CERTIFICATE

This is certify that the project titled "**Product features based sentiment analysis**" submitted by **Soumya Singh** is approved for 6 months Internship 2019 programme from 10^{th} January, 2019 to 20^{th} June, 2019, at Digital Government Research Centre, Patna. It is a record of bonafide work carried out by them under our guidance.

Head
Scientist - F
Mr. S. K. Shrivastava
DGRC, Patna

Mentor
Dr. Tapan Kant

DECLARATION

I hereby declare that the project report entitled “**Product features based sentiment analysis**” is our original work. This written submission represents our ideas in our own words and wherever others’ ideas or words have been included. I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or falsified any idea/data/fact/source in our submission.

ACKNOWLEDGMENT

I would like to thank the people who have rendered their support, supervision, guidance and sincere help during the course of this project work. I express my heartfelt gratitude towards Mr. Sailesh Kumar Shrivastav for giving us the opportunity to work as Intern at Digital Government Research Centre, Patna. I also want to thank Dr. Tapan Kant, my project mentor, for his continuous support and guidance throughout the course of this project.

I would like to thank all the employees and workers at Digital Government Research Centre, Patna for cooperating with us and for giving us such a positive work environment.

I deeply acknowledge the love, cooperation, and the moral support extended by our friends and family members during this period. Finally, I must thank the almighty for guiding and blessing all of us in this endeavor.

SOUMYA SINGH
Banasthali University,
Jaipur,Rajasthan

ABSTRACT

Nowadays, there are trend of writing opinion or reviews or blogs on social media (such as twitter, Facebook, IMDB) which has created huge amount of textual data. These textual data can be used to know the opinion about an entity or product. These opinions can be used by decision makers for making future strategies. Reading the voluminous textual data is a tedious task. In order to mine the opinion from these blogs or reviews, natural language processing (NLP) are being used. In last two decades, sentiment analysis (a branch of NLP) has gained much popularity among researchers. It includes text analysis which contains expressed opinions (positive, neutral, or negative). The approaches for sentiment analysis are (a) machine learning (b) semantic orientation (c) publicly available library (i.e. SentiWordNet). The sentiment analysis can be performed either on a document, at sentence-level, at feature-level, or at entity-level. The basic principle of sentiment analysis technique is feature extraction. However, many work has been done in mining the sentiment of a product. Our aim is to mine the features and its sentiment of a product which will reveal the opinion about the different features of a particular product. In order to solve the problem of features-opinion, a deep learning approach will be applied on product reviews. A deep learning library called keras would be used to achieve our goal. We would like to build a multi-layered model of convolutional neural network on amazon product review dataset.

ORGANIZATION PROFILE

National Informatics Center (NIC), the Ministry of Electronics and Information Technology (MeitY), launched the “Digital Government Research Centre (DGRC)” in Patna on 2nd March, 2017. DGRC is the first of its kind anywhere in the country. It is established at STPI, Patna (Patliputra Campus). IIT Patna has entered into a MoU with the NIC, which will create a framework for joint publication, patenting, internships, advisory and consultancy, under the umbrella of DGRC. DGRC is a research centre, established with a focus on the interface between technology, government and institutions.

The Digital Government Research Centre (DGRC) focuses on developing tools and technology for digital government initiatives and also helps to introduce ICT (Information and Communication Technology) solutions in improving service delivery to the common man across the nation. The main motive behind the establishment of DGRC is to involve the use of information technology, specifically the Internet, to facilitate the communication between the government and its citizens.

The purpose of DGRC is the utilization of Information Technology and other web-based technologies to improve and enhance on the efficiency and effectiveness of service delivery in the public sector. One of the most attractive outcomes of the collaboration of NIC with IIT Patna is the implementation of analytics on the huge amount of data at NIC which will be mined to discover useful patterns and rules and trends. India is getting transformed by the Digital India program launched in 2015 and IIT Patna & NIC aim to make the DGRC a crowning success.

National Informatics Centre (NIC) — The National Informatics Centre is a part of the Indian Ministry of Electronics and Information Technology. It has its headquarters in New Delhi. It has offices in all 29 state capitals and 7 union-territory headquarters and almost all districts. It is the premier science & technology organisation of the Government of India in Informatics Services and Information & Communication Technology (ICT) applications.

It plays a pivotal role in steering e-governance applications in the governmental departments at national, state and district levels. It enables the improvement of government services and also maintains a transparency in these services. Almost all Indian-government websites are developed and managed by NIC. The NIC assists in implementing information-technology projects, in collaboration with central and state governments, in the areas of communication & information technology. It also offers telecommunication networking services including wireless metropolitan-area networks (MANs) and local-area networks (LANs) with gateways for Internet and Intranet resource sharing.

NIC computer cells are located in almost all the Ministry buildings of the Central Government and apex offices including the Indian Prime Minister's office, the Indian Presidential Palace (Rashtrapati Bhavan) and India's Parliament House (Sansad Bhavan). It also provide support to grass root level administration. NIC provides the network infrastructure and e-governance support to India's central government and state governments, union-territory administrations, administrative divisions and other government bodies.

Software Technology Parks of India (STPI) — Software Technology Parks of India (STPI) is a society established in 1991 by the Indian Ministry of Electronics and Information Technology with the objective of encouraging, promoting and boosting the export of software from India. STPI has played a seminal role in India having earned a reputation as an information technology superpower.

STPI is an autonomous society that has been set up with distinct focus to boost up software export from the country. It maintains internal engineering resources to provide consulting, training and implementation services. Services cover network design, system integration, installation, operations and maintenance of application networks and facilities in varied areas. It is an export oriented scheme for the development and export of computer software, including export of professional services.

The state with the largest software export contribution has been Karnataka followed by Maharashtra, Tamil Nadu, Kerala and Telangana. STPI has a presence in many major cities of India. STPI centers also provide a variety of services including high-speed data communication, incubation facilities, consultancy, network monitoring, data centers and data hosting. STPI provides physical hosting for the National Internet Exchange of India.

The Digital India initiative — E-governance initiatives in India took a broader dimension in the mid-1990s for wider sector applications with emphasis on citizencentric services. The ICT initiatives of the Government included some major projects such as railway computerization, land record computerization, etc. which focused mainly on the development of information systems. Later on, many states started ambitious individual e-governance projects aimed at providing electronic services to citizens. Though these e-governance projects were citizen-centric, they could make less than the desired impact due to their limited features.

The Digital India campaign was launched by the Government of India in the year 2015 to ensure that government services are made available to citizens electronically by improved online infrastructure and by increasing internet connectivity, and by making the country digitally empowered in the field of technology. It is a flagship program of the Govt. of India with a vision to transform India into a digitally empowered society and economy. Digital India consists of three core components – (a) development of secure and stable digital infrastructure, (b) delivering government services digitally, and (c) universal digital literacy.

The vision of Digital India program is to transform India into a digitally empowered society and knowledge economy. The initiative includes plans to connect rural areas with high-speed internet networks. The vision of Digital India program includes the growth in areas of electronic services, products, manufacturing and job opportunities etc., and it is centered on three key areas – digital infrastructure as a utility to every citizen, governance & services on demand and digital empowerment of citizens. The Digital India campaign undertaken by the government is bringing about a massive transformation in the field of information technology.

Table Of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The emergence in the last decade of social media platforms such as Twitter, Facebook, movie review site(like IMDB),shopping website(like Amazon) enabled people to engage in social activities to express their opinions, thoughts, and emotions on a variety of topics. On such platforms, large amounts of data are produced. This represents an opportunity for companies to analyze their social influence and people opinions towards their products. Consequently, a computational framework is desirable to perform opinion mining and sentiment analysis which can adapt to the activity domain of the user [Gräbner et al., 2012].

Sentiment analysis is the task of identifying whether the opinion expressed in a text is positive or negative in general, or about a given topic. Sometimes, the task of identifying the exact sentiment is not so clear even for humans. "**SENTIMENT ANALYSIS IS ALSO KNOWN AS OPINION MINING**".

Opinion mining techniques can be applied to wide range of data. It can track the popular viewpoint or attitude of general public toward particular thing, the person or an event.There are three general level for opinion mining tasks : Document Level,Sentence Level and Phrase Level [Zhang and Zheng, 2016].

1.2 Aims and Objective

The project aims to produce real time sentiment analysis associated with a range of brands, products and topics. The project's scope is not only to have static sentiment analysis for past data, but also sentiment classification and reporting in real time [You, 2016]. As such, the system should automatically collect and analyse data from Twitter, the primary data source for this project. For example, the sentence "Brand A is awesome" has positive sentiment for Brand A. More sophisticated structures can be built, for example, the sentence "Brand A is okay but Brand B is great" has neutral sentiment for Brand A, and positive sentiment for Brand B. By the end of the project the goal is to produce up to the minute sentiment values for brands and topics. As such, a system which determines the polarity of tweets (Twitter messages) by using machine [Choi et al., 2009].

1.3 Structure

The rest of the report contains six chapters. In the next chapter, background knowledge is discussed, along with the techniques and algorithms used in development.The third chapter - Design, presents a high level view of the system produced.It also covers the design patterns applied, the requirements gathering process, both functional and non-functional, and what methodologies have been used. Then, the Implementation Chapter demonstrates a low level view of the system, covering the implementation challenges, and what heuristics have been proposed to encounter them. In addition, the system is assessed in the Testing chapter, followed by the Evaluation and Results chapter, where the system's performance is compared to other similar tools available,and achievements are presented. Lastly, the conclusion is a reflective chapter, where the completion of the objectives set at the start of the project is assessed according to the timing specifications. In addition, the knowledge gained while working on the project is demonstrated, and future.

Chapter 2

Review of Literature

2.1 Motivation

2.2 Text Mining

Text mining refers to the analysis of data contained in natural language text,(e.g. messages retrieved from twitter). It can be defined as the practice of extracting meaningful knowledge from unstructured text sources.In Digital Marketing,text mining is relevant to the analysis of the customer relationship management.This way a company can improve their predictive analytics models for customer turnover (keep track of customer opinions) [Das et al., 2018].

2.3 Natural Language Processing (NLP)

For the purpose of explaining further concepts, Twitter will be used as a running example. The data retrieved from twitter presents a certain amount of structuring, in the sense that the maximum length of a tweet is 140 characters long. The advantage of the length limit is reflected in the complexity of the analysis for an individual piece of text. However, this project aims to analyse data in a continuous manner, where a large amount of data (e.g: 200 tweets per minute) will be analysed. Furthermore, there is no certainty that all the tweets will follow a formal structure, neither that they will be grammatically correct. It is also expected that abbreviations and short forms of words, as well as slang will be encountered in the text analysed. Moreover, sentences describing the same or similar ideas may have very different syntax and employ very different vocabularies.

Chapter 3

Methodology

The system was developed for the purpose of gathering the opinion or review for sentiment analysis regarding particular product, event or a person, which is being generated on daily basis by different online sources such as social media (Twitter, Facebook), movie review site (IMDB), shopping website(Amazon), etc.

This system had to perform two functions. First, to perform real time interface. Second purpose is to run continuously in pre-established interval of time in order to perform in depth sentiment analysis for evaluating brands, product or an event. This is reflected in separation in two module of data gathering component.

The system adds series of functionalities to the processing chain. The following components build up the final version of the system:

- Data Gathering
- Data Filtering
- Sentiment Analysis

- Association Rules

3.1 Data Gathering

Data gathering or data collection is basically the collection of details or information required according to the area of interest which can be further processed for extracting or obtaining meaningful information. Main source of data gathering is online site.

According to the researcher data gathering is performing module which can be further processed.

Data gathering can have two methods either quantitative or qualitative methods that rely on random response categories. They produce result that are easy to summarize, compare and generalize.

Qualitative result is related with hypothesis of the research oriented programs which is randomly assigned different people and employ probability sampling to the participants of the task and the existing treatment. Goal of conducting quantitative research study is to determine the relationship between one thing and another thing. Quantitative research deals in numbers, logic and an objective stance [Lawless et al., 2010].

Quantitative research focuses on numeric and unchanging data and detailed, convergent reasoning rather than divergent reasoning i.e., generation of about a variety of ideas about a research problem in a spontaneous, free-flowing manner. Its main characteristics are as follows:

- The data is usually gathered using structured research instruments.

- The results are based on larger sample sizes that are representatives of the population.
- The research study can usually be replicated or repeated, given its high reliability.
- Researcher has a clearly defined research question to which objective answers are sought.
- All aspects of the study are carefully designed before data is collected.
- Data are in the form of numbers and statistics, often arranged in tables, charts, figures, or other non-textual forms.
- Project can be used to generalize concepts more widely, predict future results, or investigate casual relationships.
- Researcher uses tools, such as questionnaires or computer software to collect numeric data.

3.1.1 Qualitative Methods

The qualitative research method involves the use of qualitative data method, such as interviews, documents and observation, in order to understand and explain a social phenomenon. In information Technology and communication, there has been a general shift in research away from technological to managerial and organizational issues, and thus there is increasing interest in the application of qualitative research methods. Qualitative Research methods originated from social sciences to enable researchers to study social and cultural oriented from social science to enable researchers to study social and cultural oriented phenomena. Today , the use of qualitative method and analysis are extended almost to every research field and area. The method generally includes data sources with observation and respondent observation, interviews and questionnaires, documents and the researcher's impression and perception [Brant, 1988] [Yazan, 2015] [McNabb, 2015]. There are different research method in Qualitative Research:

- Action Research
- Case Study
- Ethnography
- Grounded Theory
- Content Analysis

3.1.2 Quantitative Methods

Quantitative research methods are research methods dealing with numbers and anything that is measurable in a systematic way of investigation of phenomena and their relationships. It is used to answer questions on relationships within measurable variables with an intention to explain, predict and control a phenomena.

An entire quantitative study usually ends with confirmation or dis-confirmation of the hypothesis tested. Researchers using the quantitative method identify one or a few variables that they intend to use in their research work and proceed with data collection related to those variables.

Quantitative method typically begins with data collection based on a hypothesis or theory and it is followed with application of descriptive or inferential statistics. Surveys and observations are some examples that are widely used with statistical association. We will see different types of quantitative research methods in the next section. For example, when a researcher is interested to investigate the effectiveness of expert system for managing e-commerce application in open source environment, the researcher will formulate the research question such as, How effective is the expert system in comparison to case-based reasoning for e-commerce module development? The researcher finds 10 software developers using e-commerce module with expert system in open source environment and 10 software developers using case-based reasoning e-commerce module in proprietary programming language environment. The researcher will administer the results and compute them using statistical approach and then summaries it [McClelland, 1994]. Here, we can say the researcher used the quantitative method for the work mentioned.

3.1.3 Quantitative Data Analysis

Data collected from questionnaires or other instruments in quantitative research methods have to be analysed and interpreted. Generally, statistical procedures are quantitative data approaches. In this section, we will look at these common statistical approaches and emphasis on a conceptual understanding for quantitative data analysis.

Statistical methods of quantitative analysis in data gathering are:

1. **Mean**— Mean is also known as average. A mean is the sum of all scores divided by the number of scores. The mean is used to measure central

tendency or centre of a score distribution generally as shown in figure 3.1. For example, the mean for the following set of integers: 3, 4, 5, 7 and 6 = 5.

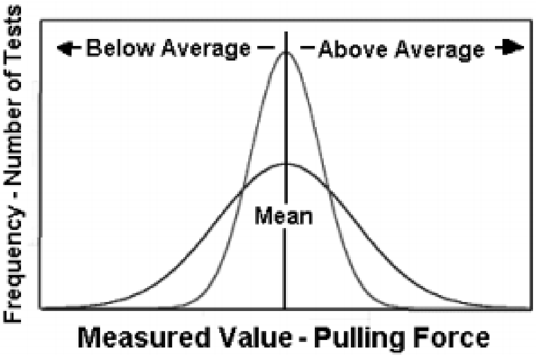

Figure 3.1: Normal Distribution

2. **Standard Deviation**— A standard deviation tells us how close the scores are centered around the mean. By referring to the below figure 3.2, when the scores are bunched together around the mean, the standard deviation is small and the bell curve is steep. When the scores are spread away from the mean, the standard deviation is large and the bell curve is relatively flat.

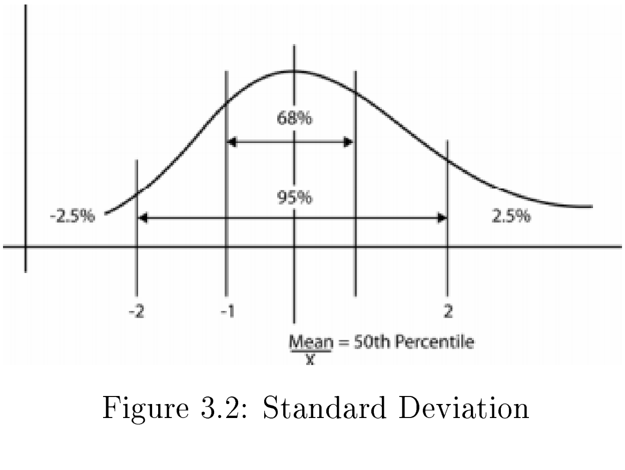

Figure 3.2: Standard Deviation

3.1.4 Data Collection Methods

In research methodology, data collection methods are given great emphasis. Data are categorized as primary data and secondary data. Data collection and research method are inextricably interdependent. A researcher who takes into account a methodology for his/her research work must consider the nature of data that will be collected in the resolution of a problem [Patton, 1990]. we can also say that the data dictate the research method of a particular field. Primary data are collected from primary sources and secondary data gathered from secondary sources. Various methods of data collection are as follows:

1. **PRIMARY SOURCES**
    - Observation
    - Interviewing
    - Questionnaire
2. **SECONDARY SOURCES**
    - Publications
    - Books
    - Research reports
    - Magazines/Newspapers

Data Collection Using Primary Sources

The choice of data collection method depends on the objective and aim of the research.whatever method you use for data collection ,always ensure that you understand clearly the purpose and the relevance of the study.The same goes for your respondent.So,you must clearly state to them so that they know the aim of the study and could give the feedback accordingly in the mode of questionnaire or interviews.Primary sources of data collection are as follows:

Observation is a systematic way of watching and listening to a phenomenon as it takes place. Observation would serve as the best approach if a researcher is interested in behavior rather than perceptions of respondents or when the subjects are so involved in it that they are unable to provide objective information about it.There are two types of observation–participant and non participant. Participants observation is when a researchers participates in the activities of the study group that is being observed in the same manner as its members without their knowledge that they are being observed.Non-participation observation,on the other hand,is when a researcher does not get involved directly in the activities of the research study but remains a passive observer [Yazan, 2015] [McNabb, 2015].

Interview is method to collect information from people is referred to as interview.Another precise definition is that any person to persons interaction between two or more individuals with a specific purpose in mind is called an interview. There are two types of interviews:

- Unstructured Interviews: This type gives complete freedom in terms of content and structure. In ICT, unstructured interviews are often deployed due to the broad nature of the field. Some unstructured interview examples are in-depth interviews, focus group, narratives and oral interviews.

- Structured Interviews: In structured interviews, we can ask a predetermined set of questions using the same wording and order of question as specified in the interview sequence. Interview sequence is a schedule that lists the set of questions,open ended or close-ended which is prepared by the researchers for use of interaction between him/her and the respondents.It is important to highlight here that the interview sequence is a research tool or instrument for collecting data whereas interviewing is a method of data collection.One of the benefits of using structured interviews is that it ensures data comparability.

Questionnaires are one of the most important technique of data collection.It is a list of written question to be answered by respondents of particular study.

Data Collection Using Secondary Sources

Secondary data is the data that has been already collected by and readily available from other sources. When we use statistical method with primary data from another purpose for our purpose we refer to it as secondary data. It means that one purpose's primary data is another purpose's secondary data.So that secondary data is data that is being reused.Such data are more quickly obtainable than the primary data.

These secondary data may be obtained from many sources, including literature, industry surveys, compilations from computerized databases and information systems, and computerized mathematical models of environmental processes.

Published Printed Sources: There are varieties of published printed sources.Their credibility depends on many factors. For example ,on the writer,publishing company and time and date when published .New sources are preferred and old sources should be avoided as new technology and researches bring new facts into light.

Books: Books are available today on any topic that you want to research.The uses of books start before even you have selected the topic. After selction of topics books provide insight on how much work has already been done on the same topic and you can prepare your literature review. Books are secondary sources but most authentic one in secondary sources.

Journals/Periodicals: Journals and periodicals are becoming more important as far as data collection is concerned.The reason is that journals provide up to date information which at times books cannot and secondary, journals can give information on the very specific topic on which you are researching rather talking about more general topics.

Magazines/Newspapers: Magazines are also effective but not very reliable.Newspaper on the other hand is more reliable and in some cases the information can only be obtained from newspapers as in the case of some political studies.

Why do we do data gathering? After discussing so much about data gathering,we must know the purpose of data gathering. Lets discuss now,Data gathering improves our decision making by helping us focus on objective information about what is happening in the process,rather than subjective opinions.For me to collect data uniformly,we will need to develop a Data Collection plan.The element of the plan must be clearly and unambiguously defined-operationally defined.We may want to pause here and review the operational definitions module before we go on.There is a need of operational definitions in order to collect useful data.Lets's say four people are collecting data on the time it takes to perform a certain process step.Unless the exact moment when each action begins and the exact moment.

Data Collection is obtaining useful information.

The issue is not: How do we collect data?

It is: How do we obtain useful data?

Data Collection can involve a multitude of decisions by data collectors. When you prepare your Data Collection plan, you should try to eliminate as many subjective choices as possible by operationally defining the parameters needed to do the job correctly. It may be as simple as establishing separate criteria and a specific way to judge when a step begins and when it ends. Your data collectors will then have a standard operating procedure to use during their Data Collection activities.

3.2 Data Filtering

After the process of gathering data, next step which should be performed is data filtering. It refers to the defining, detecting and correcting errors in raw data. Typically, data filtering will involve taking out information that is useless to a reader or information that can be confusing. Generated reports and query results from database tools often result in large and complex data sets. Redundant or impartial pieces of data can confuse or disorient a user. Filtering data can also make results more efficient. Looking at raw data isn't always informative. If you look only at individual responses or only at overall averages, you'll miss seeing some valuable trends. It's often only once you've cleaned up and organized your data into distinct sections that you start to see patterns emerge [Haykin and Haykin, 2001] [Clifford et al., 2006] [Mather and Tso, 2016]. We indulge certain methods for data filtration:

Data analysis: Learn to use basic and advanced filtering techniques to get a more detailed picture of your data.

Data quality: Find out how to use filters to sort out bad or unwanted responses so that your data is squeaky clean before you start your analysis.

3.2.1 Process That Defines the Data Filtration

Many irrelevant attributes may be present in data to be mined. So they need to be removed. Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore,feature selection techniques needs to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are to avoid over-fitting and improve model performance and to provide faster and more cost-effective models. Advantages of filter techniques are that they easily scale to high dimensional data-sets are computationally simple and fast.

3.3 Sentiment Analysis

The opinions of others have a significant influence in our daily decision-making process. These decisions range from buying a product such as a smart phone on social platform (like Amazon), or watching any movie following its review (i.e. on IMDB), all decisions that affect various aspects of our daily life. Before the Internet, people would seek opinions on products and services from sources such as friends, relatives, or consumer reports. However, in the Internet era, it is much easier to collect diverse opinions from different people around the world. People look to review sites (IMDB), e-commerce sites (e.g., Amazon, eBay), online opinion sites (e.g., TripAdvisor, Rotten Tomatoes, Yelp) and social media (e.g., Facebook, Twitter) to get feedback on how a particular product or service may be perceived in the market. Similarly, organizations use surveys, opinion polls, and social media as a mechanism to obtain feedback on their products and services. sentiment analysis or opinion mining is the computational study of opinions, sentiments, and emotions expressed in text [Abdulla et al., 2014] [He and Zhou, 2011]. The use of sentiment analysis is becoming more widely leveraged because the information it yields can result in the monetization of products and services. For example, by obtaining consumer feedback on a marketing campaign, an organization can measure the campaign's success or learn how to adjust it for greater success. Product feedback is also helpful in building better products, which can have a direct impact on revenue, as well as comparing competitor offerings.

WHY DO WE DO SENTIMENT ANALYSIS?— Sentiment analysis is an evolving field with a variety of use applications. Although sentiment analysis tasks are challenging due to their natural language processing origins, much progress has been made over the last few years due to the high demand for it. Not only do companies want to know how their products and services are perceived by consumers (and compare to competitors), but consumers want to know the opinions of others before making buying decisions. The growing need for product insights – and the technical challenges currently facing the field –will keep sentiment analysis and opinion mining relevant for the foreseeable future. Next-generation opinion mining systems need a deeper bind between complete knowledge bases with reasoning methods inspired by human thought and psychology [Algur et al., 2010] [and and and, 2015]. This will lead to a better understanding of natural language opinions and will more efficiently bridge the gap between unstructured information in the form of human thoughts and structured data that can be analyzed and processed by a machine. The SENTIMENT ANALYSIS is a complex process that has 5 different steps to analyze sentiment data. These steps are:

HOW DO WE DO SENTIMENT ANALYSIS?— Sentiment analysis is a new field of research born in Natural Language Processing (NLP), aiming at detecting subjectivity in text and/or extracting and classifying opinions and sentiments. Sentiment analysis studies people's sentiments, opinions, attitudes, evaluations, appraisals and emotions towards services, products, individuals, organizations, issues, topics events and their attributes.

- **DATA COLLECTION:** The first step of sentiment analysis consists of collecting data from user generated content contained in blogs, forums, social networks. These data are disorganized, expressed in different ways by using different vocabularies, slangs, context of writing etc. Manual analysis is almost impossible. Therefore, text analytics and natural language processing are used to extract and classify.

- **TEXT PREPARATION:** consists in cleaning the extracted data before analysis. Non-textual contents and contents that are irrelevant for the analysis are identified and eliminated.

- SENTIMENT DETECTION the extracted sentences of the reviews and opinions are examined. Sentences with subjective expressions (opinions, beliefs and views) are retained and sentences with objective communication (facts, factual information) are discarded

- **SENTIMENT CLASSIFICATION:** in this step, subjective sentences are classified in positive, negative, good, bad; like, dislike, but classification can be made by using multiple points.
- **PREPARATION OF OUTPUT:** the main objective of sentiment analysis is to convert unstructured text into meaningful information. When the analysis is finished, the text results are displayed on graphs like pie chart, bar chart and line graphs. Also time can be analysed and can be graphically displayed constructing a sentiment time line with the chosen value ((frequency, percentages, and averages) over time.

Associative classification— Associative classification mining is a promising approach in data mining that utilizes the association rule discovery techniques to construct classification systems, also known as associative classifiers. In the last few years, a number of associative classification algorithms have been proposed, i.e. CPAR, CMAR, MCAR, MMAC and others. These algorithms employ several different rule discovery, rule ranking, rule pruning, rule prediction and rule evaluation methods. This paper focuses on surveying and comparing the state-of-the-art associative classification techniques with regards to the above criteria. Finally, future directions in associative classification, such as incremental learning and mining low-quality data sets, are also highlighted in this paper.

3.4 Implementation Environment

3.4.1 Anaconda Distribution for Python

Python is a widely used high-level, general-purpose, interpreted and dynamic programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale.

Anaconda is a free and open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. Package versions are managed by the package management system Anaconda. Anaconda Cloud is where data scientists share their work. One can search and download popular Python and R packages and notebooks to jumpstart the data science work. We can also store our packages, notebooks and environments in Anaconda Cloud and share them with our team. Anaconda is professional data science platform and python IDE's where we can use as console as well as GUI.

Anaconda is Python distribution that brings a lot of useful libraries, which are not included in Python standard library as are numpy, scikit-learn, etc. Unlike Python, in Anaconda one can easily update the libraries of Python completely independent of system libraries or admin privileges. It comes with a huge set of libraries pre-installed with it even if some "libraries are not installed one can download these libraries by writing simple command in Anaconda Prompt.

An anaconda environment is instead a way to have a sand boxed installation of python installed on our computer. Having many environments lets us easily switch between different versions or packages available for import. There is also an anaconda desktop app that is a GUI for managing these environments and it lets us launch some other python work environments or IDE like Jupyter notebook and Spyder. The IDE that we have used for our project work is Spyder (Scientific Python Development Environment). Spyder is a powerful interactive development environment for the Python language with advanced editing, interactive testing, debugging and introspection features and a numerical computing environment thanks to the support of IPython (enhanced interactive Python interpreter) and popular Python libraries such as NumPy (linear algebra),matplotlib (interactive 2D/3D plotting). Spyder may also be used as a library providing powerful console-related widgets for your PyQt-based applications – for example, it may be used to integrate a debugging console directly in the layout of your graphical user interface.

3.4.2 Steps to Install Anaconda in Windows 10

The following steps are to be followed in order to install Anaconda on a Windows 10 platform.

1. Download the installer for Windows from the official Anaconda Website

2. Go to packages and select the version of Python needed to be installed.

    - Miniconda installer for Windows

    - Anaconda installer for Windows

3. Double-click the .exe file.

4. Follow the instructions on the screen. If unsure about any setting, accept the defaults. When installation is finished, from the Start menu, open the Anaconda Prompt.

5. Test the installation by running Anaconda Navigator.

6. Various IDEs are pre-installed with Anaconda like Syder, JupyterLab, Jupiter Notebook and QtConsole.

3.4.3 Python Libraries

In programming, a library is a collection of precompiled routines that a program can use. The routines, sometimes called modules, are stored in object format. Libraries are particularly useful for storing frequently used routines. Python library is a collection of functions and methods that allows us to perform lots of actions without writing our own code. The library contains built-in modules that provide access to system functionalities such as file I/O that would otherwise be inaccessible to Python programmers, as well as modules written in Python that provide standardized solutions for many problems that occur in everyday programming.

- Pandas
- Numpy
- Matplotlib

Pandas

PANDAS is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language.

The two primary data structures of pandas are series (1-dimensional) and dataframe (2-dimensional). Pandas is well suited for many different kinds of data like tabular data (with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet), ordered and unordered time series data, arbitrary matrix data (homogeneously typed or heterogeneous) with row and column label and any other form of observational / statistical data sets.

Numpy

Matplotlib

MATPLOTLIB is a Python library used for plotting. It produces publication quality figures. Matplotlib tries to make things easy. Generating variety of plots, histograms, bar charts, scatterplots etc. is possible. Matplotlib provides an object-oriented API for embedding plots into applications using GUI toolkits like Tkinter, Qt etc.

The Matplotlib code is conceptually divided into three parts:

1. The **pylab** interface is the set of functions provided by matplotlib.pylab

which allows the user to create plots with code quite similar to MATLAB figure generating code.

2. The **Matplotlib frontend** or **Matplotlib** API is the set of classes that do the heavy lifting, creating and managing figures, text, lines, plots and so on. This is an abstract interface that knows nothing about output.

3. The **backends** are device-dependent drawing devices, aka renderers, that transform the front-end representation to hardcopy or a display device.

Although it has its origins in emulating the MATLAB graphics commands, it is independent of MATLAB, and can be used in a Pythonic, object oriented way.

Matplotlib has a large community. It is mostly used in Python for drawing graphs and enabling exploration of data. It comprises of tons of plot types. The different types of plots it supports are line plots, scatter plots, bar plots, histogram and multiple plots. It is well integrated into IPython (Interactive Python). It is the de-facto standard for “command line” plotting from IPython. Matplotlib enables day-to-day data exploration.

Some features of Matplotlib are as follows:

- Easy handling of missing data (represented as NaN)
- Size mutability: columns can be inserted and deleted from a dataFrame and higher dimensional objects
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let the series, dataframe, etc. automatically align the data during computations

- Powerful, flexible group by functionality to perform split-applycombine operations on datasets, for both aggregating and transforming data
- Make it easy to convert ragged, differently-indexed data in other data structures into DataFrame objects
- Intelligent label-based slicing, fancy indexing, and subsetting of large data sets
- Intuitive merging and joining data sets
- Flexible reshaping and pivoting of data sets
- Hierarchical labeling of axes (possible to have multiple labels per tick)

Chapter 4

Results and Discussion

4.1 Motivation

The emergence in the last decade of social media platforms such as Twitter, Facebook, movie review site(like IMDB),shopping website(like Amazon) enabled people to engage in social activities to express their opinions, thoughts, and emotions on a variety of topics. On such platforms, large amounts of data are produced. This represents an opportunity for companies to analyze their social influence and people opinions towards their products. Consequently, a computational framework is desirable to perform opinion mining and sentiment analysis which can adapt to the activity domain of the user.

HERE ARISE A QUESTION WHILE READING THE ABOVE PARAGRAPH, WHICH HAS A STRONG IMPACT ON THE SOCIAL PLATFORMS i.e. "WHAT IS SENTIMENT ANALYSIS?"

Sentiment analysis is the task of identifying whether the opinion expressed in a text is positive or negative in general, or about a given topic. Sometimes, the task of identifying the exact sentiment is not so clear even for humans. SENTIMENT ANALYSIS IS ALSO KNOWN AS OPINION MINING. Opinion mining techniques can be applied to wide range of data. It can track the popular viewpoint or attitude of general public toward particular thing, the person or an event.There are three general level for opinion mining tasks : Document Level,Sentence Level and Phrase Level.

4.2 Aims and Objective

The project aims to produce real time sentiment analysis associated with a range of brands, products and topics. The project's scope is not only to have static sentiment analysis for past data, but also sentiment classification and reporting in real time. As such, the system should automatically collect and analyse data from Twitter, the primary data source for this project. For example, the sentence "Brand A is awesome" has positive sentiment for Brand A. More sophisticated structures can be built, for example, the sentence "Brand A is okay but Brand B is great" has neutral sentiment for Brand A, and positive sentiment for Brand B. By the end of the project the goal is to produce up to the minute sentiment values for brands and topics. As such, a system which determines the polarity of tweets (Twitter messages) by using machine.

4.3 Structure

The rest of the report contains six chapters. In the next chapter, background knowledge is discussed, along with the techniques and algorithms used in development.The third chapter - Design, presents a high level view of the system produced.It also covers the design patterns applied, the requirements gathering process, both functional and non-functional, and what methodologies have been used. Then, the Implementation Chapter demonstrates a low level view of the system, covering the implementation challenges, and what heuristics have been proposed to encounter them. In addition, the system is assessed in the Testing chapter, followed by the Evaluation and Results chapter, where the system's performance is compared to other similar tools available,and achievements are presented. Lastly, the conclusion is a reflective chapter, where the completion of the objectives set at the start of the project is assessed according to the timing specifications. In addition, the knowledge gained while working on the project is demonstrated, and future.

Chapter 5

Summary

5.1 Motivation

The emergence in the last decade of social media platforms such as Twitter, Facebook, movie review site(like IMDB),shopping website(like Amazon) enabled people to engage in social activities to express their opinions, thoughts, and emotions on a variety of topics. On such platforms, large amounts of data are produced. This represents an opportunity for companies to analyze their social influence and people opinions towards their products. Consequently, a computational framework is desirable to perform opinion mining and sentiment analysis which can adapt to the activity domain of the user.

HERE ARISE A QUESTION WHILE READING THE ABOVE PARAGRAPH, WHICH HAS A STRONG IMPACT ON THE SOCIAL PLATFORMS i.e. "WHAT IS SENTIMENT ANALYSIS?"

Sentiment analysis is the task of identifying whether the opinion expressed in a text is positive or negative in general, or about a given topic. Sometimes, the task of identifying the exact sentiment is not so clear even for humans. SENTIMENT ANALYSIS IS ALSO KNOWN AS OPINION MINING. Opinion mining techniques can be applied to wide range of data. It can track the popular viewpoint or attitude of general public toward particular thing, the person or an event.There are three general level for opinion mining tasks : Document Level,Sentence Level and Phrase Level.

5.2 Aims and Objective

The project aims to produce real time sentiment analysis associated with a range of brands, products and topics. The project's scope is not only to have static sentiment analysis for past data, but also sentiment classification and reporting in real time. As such, the system should automatically collect and analyse data from Twitter, the primary data source for this project. For example, the sentence "Brand A is awesome" has positive sentiment for Brand A. More sophisticated structures can be built, for example, the sentence "Brand A is okay but Brand B is great" has neutral sentiment for Brand A, and positive sentiment for Brand B. By the end of the project the goal is to produce up to the minute sentiment values for brands and topics. As such, a system which determines the polarity of tweets (Twitter messages) by using machine.

5.3 Structure

The rest of the report contains six chapters. In the next chapter, background knowledge is discussed, along with the techniques and algorithms used in development.The third chapter - Design, presents a high level view of the system produced.It also covers the design patterns applied, the requirements gathering process, both functional and non-functional, and what methodologies have been used. Then, the Implementation Chapter demonstrates a low level view of the system, covering the implementation challenges, and what heuristics have been proposed to encounter them. In addition, the system is assessed in the Testing chapter, followed by the Evaluation and Results chapter, where the system's performance is compared to other similar tools available,and achievements are presented. Lastly, the conclusion is a reflective chapter, where the completion of the objectives set at the start of the project is assessed according to the timing specifications. In addition, the knowledge gained while working on the project is demonstrated, and future.

Bibliography

[Abdulla et al., 2014] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M., Al-Kabi, M. N., and Al-rifai, S. (2014). Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, 9(3):55–71.

[Algur et al., 2010] Algur, S. P., Patil, A. P., Hiremath, P. S., and Shivashankar, S. (2010). Conceptual level similarity measure based review spam detection. In *2010 International Conference on Signal and Image Processing*, pages 416–423.

[and and and, 2015] and and and (2015). Fine-grained sentiment analysis of online reviews. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 1406–1411.

[Brant, 1988] Brant, B. (1988). *A gathering of spirit: a collection of North American Indian women.* Women's Press.

[Choi et al., 2009] Choi, Y., Kim, Y., and Myaeng, S.-H. (2009). Domain-specific sentiment analysis using contextual feature generation. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 37–44, New York, NY, USA. ACM.

[Clifford et al., 2006] Clifford, G. D., Azuaje, F., McSharry, P., et al. (2006). *Advanced methods and tools for ECG data analysis.* Artech house Boston.

[Das et al., 2018] Das, S., Behera, R. K., kumar, M., and Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. *Procedia Computer Science*, 132:956 – 964. International Conference on Computational Intelligence and Data Science.

[Gräbner et al., 2012] Gräbner, D., Zanker, M., Fliedl, G., and Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In Fuchs, M., Ricci, F., and Cantoni, L., editors, *Information and Communication Technologies in Tourism 2012*, pages 460–470, Vienna. Springer Vienna.

[Haykin and Haykin, 2001] Haykin, S. S. and Haykin, S. S. (2001). *Kalman filtering and neural networks.* Wiley Online Library.

[He and Zhou, 2011] He, Y. and Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4):606–616.

[Lawless et al., 2010] Lawless, R. M., Robbennolt, J. K., and Ulen, T. (2010). *Empirical methods in law.* Aspen Publishers New York.

[Mather and Tso, 2016] Mather, P. and Tso, B. (2016). *Classification methods for remotely sensed data.* CRC press.

[McClelland, 1994] McClelland, S. B. (1994). Training needs assessment data-gathering methods: Part 1, survey questionnaires. *Journal of European Industrial Training*, 18(1):22–26.

[McNabb, 2015] McNabb, D. E. (2015). *Research methods for political science: Quantitative and qualitative methods.* Routledge.

[Patton, 1990] Patton, M. Q. (1990). *Qualitative evaluation and research methods.* SAGE Publications, inc.

[Yazan, 2015] Yazan, B. (2015). Three approaches to case study methods in education: Yin, merriam, and stake. *The qualitative report*, 20(2):134–152.

[You, 2016] You, Q. (2016). Sentiment and emotion analysis for social multimedia: Methodologies and applications. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 1445–1449, New York, NY, USA. ACM.

[Zhang and Zheng, 2016] Zhang, X. and Zheng, X. (2016). Comparison of text sentiment analysis based on machine learning. In *2016 15th International Symposium on Parallel and Distributed Computing (ISPDC)*, pages 230–233.