International Journal of Scientific & Engineering Research, Volume 7, Issue 9, September-2016
ISSN 2229-5518

Semantic Analysis on Twitter Data Generated by Indian Users

Prof. Prashasti Kanikar, Rajitha Koppisetty, Shubadra Govindan, Sharang Bhat, Mansi Virani
Computer Engineering Department
Mukesh Patel School of Technology Management and Engineering
NMIMS University, Mumbai

***Abstract*—The rise of social media has brought about a new realm on which people are free to share their opinions. Twitter, Facebook, LinkedIn, Google+, etc. have become the most sought after platforms for people to share their views about the latest events, trends and products. This has provided opportunists with oceans of data, enabling them to understand human behaviour in ways which was thought impossible previously, called Social Media Mining. Social Media Mining entails representation, analysing and further, extraction of actionable patterns from the data collected from social media. This information enables companies - evaluating consumer behaviour or public opinion - to adopt a data based decision making process as opposed to an intuitive decision making. The existing systems however, deal with extraction of polarity and not further isolation of emotions into various classes. Besides, informal interactions on social media platforms has brought about a spin on language semantics. Unlike formal documentation, people are free to express themselves in multiple languages, by mixing two languages and most notably in India, in Hinglish, a culmination of English and Hindi. The intermixing and code-switching of text adds a new dimension in terms of language processing and mining. This paper explores the challenges faced by social media mining and proposes a system for emotion detection in English, Hindi and Hinglish languages.**

Keywords:*Sentiment, Hinglish, Dictionary Search, Nave Bayes, Sentiment Scoring, Word Cloud*

I. Introduction

Social Media presents a wide range of opportunities for individuals as well as e-commerce organizations to understand the society and their opinions. With the exponential rise in the use of internet by individuals and organizations for the purpose of communication, the evolution of social media platforms has gained business value. Twitter, Facebook, LinkedIn and public opinion websites are spearheading this revolution. Social Media Mining, introduces basic concepts and principal algorithms suitable for investigating massive social media data; it discusses theories and methodologies from different disciplines such as computer science, data mining, machine learning, social network analysis, network science, sociology, ethnography, statistics, optimization, and mathematics [4]. It involves the usage of tools to formally represent, measure, model and mine meaningful patterns from large-scale social media data.

Semantic Analysis in particular, deals with understanding of the opinions presented in the dataset to provide meaningful understanding of the overall emotion that is portrayed. This could be in the form of polarity analysis - Analysis and classification of the statements as being negative or positive; or sentiment analysis - Deeper analysis and understanding of the emotions conveyed via text, e.g., anger, sadness, happiness,etc. Language as a tool, being fluid undergoes changes and adapts to the colloquial habits of people residing in various locations. Within India, Hindi is spoken as widely, if not more than English. The combination of the two languages is finding increasing popularity, however there is no system to exploit the wide range of opinions expressed here.

The system proposed in this paper aims to perform deep sentiment analysis not only on one language, i.e., English, but also Hindi and a combination of English and Hindi referred to as Hinglish.

II. Literature Review

Existing Literature on the topic of Social Media Mining and Sentiment Analysis revealed two major problems for implementing the system[1]-

- Emotions and opinions expressed in multiple languages, often code-switched
- Complexity in isolation of emotions other than negative and positive

In India, users tweet and express sentiments using Hindi as well. Many international companies looking to establish a base in India also use Hindi as a medium to communicate with their audience and customer base. For example, Amazon India recently had promotional offers and promoted the same using Aur Dikhao hashtag on Twitter. Hinglish is also used for casual communication among friends, for example, Main temple ke pass hoon meaning I am near the temple. There are plenty of research works focusing on analyzing texts used in popular forums, like online social groups, for applications like opinion mining, sentiment analysis, etc.[2] Fig 1 shows a few Indian users who took to Twitter to share their ideas.

Code-switching is the practice of moving back and forth between two languages in spoken or written form of communication [5]. Identifying word-level languages in code-switched texts is associated with many challenge -

- People often use non-standard English transliterated forms of for example, Hindi words.
- The transliterated Hindi words are often confused with English words having the same spelling.
- Inconsistent spelling usage: Despite the availability of the standards for transliteration of Devanagari script to

Fig. 1. Example of tweet written in Hinglish

Roman script (the Hindi language is based on Devanagari script while the English language is based on Roman script), people tend to use many inconsistent spellings for the same word. For example, the most common English transliteration for the Hindi word is mai, as observed from our data set. Analysis shows that people often use mein or main as alternatives.

- Ambiguous word usage: The transliterated word main, for the Hindi word could be misinterpreted by a machine to be the English word.

To understand the sentiment of any user opinion, the semantic analysis of the statement is the first step. Semantic analysis is the procedure of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings. Moving on to semantic analysis, we delve deeper to check whether the instructions or opinions form a sensible set of instructions. Whereas any old noun phrase followed by some verb phrase makes a syntactically correct English sentence, a semantically correct one has subject verb agreement, proper use of gender, and the components go together to express an idea that makes sense. A few criteria are to be fulfilled for the program to be semantically valid. The variables, classes, functions and the like must be accurately defined. Further, the usage of expressions and variables should respect the type system as well as the access controls of the system. Semantic analysis is one of the final phases at the front end. It is also the compiler's duty to filter out any incorrect statements. The system was implemented using semantic analysis and not any other domain of language understanding, as per the requirement of the system. Below is a list of other areas of language processing as compared to semantic analysis-

- Phonetics: the study of linguistic sounds
- Morphology: the study of the meaning components of words
- Syntax: the study of the structural relationship between words see the parse tree of the football example
- Semantics: the study of meaning
- Pragmatics: the study of how language is used to accomplish goals; discourse conventions (turn taking, politeness, etc.); relation between language and context-of-use

III. Proposed System

The proposed project aims to work on Mining Sentiments and Opinions of users on the web by using text data in Hindi, English and Hinglish available on social media platforms such as Twitter, Facebook and Blogs. The project employs text mining techniques and spans across language processing. The proposed system utilizes data from social networking websites to

- Understand the thought process of the general public. The system aims to evaluate the opinions (positive/negative) and sentiments (angry/happy) of the public.
- Comprehend the social media data available in English and Hindi, i.e., the proposed system should be able to process code-switched sentences.

IV. Analysis and Design

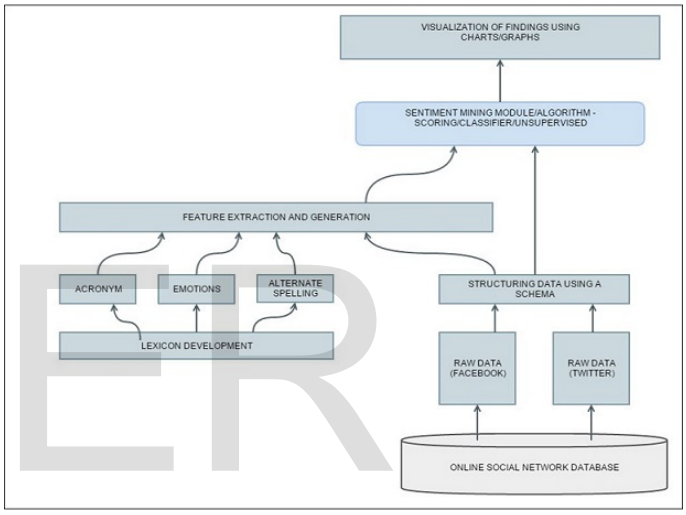


Fig. 2. Design of the system

The source for the data for the system is the Twitter corpus that is being constantly modified by the generation and addition of new tweets all over the world. This raw data consists of hashtags, spelling ambiguities, and even sentences consisting of words in different languages. The raw data needs to be structured in order to extract and retrieve the required information. This is done by removing the punctuations, hashtag signs, etc. that do not add any meaning to the tweet overall. During this step, two dictionaries are also built consisting of all the words in the tweets that add meaning to it and that determine the overall emotion/sentiment of the tweet, one for positive and another for negative. It takes all the possible words into account, including acronyms and spelling ambiguities. Hence, eventually two full- fledged dictionaries consisting of meaningful words from the tweets are created which are further used to classify the emotions as well as the tweet.

The tweets need to be classified as overall positive, negative, neutral or the degree of the emotion such as anger, disgust, etc. For searching the words in the two dictionaries, a dictionary

search algorithm is used as opposed to linear search. For determining the polarity of the tweet, a Scoring Algorithm is used. Further for classification of emotions, Nave Bayes Classifier is implemented. The algorithms are explained in detail in further sections.

For visualization of the results, a scatter plot is used which shows the overall sentiment of a Twitter handle on a scale from -5 to +5. Also the Word Cloud Algorithm gives a visual representation of words commonly associated with the handle. The WebApp being developed is divided into the following modules:

- Server
  The SERVER.R file defines the variables, place-holders and interacts with the UI.R file which resides in the client UI which is the local host. The server.R fi le contains the algorithmic implementation of the scoring engine, the Word Cloud generator and the back-end script for the data shaping techniques. As the tweets received are in a JSON format, they must be parsed as a data frame before the corpus semantics are applied on the text contained in the tweets. This back-end script parses the received tweets as a data frame which is then operated upon by the various modules. It also defines the plot outputs, plotting functions and Word Cloud format for the UI component to display. The SERVER.js files contains the Twitter credentials needed to make a GET request to the Twitter Search and Streaming APIs.
- Local Host Process-User Interface
  The UI.R file defines the User Interface and tags each element in the WebApp as an HTML tag so as to provide a structure to the elements, such as the text and plot outputs.
- Caching Mechanism
  Web Application Caching involves the storing of data generated on the fly- that is, dynamically for reuse and in turn, leaving the data closer to the end user. Caching is used at a variety of levels to improve performance and efficiency of web applications. Generally, caching reduces work load and results in users receiving content faster.

V. Implementation

The system implemented design is depicted in Fig 3. the elements of which have been elaborated in the following subsections.

A. Search Algorithm: Dictionary Search

The Dictionary Search was chosen over the Linear Search for its efficiency. Once the data is sorted, the best case performance is one comparison (the searched-for value is the mid-point value in the original array). The worst-case, where the value is not in the array, is $log_2 N$, where N is the size of the array. The feasibility of the search algorithm was measured based on the following parameters:

- Choice of compiler and execution environment
- Random events caused by other programs running during execution of the test

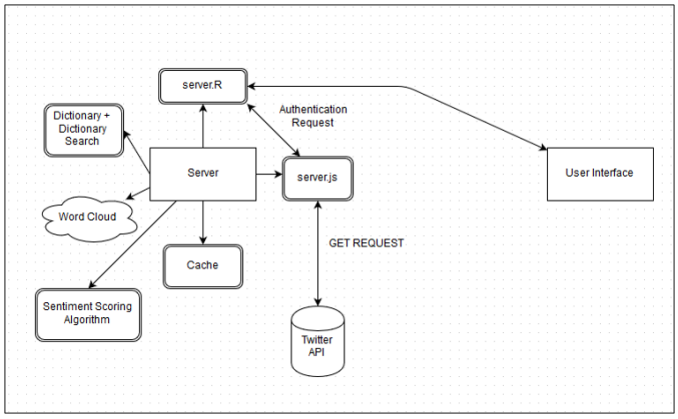


Fig. 3. Implementation Design

The dictionary search matches words in a predefined dictionary connected to the ML.R file (machine learning (ML) component within the server). The operation of the ML component can be explained with the following example. Consider the word nice. The search operation runs as shown in Fig 4. It

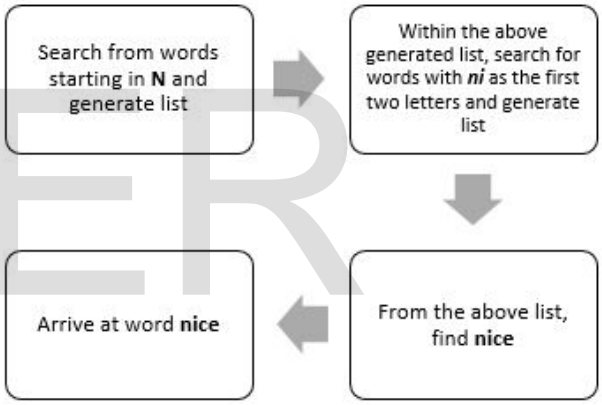


Fig. 4. Dictionary Search Process

breaks the search complexity into halves by reducing the total number of computations. The linear search on the other hand matches the search keyword in the dictionary linearly, word by word. For a dataset containing 10000 words, it runs in O(n) time as compared to a dictionary search which runs in O(log(n)) time.

B. Sentiment Scoring Engine

The general idea is to calculate a sentiment score for each tweet so one can know how positive or negative is the posted message. The sentiment scoring engine currently determines the score of each tweet as per the following algorithm:
$Score = (No.of Positive Words)(No.of Negative Words$

- If $Score > 0$, this means that the sentence has an overall 'positive opinion'
- If $Score < 0$, this means that the sentence has an overall 'negative opinion'

- If $Score = 0$, then the sentence is considered to be a 'neutral opinion'

The result of the sentiment analysis for input search query in terms of positive and negative tweets are shown in figures 5, 6.

Fig. 5. Plotting the sentiment analysis results of a search query

Fig. 6. User Interface of the System

C. Cholesky Decomposition for Determining Unknown Sentiment

Term 1	Term 2	Output Sentiment
Positive	Negative	Negative
Negative	Negative	Positive
Negative	Positive	Negative
Positive	Positive	Positive

Fig. 7. Determining Sentiment of phrase

Cholesky Factorization or Decomposition is breaking-down of a positive-definite Hermitian matrix into the product of a lower triangular matrix and its conjugate transpose. When applicable, the Cholesky factorization is roughly two times as efficient as the Lower Upper Decomposition or LU Factorization for solving systems of linear equations.

Upon encountering terms such as not bad, too good, not happy and other terms followed by or following a conjunction, the ml.R forms a matrix of known words and classifies each individual word as positive or negative. It computes sentiment polarity using a simple multiplication of polarity coefficient and produces results as per the following table.

D. Corpus Structuring

A corpus or corpora in the plural form is an enormous structured set of texts. This is usually stored and processed in the electronic form these days. These structures are used to perform statistical analysis and hypothesis testing, validating linguistic rules, or checking occurrences within a specific language territory.

Aligned Parallel Corpora are derived from multilingual corpora which are formatted for any simultaneous comparison. Translation Corpus and Comparable Corpus are the two kinds of parallel corpora. Texts in one language which are the translations of text in another language are contained in Translation Corpus. On the other hand, in a comparable corpus, texts are of the same type, they cover the same content and are not any form of translations.

Forms of text alignment that identify equivalent text segments (consisting of phrases and sentences) are prerequisites for analysis of parallel text. Parallel fragments comprising two language corpora (the second one being an element-for-element translation of the first) are used to train machine translation algorithms.

E. Naive Bayes Classification

Nave Bayes is an algorithm used to construct classifiers. Classifiers are models which assign class labels to problem instances. These problems are represented as vectors of certain feature values. The class labels are derived from a finite set. The family of algorithms for training classifiers are based on a common principle- all such Nave Bayes classifiers work with the assumption that the value of a specific feature is not dependent on the value of any other feature.[3][6]

As an example, a vegetable may be considered to be a pumpkin if it is yellow, round, large and has a green patterned outer skin. For a Nave Bayes classifier, all these factors are considered to contribute individually and independently to the probability that the vegetable in question is indeed a pumpkin. This is considered regardless of any associations between size, color and other features.

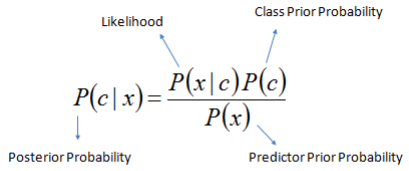
It is a classification technique, wherein-

- Data items in the data set are assigned to target classes
- Classes are predetermined using a training set
- Supervised learning target values are known

Naive Bayes Algorithm uses Bayes Rule to calculate the most probable hypothesis. It assumes that the presence or absence of a particular characteristic is independent of another.

F. Word Cloud

Word Clouds (alternatively referred to as Text or Tag Clouds) work in a modest manner. When the frequency of occurrence of a specific word in the data source increases, the



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. 8. Mathematical Formula - Naive Bayes

bigger and bolder it appears in the Word Cloud representation. The Word Cloud algorithm generates a visualization of most frequently occurring words in a text based on a predefined condition. Tags are usually single words, and the importance of each tag is shown with font size or colour. This format of representation is useful for swiftly perceiving the most prominent terms and for locating a term alphabetically to determine its relative eminence. Word Cloud visualization, when used as website navigation aids, hyper links terms to items associated with a specific tag. Consider an example of web logs. In this case, the Word Cloud frequency would correspond to the number of entries assigned to the web log. Details such as assigning of a smaller font size for lower frequencies can be specified. However, for larger values, a scaling should be made. In a linear normalization, the weight t_i of a descriptor is mapped to a size scale of 1 through f, where t_{min} and t_{max} are specifying the range of available weights.

$$s_i = [f_{max}(t_i - t_{min})/t_{max} - t_{min}]$$

for $t_i > t_{min}$; else $s_i = 1$

- s_i: display fontsize
- f_{max}: max. fontsize
- t_i: count
- t_{min}: min. count
- t_{max}: max. count

Fig. 9. Word Cloud

VI. Results

The results are based on the most recent tweets being tweeted by the users worldwide using a particular handle. The main function of the system is to classify the tweets as positive, negative or neutral and provide us a visual representation of the aggregate of the sentiments expressed by the users. Further, the emotions of the Tweets are also depicted.

VII. Conclusion and Future Scope

This project is aimed at individuals and business firms alike who require user insights, sentiments and attitudes to better market or improve their products. The added languages,Hindi and Hinglish, make this system particularly easy to use for firms that work closely with the Indian population. The various industries that this system could be include marketing, advertising, hospitality, entertain and media, e-commerce, etc. Healthcare firms could use the system to understand their service provided to their patients. Treatments could be logged onto the system to keep a track of the cycle of recovery of the patient and also the best suited methods for faster cures.
The system can be expanded to accommodate a different data source such as Facebook or LinkedIn opinions, comments or statuses to gather more data. Further, addition of languages will enable the system to cater to a larger population.

References

[1] Zafarani, Reza, Mohammad Ali Abbasi, and Huan Liu. Social media mining: an introduction. Cambridge University Press, 2014.
[2] Jhamtani, Harsh, Suleep Kumar Bhogi, and Vaskar Raychoudhury. "Word-level Language Identification in Bi-lingual Code-switched Texts." (2014).
[3] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
[4] Reza Zafarani, Mohammad Ali Abbasi, Huan Liu (2014), Social Media Mining An Introduction, Cambridge University Press
[5] Harsh Jhamtani, Suleep Kumar, Bhogi Vaskar Raychoudhury (2014), Word-level Language Identification in Bi-lingual Code-switched Texts, 28th Pacific Asia Conference on Language, Information and Computation pages 348357.
[6] Chung-Hong Lee, Hsin-Chang Yang (2005), A Classifier-based Text Mining Approach for Evaluating Semantic Relatedness Using Support Vector Machines, International Conference on Information Technology: Coding and Computing, IEEE