International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Evaluation of Sentiment Analysis over Bilingual Cross Domain Platform using Machine Learning Approaches

S. Arun Kumar, M. Sanjanaa Sri, Rishendra Ravi, Dipon Sengupta, Arhant Chatterjee

Abstract: ***Cross-Domain adaptation needs special data to get a shared characteristic with various domain. Notwithstanding, such valuable data may not generally be accessible in genuine cases. In this paper, another issue setting called Cross-Domain Sentiment Analysis in bilingual platform is addressed. It is an extraordinary instance of cross-space nostalgic examination in which diverse areas have some regular commonalities, yet in addition have their very own space explicit highlights. We influence upon normal highlights rather than beneficial data to accomplish viable adjustment. We propose a methodology, which can interface up various spaces utilizing normal highlights and at the same time decrease area divergences.***

Index Terms**: *Bilingual Analysis, Naïve Bayes Classifier, N-gram, Sentiment Analysis.***

I. INTRODUCTION

Sentiment analysis is the procedure of computationally distinguishing and arranging sentiments communicated in a bit of content, particularly so as to decide if the essayist's mentality towards a specific subject, item, and so on is sure, negative, or nonpartisan[1]. It is one of the quickest developing examination territories in software engineering, making it trying to monitor every one of the exercises in the territory. It is also called feeling mining and can be strikingly finished with the assistance of artificial intelligence and we will come to it in some time. Fundamentally, it is the way toward deciding the passionate tone behind a progression of words, used to pick up a comprehension of the frames of mind, conclusions and feelings communicated inside an online notice. Nostalgic Analysis[2] is incredibly useful in online networking observing as it empowers us to get a framework of the more broad prominent estimation behind explicit focuses. Internet based life checking apparatuses like Brandwatch Analytics make that methodology quicker and less requesting than whenever in late memory[1], as a result of progressing watching capacities. The employments of end examination are sweeping and fantastic[8]. The ability to remove bits of learning from social data is a preparation that is in actuality by and large gotten by relationship over the world. To manufacture a profound learning model for supposition examination, we initially need to speak to our sentences in a vector space[1]. We contemplated recurrence based strategies in a past post. They speak to a sentence either by a sack of-words, which is a rundown of the words that show up in the sentence with their frequencies, or by a term frequency—inverse archive recurrence (tf-idf) vector[7] where the word frequencies in our sentences are weighted with their frequencies in the whole corpus. These techniques are extremely valuable for long messages. For instance, we can portray all around unequivocally a paper article or a book by its most regular words. In any case, for short sentences, it's not precise by any stretch of the imagination. Anyway there are distinctive methodologies in which wistful investigation can be performed utilizing AI[7].

The vast majority of the Sentiment analysis is completed by focusing on a specific space to accomplish higher precision. Be that as it may, gathering a preparation information is costly and tedious for each new space since assessment communicated contrastingly in various area. Space speculations[9] still a major test in feeling investigation on the grounds that the words utilized in one area could possibly use in another area. Henceforth the majority of the highlights are concealed to that classifier which is prepared on another space. For playing out a cross-space slant examination we require a system to consolidate the data with respect to relatedness among the highlights.

First thinking about of the sentences, we train a spam/ham model[11] and train the model with the labeled information. For taking care of the component confound issue of cross-area supposition investigation we are making a glossary which contains the words that are semantically comparative. This glossary will at that point used to broaden the highlights that are available in the audit of the objective area. Stretched out highlights are attached to the first highlights of survey and afterward by following the pack of words, classifier will precisely arrange that highlights[15]. In spite of their straightforwardness and observational achievement, it isn't hypothetically clear why these calculations perform so well. Contrasted with single-space supposition order[12], cross-area assessment grouping has as of late gotten consideration with the progression in the field of area adjustment. For instance taking two spaces YouTube and Amazon[17]. The first is utilized for hunting down audit of the thing to purchase and the other to purchase the item after the best possible survey suggestion.

Revised Manuscript Received on 30 March 2019.
***** Correspondence Author
S. Arun Kumar*, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Dipon Sengupta, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Rishendra Ravi, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
M . Sanjanaa Sri, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Arhant Chatterjee, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

II. RELATED WORKS

Sentiment Analysis has been a vast area of research and has gained much popularity among data scientists and machine learning algorithm designers in recent. Thus, contributing research works can be visited along in order to perform the literature survey comprising of the existing system architecture and hence draws us towards our area of research. In a research by Godbole[3], he talked about the Large Scale Sentiment Analysis for News and Blogs. The algorithm followed here are: (i) algorithm construction of sentiment dictionaries. (ii)sentiment index formulation. (iii) Evaluation of significance. The paper aims to predict how sentiment can fluctuate by statistic assemble news source or geographic area. By extension of spatial examination of news elements to slant maps can be utilized for recognizing geological areas of positive or unfriendly suppositions for given substances. Thus, it can be remarked that proposed arrangement can likewise be utilized dissecting how much our slant records anticipate future changes in notoriety or market conduct. Another work by Nikhail Bautin[4] in which he talks with reference of the same blogs and news, to show that: (a) element feeling scores acquired by our strategy are factually essentially associated crosswise over nine dialects of news sources and five dialects of a parallel corpus; (b) the nature of our assessment investigation strategy is to a great extent interpreter autonomous; (c) in the wake of applying certain standardization strategies, our element notion scores can be utilized to perform important culturally diverse examinations. The different types of algorithm used are: (i) Cross-language analysis across news streams. (ii) Cross language analysis across parallel corpora. (iii) Analysis of translator-specific artifacts (iv) Normalizing for cross-cultural language effects. The outcome demonstrated that the technique for computing element slant scores is steady regarding shifting dialects and news sources and the scores of two distinct interpreters were thought about. Getting into more extensive imminent and a bigger space we get Semantic Sentiment Analysis of Twitter.

Notion examination over Twitter information and other comparative microblogs[5] faces a few new difficulties because of the regular short length and sporadic structure of such substance. It includes examining a novel arrangement of highlights got from the semantic applied portrayal of the substances that show up in tweets. The semantic highlights comprise of the semantic ideas (for example "individual", "organization", "city") that speak to the elements (for example "Steve Jobs", "Vodafone", "London") removed from tweets. The Algorithms utilized are: 1. Semantic Replacement: Here, all substances in tweets are supplanted with their relating semantic ideas. This prompts the decrease of the vocabulary measure. 2. Semantic Augmentation: This technique increases the first element space with the semantic ideas as extra highlights for the classifier preparing. The extent of the vocabulary for this situation is broadened by the semantic ideas introduced.3. Semantic Interpolation: A general addition technique, which can add discretionary kind of highlights, for example, semantic ideas, POS successions, slant subjects and so forth. In the paper distributed in the year 2018, A General Domain Specific Feature Transfer Framework for Hybrid Domain Adaptation[6], Heterogeneous area adjustment needs strengthening data to connect up various areas. Nonetheless, such valuable data may not generally be accessible in genuine cases. It is an exceptional instance of heterogeneous space adjustment, in which diverse areas share some basic highlights, yet additionally have their very own space explicit highlights. The interpretations between normal highlights and space explicit highlights are appeared. Cross-utilization of the educated interpretations to exchange the area explicit highlights of one space to another space. DSFT[15] takes the top off technique such that, it benefits area adjustment. Calculations utilized are: (i) A conveyance inconsistency metric MMD. (ii) A straight case DSFT structure. (iii) A non-straight case DSFT structure. (iv)Complexity examination of non-straight and direct DSFT structure. (v)Property and versatile examination of DSFT. This structure, DSFT, profits by both the interpretation term and the disparity term. It learns the interpretations between normal highlights and area explicit highlights and uses the educated interpretations to develop a homogeneous element space, in which cross-space information is exchanged and the area difference is limited.

III. PROBLEM STATEMENT

The present module that is in existence handles the sentimental analysis within the same domain. Now implementing this across different domains remains a challenge. The important part that will be handled while adapting cross-domain will be language and finding the feature extraction with respect to both the domains.

IV. SYSTEM OVERVIEW

In the present system, the modules into existence are Basic Pre-processing, Intermediate Pre-processing, Normalization, Sentiment Analysis[1]. With the new system to adapt Cross Domain Sentiment Analysis, we have used an approach, which helps us to deal with the cross-domain adaptation. The flow diagram is as illustrated in Fig 1. The initial step is Data collection[2]. Data can be collected through various means such as Sysomos and Social-Studios which are front-end data extraction modules and hitting the website with the help of API and extracting the fields according to the requirements for our data analysis[5]. After the data has been extracted we store it in the database which has easy accessibility such as, MongoDB. Data extracted is in a raw form and has to be processed according to our analysis needs.

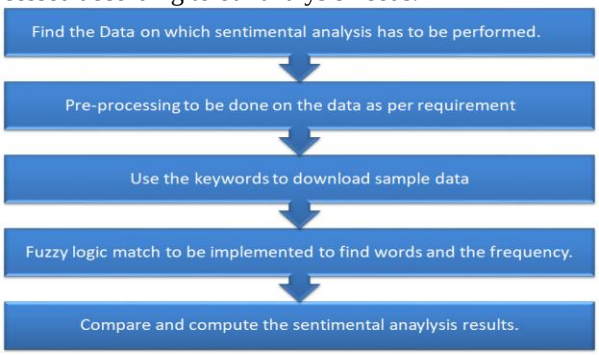

Fig 1 : Flow Diagram of the system

4.1 Proposed System

The proposed system architecture which we want to pursue in this paper fulfils the aspect of sentiment analysis over bilingual datasets across various domains. To obtain this, we consider two different domains which are directly connected to the language database from which they extract the data directly to perform feature extraction. Now explaining the architecture diagram for the process as shown in Fig 2.

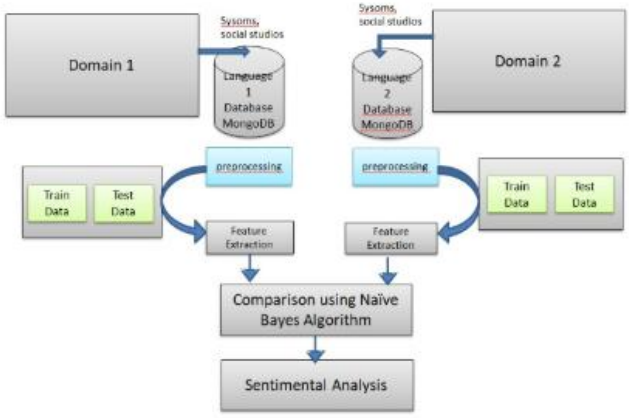

Fig 2: Architecture Diagram

At first the domains have to be selected on which cross domain sentiment analysis has to be performed[2]. After the selection of the domains pre processing has to be done which is the first step of machine learning approach towards sentimental analysis. In pre processing basically the following are carried out. For bilingual approach we have mongoDB databases which already have fed in data for different languages available across domains for example : English and Hindi or English and Tamil can be considered as Bilingual. So we have to make sure that all the datasets that are available can be used for cross domain bilingual sentiment analysis. So after the basic pre-processing of data[13] we see to it which language is required for the particular unknown data which we have processed. Now the challenge is to overcome the problems previously faced in single domain analysis when different languages were used for unknown datasets. In our approach this challenge is met when we check the data accordingly and for feature extraction we relate it to that particular language database and it can be achieved using character set comparison to find out which language dataset has to be called for to solve the issue. After the preprocessing is done and the bilingual part is accessed now feature extraction is to be carried out for each domain separately. Now what feature extraction[16] does is as explained :highlight extraction begins from an underlying arrangement of estimated information and manufactures inferred values (highlights) planned to be instructive and non-excess, encouraging the ensuing learning and speculation steps, and now and again prompting better human elucidations. Highlight extraction is a dimensionality decrease process, where an underlying arrangement of crude factors is diminished to increasingly sensible gatherings (highlights) for preparing, while still precisely and totally depicting the first informational collection. Presently include extraction is completed utilizing the two essential standardization methods Lemmatization: It is used to reduce a different inflectional form of the word to its root or headword which called as its 'lemma. A 'lemma' is simply "Dictionary form" of a word.

Stemming: Decreases terms to their stems in data recovery. It is the unrefined slashing of attaches and is language subordinate for example mechanize, programmed or computerization all inspire diminished to automate.

Further to the feature extraction, lies the analysis of datasets gathered as a result of preceding process[17]. Thereafter in the current phase we perform sentiment analysis over the datasets obtained from both the domains after normalization, individually, using Fuzzy Match Logic[17] along with the available mongoDB datasets, which outputs a certain probability value linked to the polarity of the data. Furthermore, after fetching individual results from both datasets, we merge them, to again perform a combined analysis over the probability values using Fuzzy Match Logic. This outputs our final result for sentiment analysis over bilingual cross domain platform.

The problem being addressed in this paper is that the sentiment towards a particular entity is not constrained to just one language or one domain, so we needed a platform that can address sentiments across languages and domains collectively[18]. Also, a major challenge was merging bilingual datasets before performing sentiment analysis, which increases the complexities of the algorithm being used. Thus, we implement sentiment analysis over bilingual cross domain platform using above mentioned architecture.

V. MODULE DESCRIPTION

As mentioned in [17], ML approaches have proven to be more efficient than lexicon based approaches in terms of syntactic feature selection and co-occurrence of data. Thus, we use pre-tested supervised ML approaches to perform a sentiment analysis over bilingual cross domain platform. The module setup features can be derived from Fig 3.

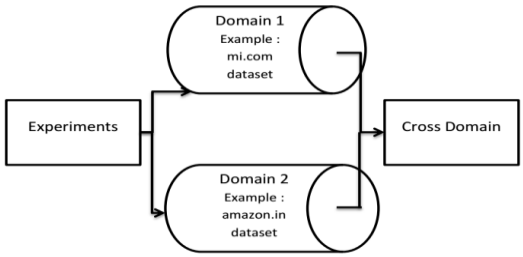

Fig 3: Module Setup

5.1 Basic Data Pre-processing:

The basic pre-processing[1] steps would include the following elements. Text case handling: Convert all text to lowercase. Whitespace management: Remove and trim extra spaces in between characters. Remove punctuation: Remove all punctuation characters Exaggeration expressions. For example: Expressions such as ' I'm extreeeeemly excited for this product' is converted to ' I'm extremely excited for this product', Email address, websites, phone number Remove the emails, website reference and phone numbers from the text. Handling Abbreviations: Handle abbreviations by removing or replacing them with appropriate context. Stopwords removal: Removing generic words of the English language which do not provide any context towards meaningful sentences.

Slang words expansion: Modify and expand the slang words used on the text since sources are from social media as well.
Digits handling: Handling digits by removing them from the given context of words as the numbers usually do not represent content significant enough for the spam/ham model or the scoring model.
Algorithm for Pre-processing is explained ahead. The data exported is first sent into the HTML parser. This parser removes all the tags which are considered to make the data sparse. For example when we the parser encounters something like 'raw milk'. The resultant data would be Data : raw milk. As we see the above data gives the insights into what the raw data is about. After that, the data got from HTML parser is then next sent to Basic cleaning. The various types of data that are handled are: email addresses and Website link. Example of the are:
"next1\nnext2\nhttp://google.com/la1/lah1/\nnext3\nnext4\nhttp://google.com/la2/lah2/\nnext5\nnext6"
The resultant output will be:next1 next2 next3 next4 next5 next6
Further we see, Removing the digits is also handled in this. Then, final cleansed data is stored for the purpose of analysis.

5.2 Intermediate Pre-Processing:

After the data is got from Preprocessing the data has to be verified, we involve the Intermediate pre-processing[17], which includes handling the elements that appear and have to classified post the basic clean up of text.

Algorithm for Intermediate Pre-Processing is expressed ahead. First, named Entity Recognition: Named-Entity Recognition (NER) is a subtask of data extraction that looks to find and order named element makes reference to in unstructured content into pre-characterized classifications, for example, the individual names, associations, areas, medicinal codes, time articulations, amounts, money related qualities, rates, and so forth. Most research on NER frameworks[19] has been organized as taking an unannotated square of content, taking into case of one such information: Jim purchased 300 offers of Acme Corp. in 2006. Furthermore, delivering an explained square of content that features the names of elements: [Jim]Person purchased 300 offers of [Acme Corp.]Organization in [2006]Time. In this precedent, an individual name comprising of one token, a two-token organization name, and a worldly articulation have been identified.

Further, during the preprocessing the data might have lost is originality when various characters are being removed from the data. To cross verify such instances, we do a spell check. Spelling correction for the mistakes made on certain basic words have to be converted into an appropriate format which is being chosen based on high relevance to the nearest word.
Thereafter, the word check has been done the data might have to be re-written with some sensible words to make the sentence appropriate. Substitute similar words using word2vec[18] to replace words in the given data that carry the same meaning but represented in different ways.

5.3 Normalization:

Stemming: Lessens terms to their stems in data recovery. It is the unrefined slashing of joins and is language subordinate for e.g. automate, automatic or automation all get reduced to automate.
Lemmatization: *Lemmatization* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma* .
Example: the boy's toys are various types will be lemmatized as: the boy toy be vary color.
Eg: For instance, in English, the action word 'to sleep' may show up as 'sleep', 'slept', 'sleeps', 'sleeping'. The base structure, 'sleep', that one may turn upward in a lexicon, is known as the lemma for the word[14].
HTML/XML/JSON Parsing :It clears all Html tags,jquery, and hex codes present in scaped data.Example: " i am a sample text<\div> !function(){e.getId()} "to " i am a sample text" and converting "</Xa0i am good and/Xu0nice" to "i am good and nice".

5.3 Model Building :

Once the data has been cleaned and the features have been highlighted and selected we will proceed further towards modeling with the data, during this phase we will identify conversations that are relevant to the context and identify the polarity in the conversations[14].

5.3.1 Spam/Ham models:

A confusion matrix[Fig:5] is a table that is regularly used to portray the execution of a characterization model (or "classifier") on a lot of test information for which the genuine qualities are known. It permits the representation of the execution of a calculation. It permits simple distinguishing proof of confusion between classes[13] for example one class is usually mislabeled as the other. Most execution measures are processed from the disarray framework.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Fig 5:Confusion Matrix

With the above table we will able to find out who well our Machine is trained.
TP-True Positive
FN-False Negative
FP-False Positive
TN-True Negative

5.3.2 N-Gram:

A n-gram is a coterminous grouping of n things from a given succession of content. Given a sentence, s, we can build a rundown of n-grams from s by discovering sets of words that happen by one another. For instance, given the sentence "I am Sam" you can build bigrams (n-grams of length 2) by finding successive sets of words[18].

5.4 Sentimental Analysis Model Building:

First step involves, training the Classifiers: From the data downloaded from the websites we chooses the column for training the sentimental analysis model. Taken into example are few tweets which will help us classify and train the model, as depicted by Fig 6.

Sentiment	Tweets
Negative	@united is the worst. Nonrefundable First class tickets? Oh because when you select Global/FC their system auto selects economy w/upgrade.
	@united I will not be flying you again
Neutral	@VirginAmerica my drivers license is expired by a little over a month. Can I fly Friday morning using my expired license?
	@VirginAmerica any plans to start flying direct from DAL to LAS?
Positive	@VirginAmerica done! Thank you for the quick response, apparently faster than sitting on hold ;)
	@united I appreciate your efforts getting me home!

Fig 6: Classifier Training

Next step encounters, extraction of feature vector: One vital advance in developing a classifier is selecting the highlights of the information that are applicable, and how to encode those highlights. For instance, we can utilize the completion letter of the names as an element and fabricate a classifier to recognize sexual orientation with these unmistakable highlights. In particular, names finishing in an, e and I are probably going to be female, while names finishing off with k, o, r, s and t are probably going to be male. Additionally, we can utilize the nearness or nonattendance of words that show up in tweet as highlights. In the preparation information, we can part each tweet into words and add each word to the element vector[15]. A portion of the words probably won't demonstrate the feeling of a tweet and we can sift them through. At that point consolidate singular component vector into an expansive rundown that contains every one of the highlights and evacuate copies in this rundown. Adding singular words to the element vector is alluded to as 'unigrams' approach[16]. Here, for effortlessness, we will just consider the unigrams and underneath are a few instances of highlights extricated from tweets.

5.4.1 Cross Domain Sentiment Analysis Approaches:

There are two fundamental strategies for SA progression, which are Lexicon-Based Approaches (LBA) and Machine Learning(ML) approaches[17]. LBA decides the supposition or extremity of supposition by means of some capacity of conclusion words in the report or the sentence. It depends on syntactic or co-event designs and furthermore a seed rundown of supposition words to discover other conclusion words in a substantial corpus. In ML approaches, an administered or unsupervised strategy is utilized to recognize supposition dependent on calculations that have been effectively demonstrated in organized information mining strategies[18]. Now explaining about Lexicon approach: Our collection which if analyzed by using a lexicon approach combined with a linguistic analysis in order to detect sentiments, during a period of time, in social and political tweets. The lexicon approach starts with a list of positive and negative words, which are already pre-coded for polarity[19]. A linguistic analysis, in contrast, exploits the grammatical structure of text to predict its polarity, in conjunction with the lexicon . Words contained in a tweet are classified into positive or negative by using the previous lexicon. Nevertheless, this methodology does not takes into account the sarcasm which transforms the polarity of an apparently positive or negative utterance into its opposite . But by analyzing a big corpus the sarcasm rest minimum and do not contributes to inflate in a big amount the percentage of the total results. The corpus is pre-processed in order to extract stop-words, punctuation, links, etc. Then, the Spanish and the English translated lexicon[19], respectively, are used to count for each tweet and for each corpus the number of positive and negative words contained in each tweet. Now the types of algorithms used for cross domain sentiment analysis in machine learning is given below :

1. Random Forest Classifier (RFC)

RFC is a strategy proposed by Breiman and is a mix grouping technique dependent on measurable learning hypothesis . It depends on outfit technique different choices tree arrangement. Outfit techniques have been known as the most compelling improvement in Data Mining and ML in the past decade . It is a procedure for consolidating numerous models into the one that produces progressively exact outcomes. Different tests are drawn utilizing resampling strategy and order trees are constructed comparing to each bootstrap test[15] . RFC assemble expectation outfits utilizing choice trees are produced haphazardly in chosen subspace of information. The principle thought is to get decorrelated choice trees from irregular information in datasets and creates conglomeration from the outcomes. RF has been connected in numerous areas, for example, debacle the executives , biomedical and science . Specialists that have connected RFC classifiers have done as such when they have to break down huge datasets. RFC is proposed to defeat overfitting model by utilizing choice tree as a method. Numerous endeavors have been made to utilize RFC in content mining[15]. Printed information are normally described by high dimensional highlight spaces, it is fundamental that the picked classifier performs well inside this setting . The propensity for the classifier to over fit the preparation information amid arrangement is high due to the high space dimensionality for printed information.

Advantages of using this classifier comprises incremental robustness for increased number of datasets, accuracy for decorrelated data from random datasets. But a major disadvantage of using this classifier over textual data is the tendency of overfitting, which increases the complexity of the algorithm with increase in number of datasets[15]. The performance comparison of this classifier along with others will be dealt later in this paper.

2.Stochastic Gradient Descent (SGD)

SGD otherwise called steady inclination plunge is a stochastic guess of the slope plummet advancement strategy for limiting a target work that is composed as an aggregate of

differentiable capacities. SGD endeavors to discover essentials or on the other hand maximums by iteration. SGD is a basic yet extremely effective way to deal with discriminative learning of straight classifiers under curved misfortune capacities, for example, (straight) Support Vector Machines and Calculated Regression[16]. Despite the fact that SGD has been near in the ML people group for quite a while, it has gotten a extensive measure of consideration only as of late with regards to expansive scale learning. The upsides of Stochastic Gradient Plummet are proficiency and simplicity of usage .Notwithstanding, impediments of SGD are that SGD requires a number of hyper-parameters, for example, the regularization parameter and the quantity of cycles. It is likewise delicate to include scaling.

Due to the inefficiency of this algorithm over feature based data as mentioned in the previous paragraph, we cannot use it as a base for sentiment analysis.

3.Multinomial Naive Bayes (MNB)

Naive Bayes classifiers (NBC) are basic probabilistic classifiers dependent on applying Bayes' hypothesis with solid (naive) freedom presumptions between the highlights . NBC can forecast class enrollment trials, for ex: the similarity that a given tuple has a place with a specific class by accepting that the parameter estimation of a given class is autonomous of the estimations of alternate parameters. MNB is a specific rendition of Naive Bayes that is structured more for content records. In the MNB, highlight vectors speak to the frequencies with which certain occasions have been created by a multinomial where is the likelihood that occasion it happens[20]. MNB accept that all characteristics (i.e., highlights) are autonomous of one another given the setting of the class, and it disregards all conditions among characteristics. One preferred standpoint of the MNB demonstrate is that it can make probability calculations up to the mark and thus it is used in cross domain sentiment analysis while merging the two single domain results.

The main advantages of preferring this algorithm over others is its efficiency over noisy textual data and independent data, since it has a probabilistic approach. According to Ang et. el[20], the accuracy of NBA over twitter datasets consisting thousands of tweets covers an accuracy of 84.06%.

The approach over NBA in this paper can be explained as follows: We consider Nd=10 datasets each from two domains(A and B) over which individual sentiment analysis is being performed. Let i refer each dataset and j refer the class of data(Positive and Negative). Thus we apply Bayes' Theorem over these datasets using below mentioned formula to find individual probabilities of each dataset:

$$P(i/j) = [P(j/i)*P(j)] / P(i)$$

Once we obtain overall probabilities(P(A) and P(B)) from both the domains, we apply the same formula over them to find out the final probability over cross domain to fetch the result of sentiment analysis. Thus after explanation of why Naive Bayes Approach is chosen is explained previously and now the sequence diagram is illustrated in Fig 7. The whole process is explained in the diagram step by step.

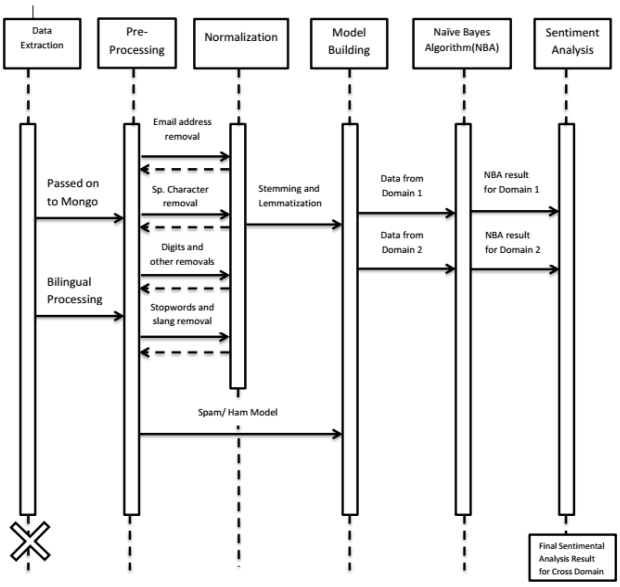

Fig 7: Sequence Diagram for sentiment analysis over cross domain platform.

Bilingual Analysis:

Bilingual Analysis has been adapted to find a common passage of different domain which has different languages. For example: Given below is the description of the product Burt's Bees 100% Natural Tinted Lip Balm. The below attached image is that of the German market, as given in Fig 8.

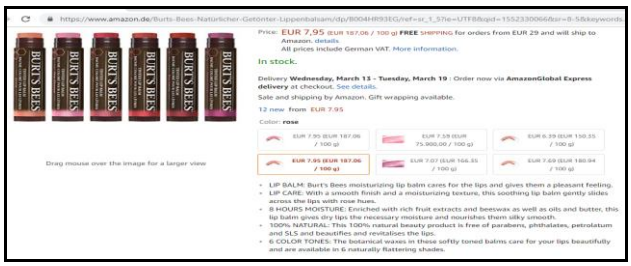

Fig 8: German Market

Another image which is of the same product from the Chinese market as shown in Fig 9:

Fig 9: Chinese Market

When it comes to analyzing such products we need a common domain through which this can be done. We translate the pages to common languages and set out our analysis. For translating the pages the stepwise algorithm is discussed. Firstly, we analyze the page and the language, we find the part to which needs to be translated so that we can translate the only the which are needed for us. Next, there are multiple translating domains which will help us translate the pages. We hit the API of such domains and then store the respective results for our analysis. For example: Yandex is domain which is used for German translation as shown in Fig 10.

Fig 10: German Translation

Whereas Baidu is used for Chinese translation as shown in Fig 11:

Fig 11: Chinese Translation

After that, once the data has been translated to the common language we can use it for sentimental analysis.

VI. EXPERIMENT AND RESULT ANALYSIS

In this area we give a short depiction of every data set we utilized. To explore to grouping issue we needed to analyze the utilization of comparative information with a French commented on corpora of tweets managing governmental issues and with a gathering of tweets managing corporate substances notoriety. The corpus investigated concerns 800 tweets containing #AMLO that were removed between the time of 9, 10 and 11 June 2012. AMLO is acrynom for Andr´es Manuel L´opez Obrador. Description of the Spanish political set:

Table 1. Class distribution in both complete and French sub-part collection

Class	Class-Distribution	Class-Distribution (French)
Negative	0.41	0.37
Neutral	0.29	0.30
Positive	0.30	0.33

As appeared Table 1, classes are very much offset with just a marginally distinction with negative tweets for the total accumulation just as for the French subpart.

Statistics on the French political: Table 2 demonstrates that the primary inclination is negative mind a not many number of unbiased tweets. The principle reason is that governmental issues in France release interests between individuals.

Table 2. Class distribution in the French political collection

Class	Class-Distribution
Negative	0.60
Neutral	0.12
Positive	0.28

Statistics on the Spanish annotated set: Table 3 demonstrates that the principle propensity of the RepLab set is sure. Intersection this point with the negative view from the French accumulation ought to give an intriguing outcome.

Table 3. Class distribution in the Spanish reputation collection

Class	Class-Distribution
Negative	0.24
Neutral	0.28
Positive	0.48

Tables 4 and 5 summarize the experimental results of our proposal concerning the tweets polarity.

Table 4. Polarity classification results using French set

Method	F-Score	Accuracy
Baseline	**0.39**	**0.42**
Cosine	0.24	0.36
SVM	0.33	0.37

Table 5. Polarity classification results using Spanish set

Method	F-Score	Accuracy
Baseline	0.50	0.51
Cosine	**0.74**	**0.74**
ElhPolar Lexicon	0.25	0.32
Translated Lexicon	0.21	0.33
SVM	0.17	0.31

Table 5 demonstrates characterization execution over Spanish substance as indicated by F-Score and Accuracy. A fascinating outcome is the frameworks' positioning while the Cosine closeness was beaten with the French sets it is intriguing to see that we can acquire very great arrangement results that are near between annotator understandings saw in the writing, while SVM execution significantly diminishes. Benchmark execution is additionally very intriguing since his execution increments, however, remain lower than the Cosine. It was fundamentally better on the French set. Both vocabulary approaches (ElhPolardictionary and Bing Liu made an interpretation of one) appear to not accommodate our information collection or this sort of examination since they don't perform well.

VII. CONCLUSION AND FUTURE WORKS

In this paper, we formulate a method in which we conduct sentimental analysis for cross domains in a bilingual platform. It generally tries to learn the similitudes and regular highlights and then compare it domain wise. After the point is reached where the sentimental analysis can be done the languages are checked and mapped according to the algorithms given. Thus we have achieved solving the problem of cross domain sentiment analysis using naive bayes approach and the bilingual part is also taken care of as explained with examples. Future work will be based on applying it with

video and text comparison. And it will be based on image categorization, as image comparison with help of advanced machine learning techniques as in future better ways of handling bilingual videos or texts in pictures can be handled.

REFERENCES

1. Pengfei Wei, Yiping Ke, Chi Keong Goh, 2018: A General Domain Specific Feature Transfer
2. Framework for Hybrid Domain Adaptation , Nanyang Technological University, Singapore,
3. Rolls-Royce Advanced Technology Centre, Singapore.
4. Hassan Saif, Yulan He and Harith Alani,2012: Semantic Sentiment Analysis of Twitter Knowledge Media Institute. The Open University, United Kingdom.
5. Godbole, N.; Srinivasaiah, M.; and Skiena, S. 2007. Large-Scale Sentiment Analysis for News and Blogs. In ICWSM'07.
6. Mikhail Bautin, Lohit Vijayarenu, Steven skiena, 2008 : International Sentiment Analysis for News and Blogs, Association for the Advancement of Artificial Intelligence.
7. Mehler, A., Bao, Y., Li, X., Wang, Y., Skiena, S.: Spatial analysis of news sources. IEEE Trans. Visualization and Computer Graphics 12 (2006)
8. Benamara, F.; Cesarano, C.; Picariello, A.; Reforgiato, D.; and Subrahmanian, V. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In ICWSM'07.
9. Ethem Alpaydin. 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.
10. Claude E. Shannon and Warren Weaver. 1963. A Mathematical Theory of Communication. University of Illinois Press, Champaign, IL, USA.
11. Yi, J.; Nasukawa, T.; Bunescu, R.; and Niblack, W. 2003.Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In ICDM '03, 427. Washington, DC, USA: IEEE Computer Society.
12. Brian Heredia, Joseph Prusa 2016 Cross-Domain Sentiment Analysis: An Empirical Investigation
13. Tiangu Zhang, Xiaoshan Yang 2015 Cross-Domain Feature Learning in Multimedia.
14. Bowen Zhang, Min Yang, Xiaoshun Chen 2017 Cross Domain Sentiment Classification by Capsule Network with Semantic Rules.
15. P. Sanju, T.T. Mrinalinee 2013 Cross Domain Sentiment Classification Using Sentiment Sensitive Thesaurus.
16. Wenpeng Yin, Quiang Qu, Wenting tu 2016 Neural Attentive Network for Cross Domain Aspect-Level Sentiment Classification.
17. L Brieman 2001 Random Forests in Machine Learning.
18. N. A. Jabeseeli and E. Kirubakaran, A Survey on Sentiment Analysis of (product) Reviews, 2012.
19. S. Mahalakshmi and E. Sivasankar, Cross Domain Sentiment Analysis using Different Machine Learning Techniques, Conference on Fuzzy and Neuro Computing, 2015.
20. O. Abdelwahab, M. Bhagat, C.J Lowrance, Effect of Training Set Size on SVM and Naive Bayes for Twitter Sentiment Analysis, 2015.
21. K. R McKeown and V. Hatzivassiloglou, Predicting the Sentiment Orientation of Adjectives , 1997.
22. T. Wilson, J. Wiebe and P. Hoffmann, Recognizing Contextual polarity in Phase-Level Sentiment Analysis, 2005.

AUTHORS PROFILE

S. Arun Kumar originating from Chennai is currently working as Assistant Professor in SRM Institute of Science and Technology. He completed his B.E from Anna University and M.Tech from SRM University(now SRM Institute of Science and Technology). His main areas of interest are Cloud Computing, Network Security and Privacy Preservation.

Dipon Sengupta originating from West Bengal is currently pursuing his B.Tech Degree from SRM Institute of Science and Technology. He has been to National Taipei University of Technology for Semester Abroad Program. His main areas of interest are Information Security and Internet of Things.

Rishendra Ravi originating from Bihar is currently pursuing his B.Tech from SRM Institute of Science and Technology. His main area of interest is Machine Learning and Web Development.

M . Sanjanaa Sri hailing from Chennai, pursuing B.tech from SRM Institute of Science and Technology. She has been to National Tsing Hua University for Semester Abroad Program. She has been a part of a Major project in Taiwan. Her main area of interest is Internet of things (IoT).

Arhant Chatterjee originating from Jharkhand is currently pursuing his B.Tech from SRM Institute of Science and Technology. His main area of interest is Machine Learning(ML).