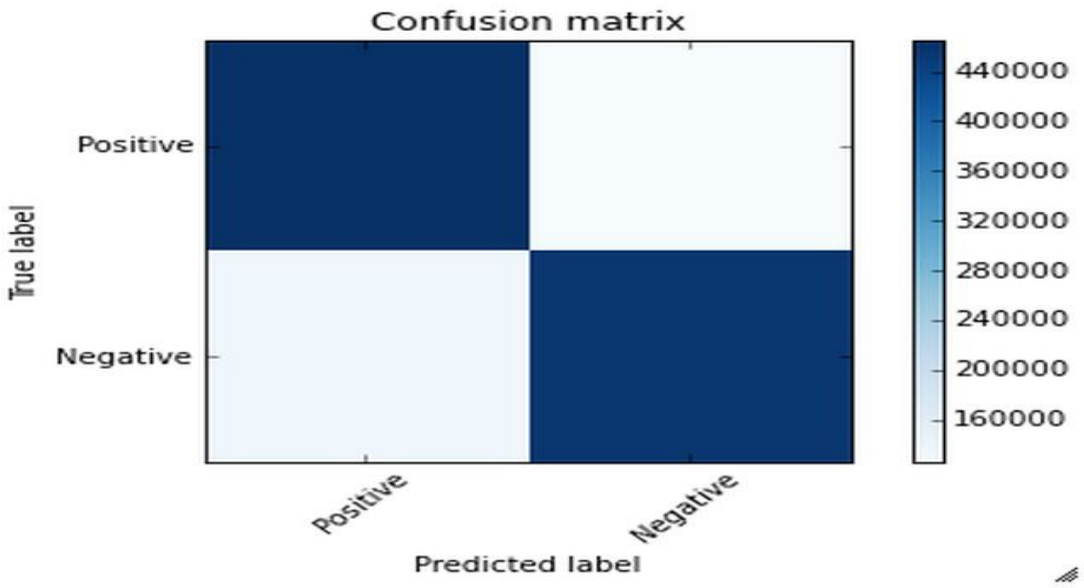


Figure 2.4.2.4: Colormap of the confusion matrix related to the naive bayes classifier used.

Hopefully we can distinguish that the number of true positive and true negative classified tweets is higher than the number of false and positive and negative tweets. However from this result we try to improve the accuracy of the classifier by experimenting different techniques and we repeat the same process using the k- fold cross validation to evaluate its averaged accuracy.

2.4.3 Improvements

From the baseline, the goal is to improve the accuracy of the classifier, which is 0.77, in order to determine better which tweet is positive or negative. There are several ways of doing this and we present only few possible improvements (or not).

First we could try to removed what we called, stop words. Stop words

usually refer to the most common words in the English language (in our case) such as: "the", "of", “to” and so on.

They do not indicate any valuable information about the sentiment of a sentence and it can be necessary to remove them from the tweets in order to keep only words for which we are interested. To do this we use the list of 635 stopwords that we found. In the table below, you can see the most frequent words in the data set with their counts,

```
[('||target||', 780664),
 ('i', 778070),
 ('to', 614954),
 ('the', 538566),
 ('a', 383910),
 ('you', 341545),
 ('my', 336980),
 ('and', 316853),
 ('is', 236393),
 ('for', 236018),
 ('it', 235435),
 ('in', 217350),
 ('of', 192621),
 ('on', 169466),
 ('me', 163900),
 ('so', 158457),
 ('have', 150041),
 ('that', 146260),
 ('out', 143567),
 ('but', 132969)]
```

Table 2.4.3.1: Most frequent words in the data set with their corresponding count.

We can derive from the table, some interesting statistics like the number of times the tags used in the pre- processing step appear,

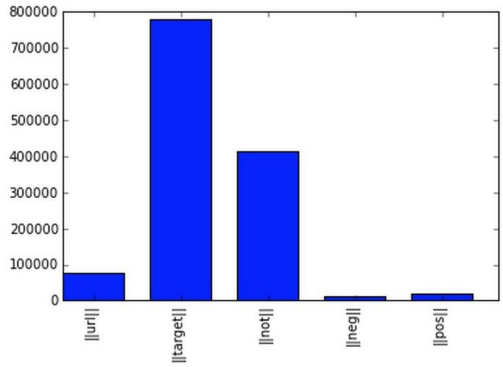


Figure 2.4.3.1: Tags in the data set with their corresponding count.

Recall that ||url|| corresponds to the URLs, ||target|| the twitter usernames with the symbol "@" before, ||not|| replaces the negation words, ||pos|| and ||neg|| replace the positive and negative smiley respectively. After removing the stop words we get the results below,

```
Total tweets classified: 1183958
Score: 0.758623708326
Confusion matrix:
[[437311 154015]
 [136343 456289]]
```

Figure 2.4.2.2: Result of the naive bayes classifier with stopwords removed.

Compared to the previous result, we lose 0.02 in accuracy and the number of false positive goes from 126305 to 154015. We conclude that stop words seem to be useful for our classification task and remove them do not represent an improvement.

We could also try to stem the words in the data set. **Stemming** is the process by which endings are removed from words in order to remove

things like tense or plurality. The stem form of a word could not exist in a dictionary (different from Lemmatization). This technique allows to unify words and reduce the dimensionality of the dataset. It's not appropriate for all cases but can make it easier to connect together tenses to see if you're covering the same subject matter. It is faster than **Lemmatization** (remove inflectional endings only and return the base or dictionary form of a word, which is known as the lemma). Using the library NLTK which is a library in Python specialized in natural language processing, we get the following results after stemming the words in the data set,

```
Total tweets classified: 1183958
Score: 0.773106857186
Confusion matrix:
[[462537 128789]
 [138039 454593]]
```

Figure 2.4.2.2: Result of the naive bayes classifier after stemming.

We actually lose 0.002 in accuracy score compared to the results of the baseline. We conclude that stemming words does not improve the classifier's accuracy and actually do not make any sensible changes.

2.4.4 Language Models

Let's introduce language models to see if we can have better results than those for our baseline. Language models are models assigning **probabilities to sequence of words**.

Initially, they are extensively used in speech recognition and spelling correction but it turns out that they give good results in text classification.

The quality of a language model can be measured by the empirical perplexity (or entropy) using:

$$Perplexity = T\sqrt{\frac{1}{P(w_1,\ldots,w_T)}}$$
$$Entropy = log_2 Perplexity$$

Formula 2.4.4.1: Perplexity and Entropy to evaluate language models.

The goal is to minimize the perplexity which is the same as maximizing probability.

An **N- Gram model** is a type of probabilistic language model for predicting the next item in such a sequence in the form of (n - 1) order Markov Model. The Markov assumption is the probability of a word depends only on the probability of a limited history (previous words).

$$P(w_i|w_1,\ldots,w_{i-1}) = P(w_i|w_{i-n+1},\ldots,w_{i-1})$$

Formula 2.4.4.2: General form of N- grams.

A straightforward maximum likelihood estimate of n- gram probabilities from a corpus is given by the observed frequency,

$$P(w_i|w_{i-n+1},\ldots,w_{i-1}) = \frac{count(w_{i-n+1},\ldots,w_i)}{count(w_{i-n+1},\ldots,w_{i-1})}$$

Formula 2.4.4.2: MLE of N- grams.

There are several kind of n- grams but the most common are the unigram, bigram and trigram. The **unigram model** make the assumption that every word is independent and so we compute the probability of a sequence using the following formula,

$$P(w_1, w_2, \ldots, w_n) = \prod_i P(w_i)$$

Formula 2.4.4.3: Unigram.

In the case of the **bigram model** we make the assumption that **a word is dependent of its previous word**,

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx P(w_i|w_{i-1})$$

Formula 2.4.4.3: Bigram.

To estimate the n- gram probabilities, we need to compute the **Maximum Likelihood Estimates**.

For Unigram:

$$P(w_i) = \frac{C(w_i)}{N}$$

Formula 2.4.4.4: MLE for unigram.

For Bigram:

$$P(w_i, w_j) = \frac{count(w_i, w_j)}{N}$$

$$P(w_j|w_i) = \frac{P(w_i, w_j)}{P(w_i)} = \frac{count(w_i, w_j)}{\sum_w count(w_i, w)} = \frac{count(w_i, w_j)}{count(w_i)}$$

Formula 2.4.4.5: MLE for bigram.

Where N is the number of words, C means count, w_i are words. and

There are two main practical issues:

- We compute everything in log space (log probabilities) to avoid underflow (multiplying so many probabilities can lead to too small number) and because adding is faster than multiplying ($p_1 \times p_2 \times p_3$) = ($log_{p}1 + log_{p}2 + log_{p}$)
- We use smoothing techniques such as Laplace, Witten Bell Discounting, Good- Turing Discounting to deal with unseen words in the training occurring in the test set.

An N- gram language model can be applied to text classification like Naive Bayes model does. A tweet is categorized according to,

$$c* = arg\max_{c \in C} P(c|d)$$

Formula 2.4.4.6:Objective function of n- gram.

and using Baye's rule, this can be rewritten as,

$$c* = arg\max_{c \in C} \{P(c)P(d|c)\}$$

$$c* = arg\max_{c \in C} \left\{ P(c) \times \prod_{i=1}^{T} P(w_i|w_{i-n+1}, \ldots, w_{i-1}, c) \right\}$$

$$c* = arg\max_{c \in C} \left\{ P(c) \times \prod_{i=1}^{T} P_c(w_i|w_{i-n+1}, \ldots, w_{i-1}) \right\}$$

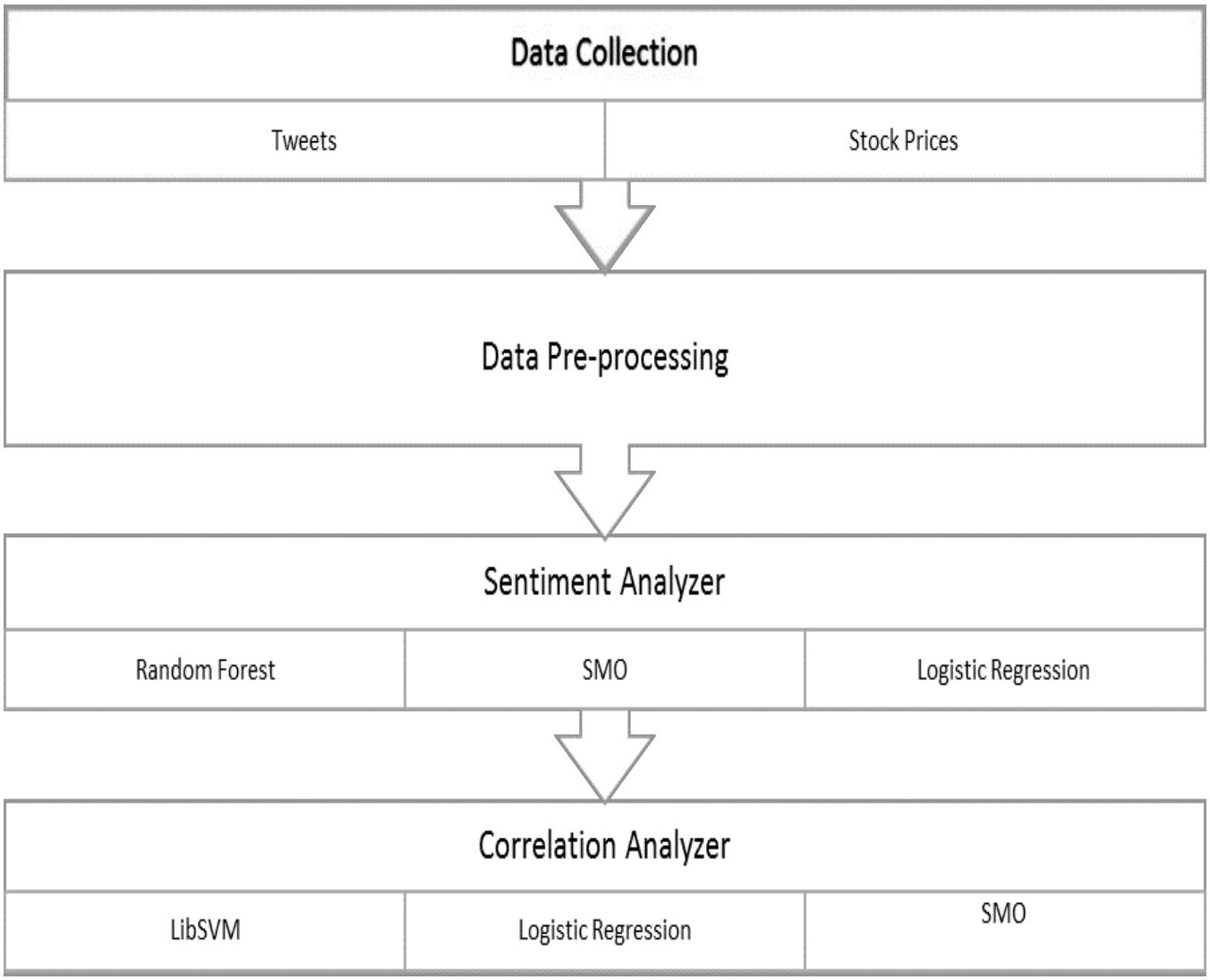


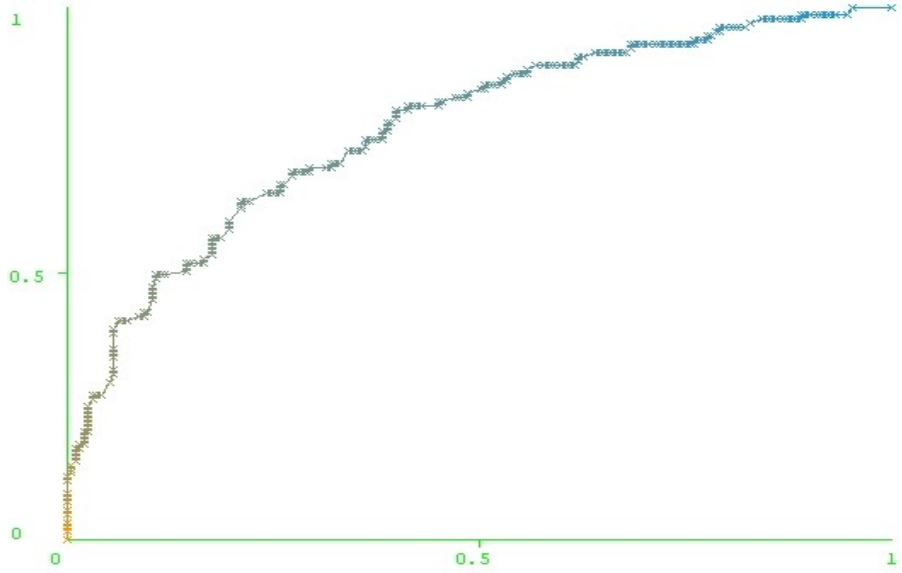
Fig. 1: Flow Chart of the proposed analysis

TABLE I: Sample tweets sentiment labeling by the model

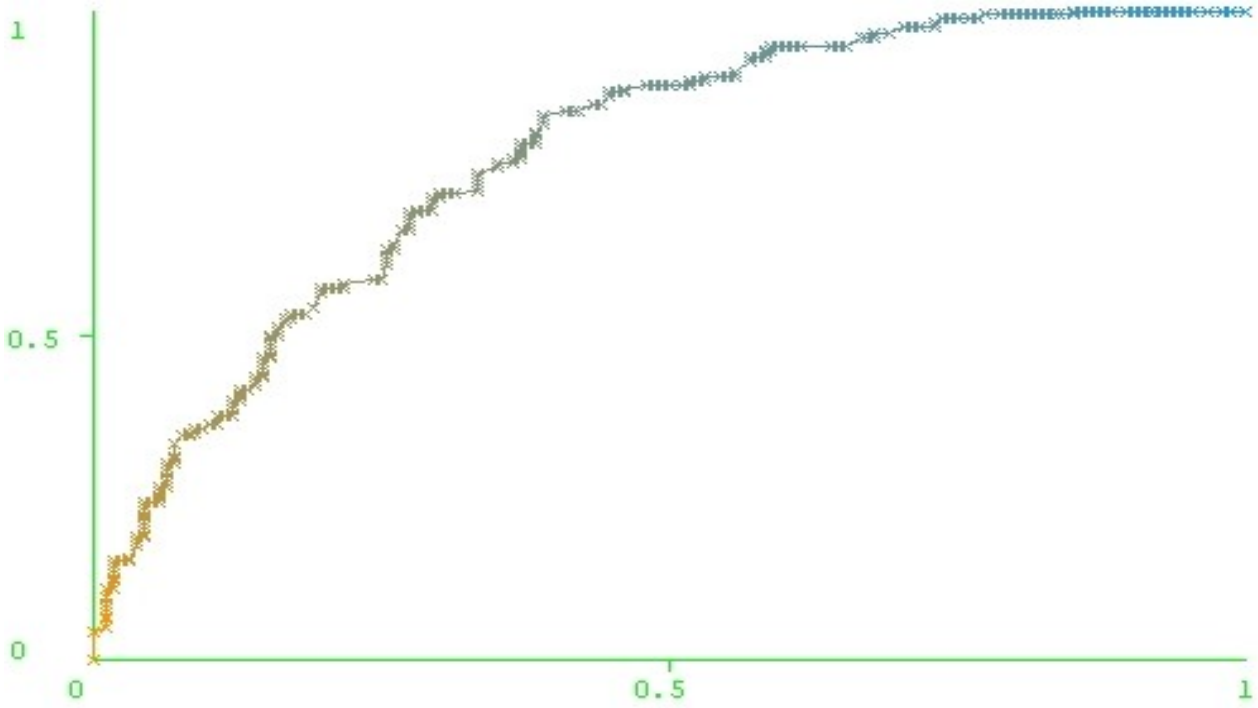
I’m really excited that today is my first day at @Microsoft as a Technical Evangelist	1
Ultrabooks in the mainstream (*AAPLiPadP roincontention*)/ MSFT	1
About the Surface team, it’s one of the most secret team at @Microsoft , except VPs, no one knows what they do, so don’t expect leaks ;)	1
We broke all your devices, we force you to update whenever we want, and we spy on you all the time. Happy Anniversary!” - @microsoft	0
My grandfather, a depression era company man, hated that $ MSFT didn’t pay dividends (90s). I never could explain ”new” idea of growth/cash.	2
Thanks @Microsoft , delivered my Surface Pro 3 power plug to a newsagents 20 miles from my home with no notification.	0
10 swift lessons. My post on @SwiftKey @Microsoft @georgewhitehead @OctopusVentures @IndexVentures @Accel	0
@Microsoft Acquires MinecraftEdu, Tailored for Schools http:// nyti.ms/1U9a1oV (via: @nytimesbusiness	0
My surface’s touchscreen stopped working... I have homework... @Microsoft expand your customer service hours please	0
MSFT trailing revenues are declining since 2015, net income is up thanks to higher margins	2

Machine Learning Algorithm	Word2vec				N-gram			
	Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
Random Forest	70.18%	0.711	0.702	0.690	70.49%	0.719	0.705	0.694
Logistic Regression	62.42%	0.621	0.624	0.621	57.14%	0.580	0.571	0.574
SMO	62.42%	0.617	0.624	0.618	65.84%	0.658	0.658	0.657

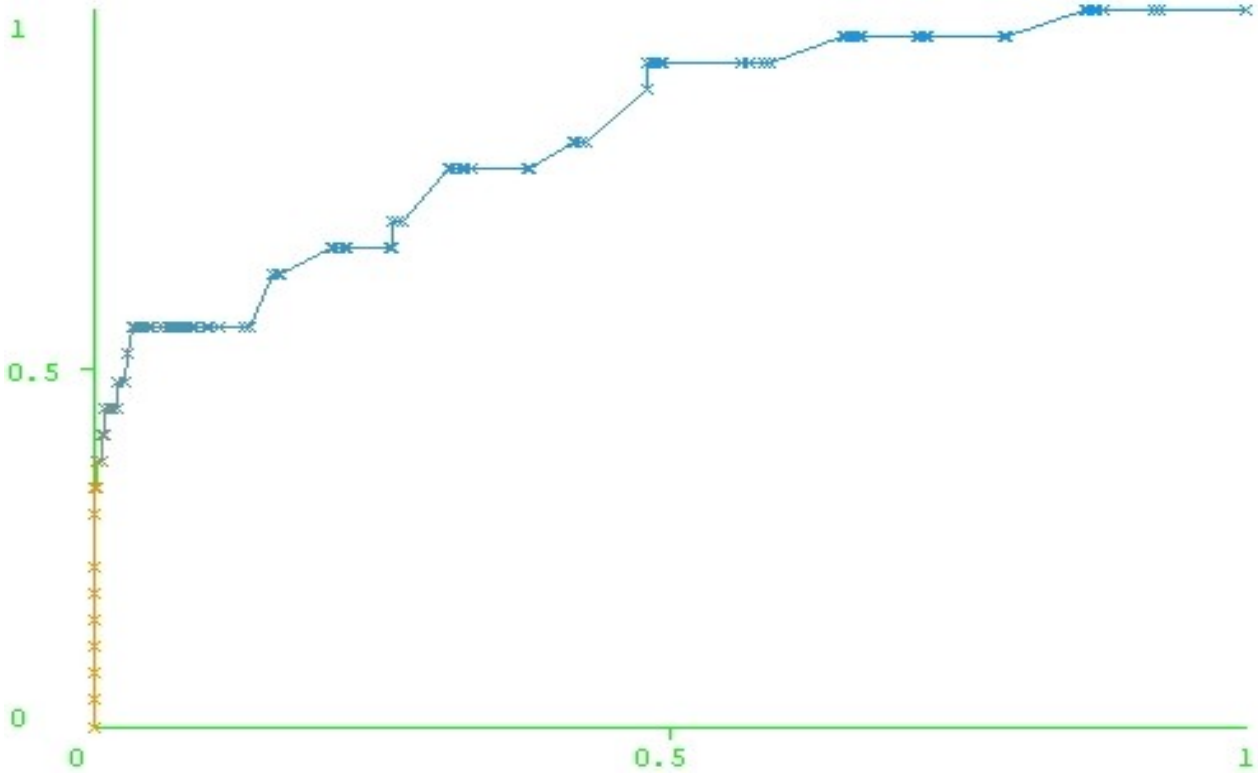
(a) Figure-1 ROC for Positive sentiment classification Area=0.772



(b) Figure-1 ROC for Positive sentiment classification Area=0.773



(c) Figure-3 ROC for Negative sentiment classification Area=0.828



An important note is that n- gram classifiers are in fact a generalization of Naive Bayes. A unigram classifier with Laplace smoothing corresponds exactly to the traditional naive Bayes classifier.

Since we use bag of words model, meaning we translate this sentence: "I don't like chocolate" into "I", "don't", "like", "chocolate", we could try to use bigram model to take care of negation with "don't like" for this example. Using bigrams as feature in the classifier we get the following results,

```
Total tweets classified: 1183958
Score: 0.784149223247
Confusion matrix:
[[480120 111206]
 [138700 453932]]
```

Formula 2.4.4.8: Results of the naive bayes classifier with bigram features.

Using only bigram features we have slightly improved our accuracy score about 0.01. Based on that we can think of adding unigram and bigram could increase the accuracy score more.

```
Total tweets classified: 1183958
Score: 0.795370054626
Confusion matrix:
[[486521 104805]
 [132142 460490]]
```

Formula 2.4.4.9: Results of the naive bayes classifier with unigram and bigram features.

and indeed, we increased slightly the accuracy score about 0.02 compared to the baseline.

3. Conclusion

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of s corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Chinese.

In this project we tried to show the basic way of classifying tweets into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes classifier, or trying another classifier all together.

- References

Alexander Pak, Patrick Paroubek. 2010, Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

- Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision.
- Jin Bai, Jian- Yun Nie. Using Language Models for Text Classification.
- Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. Sentiment Analysis of Twitter Data.
- Fuchun Peng. 2003, Augmenting Naive Bayes Classifiers with Statistical Language Models

4. Appendix

5.1 Minutes of the first meeting

Date:

08/06/201

5 Time :

13:40

Place : Room 3512

Attending :

Marc Lamberti

Professor David Rossiter

Recorder : Marc Lamberti

5.1.1. Approval of minutes

This is first formal group meeting, so there were no minutes to approve.

5.1.2. Discussion Items

- Information gathering
- Research papers
- Set the objectives of the project

5.1.3. Meeting adjournment and next meeting

- Make the data set
- Define the choice of techniques for pre- processing and machine learning

5.2 Minutes of the second meeting

Date:

22/06/201

5 Time:

13:40

Place :Room 3512

Attending:

Marc Lamberti

Professor David Rossiter

Recorder : Marc Lamberti

5.2.1. Approval of minutes

Minutes approved.

5.2.2. Discussion Items

- Methods used for pre- processing the data
- The set of tweets to use
- Methods used for machine learning

5.2.3. Meeting adjournment and next meeting

- Pre- processing of the data set

5.3 Minutes of the third meeting

Date:

06/07/201

5 Time :

13:40

Place :Room 3512

Attending :

Marc Lamberti

Professor David Rossiter

Recorder :Marc Lamberti

5.3.1. Approval of minutes

Minutes approved.

5.3.2. Discussion Items

- Pre- processing of the data set
- Features extracted
- Features selection
- Features transformation

5.3.3. Meeting adjournment and next meeting

- Make prediction using Machine Learning
- Make the report

5.4 Minutes of the fourth meeting

Date:

21/07/201

5 Time :

13:40

Place :Room 3512

Attending :

Marc Lamberti

Professor David Rossiter

Recorder :Marc Lamberti

5.4.1. Approval of minutes

Minutes approved.

5.4.2. Discussion Items

- Prediction using Naive Bayes and Language Models
- Results
- Improvement
- Report

Chapter-6

LEARNING EXPERIENCE

The study is an excellent opportunity to build professional connections. Unlike networking events, the people us connect with during an internship spend time with in a professional setting and become familiar with are work. Therefore, this connection will give us a strong recommendation in the future.

During the month of study in Varcons Technologies Pvt Ltd for period of 4 weeks my learning experience is:

- Observed and understand the professional environment which has helped to my knowledge and skills
- Came to know importance of planning, organizing, staffing, directing and controlling how they are doing the managerial functions and adopting in the organization.

- Get to know the promotional schemes like advertising, sales promotion, digital promotion, expanding of branches. done in company
- Understood the workflow model how it helps in organization work determined authority and responsibility for staff.
- Manage to know information technology and system have reduced the time of an activity.
- Got knowledge of the selection process of employees in organization.
- Understand how competition are going in the market.

To conclude this study has given me practical exposure in the study of organization. It was a great experience working. The main purpose of the organizational study is to make student acquainted with the practical knowledge the overall functioning of the organization.

The study opportunities provided but “VARCONS TECHNOLOGIES PVT LTD” Bangalore has helped me fulfil not only my course objective but also enabled me to learn about the organization. The expose received during the course of my study at “VARCONS TECHNOLOGIES PVT LTD” Bangalore has enabled me to understand the corporate world even better. The skills will learn during this period will definitely help in my future. all in all, that is good experience interning at “VARCONS TECHNOLOGIES PVT LTD” Bangalore.

REFERENCE

WEBSITE

- Varcons Technologies Pvt Ltd

  https://www.varconstech.com/

OTHERS

- Company brochure book
- Company annual reports