I.J. Intelligent Systems and Applications, **2017, 7, 31-40**
Published Online July 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.07.04

Challenges with Sentiment Analysis of On-line Micro-texts

Ritesh Srivastava
Computer Engineering Division, NSIT, Delhi University, New Delhi-INDIA
E-mail: ritesh21july@gmail.com

M.P.S. Bhatia
Computer Engineering Division, NSIT, Delhi University, New Delhi- INDIA
E-mail: mpsbhatia@nsit.ac.in

***Abstract*—**With the evolution of World Wide Web (WWW) 2.0 and the emergence of many micro-blogging and social networking sites like Twitter, the internet has become a massive source of short textual messages called on-line micro-texts, which are limited to a few number of characters (e.g. 140 characters on Twitter). These on-line micro-texts are considered as real-time text streams. On-line micro-texts are extremely subjective; they contain opinions about various events, social issues, personalities, and products. However, despite being so voluminous in quantity, the qualitative nature of these micro-texts is very inconsistent. These qualitative inconsistencies of raw on-line micro-texts impose many challenges in sentiment analysis of on-line micro-texts by using the established methods of sentiment analysis of unstructured reviews. This paper presents many challenges and issues observed during sentiment analysis of On-line Micro-texts.

***Index Terms*—**Sentiment analysis, On-line micro-texts, Natural language processing, Text Mining, Machine learning.

I. Introduction

The sentiment analysis (SA) is defined as an application of natural language processing in which automatic extraction of writer's opinions is performed in a manner to classify (positive or negative) and quantify his or her feelings about the features (e.g. event, product, personality, etc.) present in an electronic document [1, 2].

With the increase in e-commerce and blogging sites during the early years of 2000 decade, the sentiment analysis has drawn the interest of many computational linguistic researchers The SA has gained an appreciated performance regarding accuracy in the midst of the decade of 2000. During this span of time, the raw textual data were mostly scraped from e-commerce and blogging sites in the form of unstructured reviews (e.g. Movies or Products reviews). These reviews were quite consistent enough regarding grammar and spelling. Hence the sentiment analysis of such data is easy to perform and rarely use any intense pre-processing steps to make the raw feed consistent.

In parallel, in mid years of the decade of 2000, some new sources of information emerged in the form of micro-blogging and social networking sites. Due to this, the internet becomes a massive source of short textual messages called "on-line micro-texts", which are limited to a few number of characters (e.g. 140 characters on Twitter). These on-line micro-texts are extremely subjective and contain opinions about various events, social issues, personalities, and products. The massiveness and subjectivity of this enormously big source of data have attracted most of the researchers working in the area of Natural Language Processing (NLP), especially, in sentiment analysis. However, despite being so voluminous in quantity, the qualitative nature of these micro-texts is very inconsistent. These qualitative inconsistencies of raw on-line micro-texts imposed many challenges in the use of established methods of sentiment analysis of consistent textual information.

In early decade of 2000, since the e-commerce and blogging sites were the primary source of information, that's why sentiment analysis was performed offline, without involving any intense preprocessing steps as the texts were consistent enough. At that time, the sentiment analysis favors classical methods of NLP such as Heuristics, Rule Base approaches, chunking of opinion phrases, Dictionary lookup, Parsing, POS tagging, Statistical methods, etc. and involves following steps [1-4]:

(i) Identification of product specific features.
(ii) Identification of opinion words.
(iii) Determination of semantic orientation of contextual polarity opinion words.
(iv) Determination of strength and polarity of opinion words if modified by modifiers.
(v) Association of product features with corresponding opinion words.
(vi) Summarization or overall ranking of the product, based on the scores or strengths and polarity of opinions corresponding to each feature identified in the review.

By the mid of 2000 decade to till date, the micro-blogging sites (e.g. Twitter), instant messenger sites and social networking sites become the largest source of information in the form of on-line micro-texts for sentiment analysis. Unlike to sentiment analysis of unstructured reviews, the sentiment analysis of on-line micro-texts highly favors the Machine Learning (ML) and Statistical approaches of NLP. Contrasting on-line micro-texts with the nature of the textual reviews, these on-line micro-texts are treated as real-time text streams, which needed to be processed in real-time. Since on-line micro-texts are very noisy, it is tough to analyze them in real time.

In this paper, we tried to classify challenges associated with the On-line micro-texts, because of which it is very complicated to apply classical methods of sentiment analysis as Heuristics, Rule Base approaches, chunking of opinion phrases, Dictionary lookup, Parsing, POS tagging. Furthermore, these methods are very time consuming and due to which application of such procedures may affect the real-time analysis process of on-line micro-texts. However, the availability of huge amount on-line micro-text on WWW justifies the success of ML and statistical methods as well they also ensure fast processing with respect to the classical methods of sentiment analysis.

A. Tweets as On-line Micro-texts

In a manner, to illustrate the challenges associated with sentiment analysis of on-line micro-text, we take Twitter's data (Tweets) as on-line micro-text. Twitter is an on-line social networking site and a micro-text (micro-blogging) service, created in March 2006 by Jack Dorsey and launched in that July 5. The obvious reasons for taking Twitter data for our analysis are:

(i) Twitter is a most popular short message service.
(ii) It is voluminous and versatile collection of real-time data (150 million Tweets are generated per day)
(iii) Despite some Twitter specific terminologies (e.g. @, #), it also possesses very high with other sources of micro-texts e.g. SMS, Instant Chat and contents of other social networking sites.
(iv) In addition, due to the popularity of Twitter, a spontaneous use of Twitter-specific terminologies has been observed in other sources of micro-texts too.

Tweets are the basic atomic building blocks of all things in Twitter limited up to 140 characters [5]. Users Tweets are also known more generically as "status updates." Table 1 provides a quick and brief look up of the terminologies and concepts related to Twitter.

The rest of the paper is organized as follows: Section II, summarizes related work in area sentiment analysis in general; however, some parts of this section are specific to the work done for sentiment analysis of on-line micro-texts. Section III explains the data gathering process. Section IV specifically discusses the challenges associated with the on-line micro-texts. Section V provides result analysis of our observations on on-line micro-texts. Finally, the paper concludes the observations in Section VI.

Table 1.Terminologies/ Concepts used in Twitter

Terminologies/ Concepts used in Twitter	Description
Tweets	Tweets are the basic atomic building blocks of all things in Twitter limited up to 140 characters
@	The @ sign is used to call out usernames in Tweets
#	The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages.
RT	An abbreviated version of "Retweet."
URL Shortener	URL shorteners are used to turn long URLs into shorter URLs
DM	Direct Message

II. Related Work

As discussed in previous sections, due to the rapid growth in Twitter users, the numbers of tweets per day are increasing drastically, and hence sentiment analysis of tweets has gained more attention. This section describes some of the pioneered work done for sentiment analysis of Twitter's data. However, we present related work done throughout the paper.

The sentiment analysis is not new for linguistic researchers, but attention to fine-grained opinion mining has increased after [6-11]. Many classical, statistical and machine learning approaches have described in these literature. An unsupervised approach is proposed in [11] for the classification of movie or product reviews. In this work, semantic orientation score of subjective phrases is calculated using PMI values between predefined seed words and the phrases. OPINE [10] uses web-based search for computing Pointwise Mutual Information (PMI) score between the phrase and meronymy discriminators associated with product class using Equation (1), where f is fact and d is discriminator.

$$PMI(f,d)=\frac{(hits(f\wedge d))}{(hits(d)hits(f))} \tag{1}$$

The OPINE approach is based on [11]. Use of predefined rules and binary grammatical dependencies for sentiment analysis is explored in [1] and [2].

Although, the work of [1, 7-9, 11] are amongst the pioneered work in the area of sentiment analysis, but since these works are mostly carried out on product or movie reviews, hence they do not grab very well for the sentiment analysis of on-line micro-texts. There are various challenges associated with sentiment analysis of on-line micro-texts, which prevent text mining people from adopting these proven approaches directly without major modifications.

Furthermore, due to the popularity of Twitter, in the very last years of 2000 decade, many works were carried out by researchers, which are peculiar to the sentiment analysis of tweets. These research are mostly based on machine learning approaches. A distant supervision-based approach using Na ïve Bayes, SVM and Maximum Entropy used in [12] for classification of tweets in positive and negative sentiments. In this work, emoticons such as ":) and "(:" are used for creating training data using Twitter API and learned data for following combinations: unigrams and bigrams, unigrams and bigrams and part-of-speech (POS) tags. Use of Twitter-specific features like hashtags (#) and emoticons with POS tagged n-gram features is proposed in [12]. In this, the sentiment analysis is carried out in two steps: sentiment classification then polarity classification. Target-dependent sentiment classification of tweets is taken care in [13], this work utilized graph base optimization for associating the target with opinion and does classification using Support Vector Machine (SVM). A method for alleviating data sparsity of Twitter data has proposed in [14] which works for semantic smoothing of tweets by extracting and associating the hidden semantic concept with corresponding tweets and utilized methods such as Maximum Likelihood Estimation (MLE) and Laplace Smoothening for accomplishing the task. MOA-TwitteReader [15] is based on Massive On-line Analysis (MOA) [16] framework. MOA-TwitteReader identifies the changes in the frequency of most used terms and classifies tweets in real time. This system utilizes the concepts of incremental TF-IDF and Adaptive Sliding Window (ADWIN).

III. Data Collection and Tools for Carrying out Observations

We were collected 37,800 using Twitter's Search and REST API [17] between 3:00 PM to 4:00 PM from 1st February to 5th of February 2015 for four very frequent terms related to Indian politics about that time in India, namely, "BJP" (abbreviation for "Bharatiya Janata Party"), "NAMO" (used for "Narendra Modi", Prime Minister of India), "AAP" (abbreviation to "Aam Aadmi Party"), and "Congress" . All tweets were collected in JSON format and parsed for extracting various information to correspond each tweet such as text, hashtags (#), created_at, tweet_id, user_id, user_name, etc. About 1000 tweets are selected randomly from every day's collection in a manner to get a uniform representative of the complete data collection. Finally, we chose 5000 tweets to perform observations with a total of 79,502 tokens.

The observations were conducted by using various popular tools, mainly NLTK [18], Python language [19] and R [20] to analyze challenges involve in the analysis of on-line micro-texts.

IV. Challenges Involved in Analysing On-Line Micro-Texts

This section presents a discussion on challenges involved in processing and analysis of on-line micro-text (especially twitter's data). For this, based on our empirical observations, we have categorized the challenges as depicted in Fig. 1. Some challenges are obvious and can be easily identified. Although, few of these challenges have been already explored in many kinds of literature [21-25] for example use of emoticons and lingoes. However, some challenges are not very obvious and can be explored and justified only by performing intense experiments on the data. Some of these challenges are described in detail in this section.

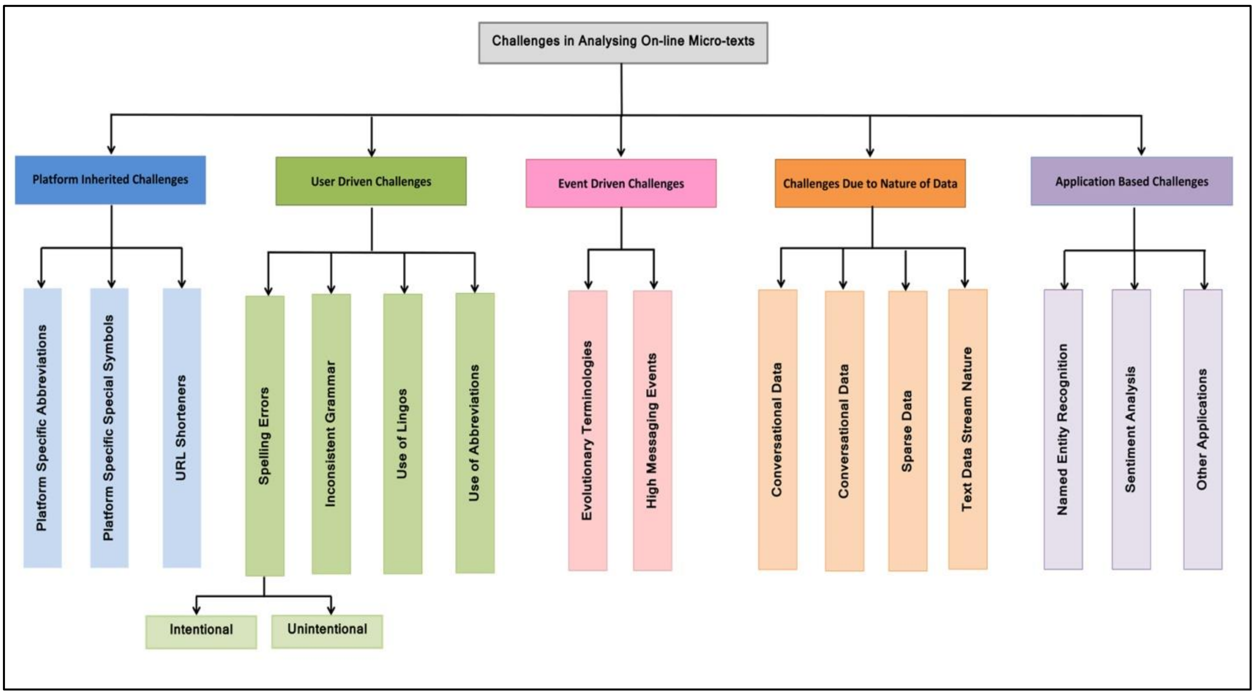
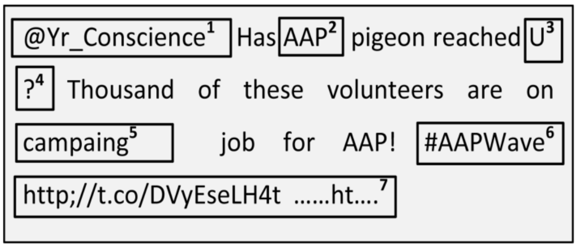


Fig.1. Classification of challenges associated with on-line micro-text

A. Platform Inherited Challenges

Platform Inherited Challenges are generated due to the noises that are introduced in micro-texts by the use of platform specific symbols and abbreviations. These symbols and abbreviations are used for special purposes in micro-texts. A brief description of some twitter specific symbols is already mentioned Table 1. The use of Twitter-specific symbols and concepts are now very frequent in other sources of micro-texts also. Fig. 2 depicts a general construct of a tweet and showing all possible noises by the help of boxes.

Although, the platform specific symbols, concepts, and abbreviations are very informative in many senses (e.g. used in graph-based optimization [13]), however, they are also considered as major sources of noise. The processing of these noisy texts imposes challenges in the syntactic and semantic analysis of the text. Syntactic analysis is of any text is crucial to exploit the grammatical construct of the text.



@Yr_Conscience[1] Has AAP[2] pigeon reached U[3] ?[4] Thousand of these volunteers are on campaing[5] job for AAP! #AAPWave[6] http;//t.co/DVyEseLH4t ......ht....[7]

Fig.2. Example of a tweet: ^{1}@, ^{2} Abbreviation, ^{3} Acronym, ^{4} Special characters, ^{5} Misspelled word, ^{6} Hashtag, ^{7}URL

B. User Driven Challenges

There are various kinds of noises, which are very common in the textual contents scraped from SNSs, blogs, e-commerce sites, and SMS. These noisy texts required intense pre-processing for normalizing them before their actual use. Text normalization is a process of making un-normalized text consistent. Text normalization is very essential for the analysis of on-line micro-texts Some basic operations of text normalization include. converting of all letters to lower or upper case, converting numbers (dates, currencies, temperature) into words, removing punctuation, expanding abbreviations, removing stop words (the, of, to, for, from, etc.) using stop words list, stemming, removing repeating characters (e.g. "I looooove it!" → "I love it!"), lingo and slang resolving (e.g. “l8ter” → “later”). The observation in [24] claimed that 31% of Twitter data and 92% of SMS messages needed normalization. Some basic user-driven noises are described in the remaining part of this section.

1) Unintentional spelling errors

Unintentional spelling errors are those spelling mistakes that are mistakenly made by writers of micro-texts. The spelling error correction problem of on-line micro-texts contains many similarities with the conventional spelling error correction problem. These spelling error correction problems include (i) non-word errors such single-error misspelling and multiple-error misspelling and (ii) real-word errors such as context-dependent spelling error. The single-error misspelling can be further divided into four different classes; Missing of a letter (alphabet = alpabet), Addition of a letter (alphabet = alphabete) and Single transpose (alphabet = alhpabet) and single substitution (alphabet = alphabed). The non-word spelling correction methods are generally based on the spelling lexicon lookup method and the raking of the all possible correct words corresponding to the misspelled word. The ranking methods are further based on certain similarity measures such as Skeleton Levenshtein Distance (Edit Distance) [26] and probabilistic measures [27]. For solving multiple-error misspelling, Khothari et al. [28] utilized string based similarity measure for computing similarity between two strings.

2) Intentional spelling error

In our dataset, we observed many situations where the writer does intentional spelling mistake. One of the intentional spelling mistakes is the repetition of letters e.g. "I looooove it!" → "I love it!”. Writers usually do repetition of words to emphasize his or her utterance and to express an intense feeling about the topic or event. One of the obvious solutions for such noise is the removal of the repeated letters until correct word will be formed. To identify such words, we simply detected more than one repetition of same characters in non-standard words.

3) Lingoes and Slangs

Nowadays the use of lingos (e.g. “gr8” → “great”, “bcos→ because”, “gud→ good”) is very fascinating and intentionally used by writers. However, this kind of writing for a word imposes challenges in the processing of the text. Letter-to-Phoneme (L2P) approach is suggested in many kinds of literature for solving lingos and slangs [29-31]. The CMU Pronouncing Dictionary (CMUDict) [32] is a machine-readable pronunciation dictionary contains over 125,000 words and their transcriptions.

4) Use of Abbreviations and Shortening of Phrases

Due to the limitations of characters in SMS and micro-blogging sites, the use of abbreviation is very common. Since we have collected our data set by making a search on following terms namely “BJP”, “AAP”, “Congress” and “NAMO”, that’s why we have all tweets were having abbreviations BJP (Bhartiya Janta Party) and AAP (Aam Aadmi Party). However, the main concern of our observation is to target those abbreviations that aren’t representing any standard entities like; Institute of Electrical and Electronics Engineers (IEEE), State bank of India (SBI), United State (US), etc., but on those which are shorten form of frequent phrases e.g. ASAP→”As soon as possible”. Generally, corpus lookup is suggested for world-to-phrase mapping.

The abbreviation detection and resolution is a two-phase task. First, is abbreviation detection and the second is abbreviation resolution phase. The most popular method for abbreviation detection is based on certain heuristics such as the continuous occurrence of capitalized characters [33], use of bracket symbols [34]. Other methods include Part-of-Speech tagging, Rule-based, and domain specific corpus etc.

In on-line micro-texts like tweets, the conventional methods of abbreviation are very hard to apply directly. Furthermore, the frequent use of newly evolved terminologies restricts the completeness of any corpus and rule-based method. In addition, the use of inconsistent grammar in on-line micro-texts limits the exploitation of grammatical constructs for abbreviation detection. To conduct our observation, we simply detected abbreviations by utilizing a hybrid method, which includes continuous capitalization detection, bracket detection, matching of the manually crafted list of frequently used abbreviations from internet and tweet tagger [35].

5) *Inconsistent Grammar*

In sentiment analysis (SA), the grammatical consistency of a text plays very important role. A grammatically consistent sentence can be analyzed for various issues in sentiment analysis such as; target and sentiment association, semantic role labeling and identifying grammatical dependencies. Due to the inconsistency in grammar, many classical methods of NLP do not suit in the processing of micro-texts. However, machine learning approaches are doing well. The voluminous availability of data justifies the success of machine learning methods. While using machine learning methods, short of binary dependencies and ternary dependencies are also exploited by learning the data in chunks of bigrams and trigrams e.g. “ not good”, and “extremely very good” [2]. In our observation, we identified that the problem in utilizing the grammatical information or constructs is tough in micro-texts e due to the excessive use of platform-specific conventions e.g. use of hashtags, emoticons, URLs etc.

6) *Excessive use of emoticons*

An emoticon is a pictorial representation of a facial expression e.g. “:)” and “:(“,. Writers usually use it to express their feeling about the topic of discussion. The main reason for using emoticons are their high expressiveness and ease in using them. Of course, the limitation of characters in micro-texts posts is also a reason for the excessive use of emoticons. The use of emoticons also introduces considerable noise in the text. Despite inability in assigning any grammatical binding of the emoticons with any other words in the micro-texts, they still play an important role in the sentiment analysis of a tweet or SMS. In that particular role, these emoticons are used as sentiment polarity labels (i.e. + or - depending on feelings they expressed) of the dataset and hence provide a way of performing supervised learning. We used ark-tweet-tagger [35] for detecting emoticons.

C. Event Driven Issues

There are many events, which are temporal and evolutionary. Sometimes they become so popular that everyone makes a comment on them. Such events also produce a significant impact in the writing style of writers. Since the popular events of the world also attract millions of Twitter users, due which the rate of tweets increased explosively. Such events impose a significant impact on the accuracy and real-time compatibility issues of sentiment analysis as long as on-line sentiment analysis is concerned. We observe such event-driven issues in two categories: (i) Evolutionary Terminologies (e.g. Hashtags) and (ii) High Messaging Events.

1) *Evolutionary Terminologies*

Evolutionary terminologies are those terminologies, which evolve with time and do not belong to any existing dictionary. Such terminologies do not have past evidence. However, these are highly motivated and evolved due to recent occurrence of events or newly developed concepts (e.g. recent version of any operating system). In Twitter posts, these newly evolved terminologies and concepts are represented by hashtags. Officially, these hashtags are originated or constructed by any users [36], however, in many cases, the hashtags are created and offered by some official bearers of the hashtags (e.g. New channels, sports channels, movies, etc.) in a manner to invite comments on it. The sentiment analysis of the tweets which are having official hashtags is quite easy because all comments can be specified by the hashtags. Whereas those hashtags or terminologies that are created by the users are not easy to unify, i.e. many user created terminologies may represent the same event, concept or personality as shown in Fig. 3. In this scenario, it is very often to miss those tweets, which are discussing the same idea or event and having a direct or indirect impact on sentiment analysis of the particular topic under consideration. Hence, it becomes very challenging to do accurate sentiment analysis with nonstandard evolutionary terminologies.

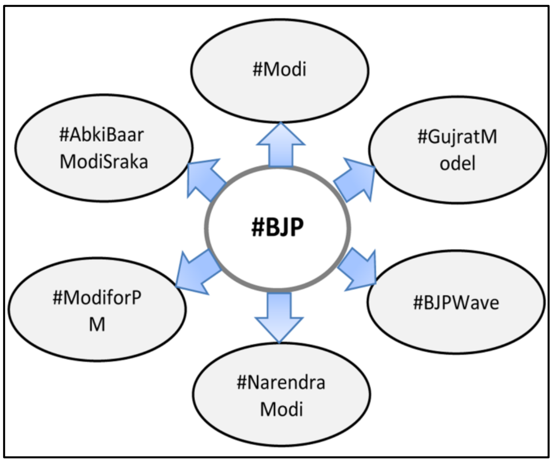


Fig.3. Hashtags representing the same concepts as #BJP

In our data, we get 453 distinct hashtags out of 6371 hashtags. We selected randomly 200 tweets, which are having more than one hashtags including #BJP and manually analyzed that there six more terms or hashtags related to BJP as depicted in Fig. 3. Any comment made for any of these six terms, indirectly associated with BJP hence for accurate sentiment analysis all seven terms including BJP are needed to be analyzed.

2) High Messaging Events

As already mentioned the real-time compatibility is a challenging issue for on-line sentiment analysis for doing decision making in real time. However, there are some events, which become extremely trendy for a very short duration of time (e.g. final of football world cup and Olympics) and draw the attention of millions of users instantly, due to which is not possible to process them in real time. Handling such explosive situations need very fast processing and classification algorithms.

D. Challenges Due to Nature of Data

One of the ways to classify textual data is to do the classification by the genre of the style of writing e.g. story, technical writing, reviews, speech, blogs, short messages and dialog or conversations. Each of these styles of writing introduces a different kind of challenges in the sentiment analysis because all of these styles hold a different amount of consistencies within the text. For example, stories are highly consistent with respect to the short message. The conversation contains a different way of understanding.

1) Conversation and dialogue

In a conversation, a kind of joint participation in which of two or more active participants communicates to each other interactively, however, in the dialogue is the conversation between two participants. It is hard to analyze the conversations and dialogues as compared to stories or other consistent texts.

Many time the On-line micro-texts, mainly Tweets carried a conversational nature due to which it becomes tough to associate topic of conversation as the time evolve. However, due to the recent practice of including hashtags by users reduces the complexity associated for associating topic with the texts.

2) Challenges due to native topics and use of mixed Languages

Undoubtedly, English is one of the most used languages over Twitter; however, many times writers mix their native languages as well as other languages with the tweets. Additionally, many times writers write the comments using transliteration. There are many possibilities of writing in mixed languages. Some combination of multi-language writing is given in Table. 2. The appearance of any foreign scripts makes the micro-texts highly inconsistent. Dealing such issues in real time is still very difficult and have low accuracy [32].

For the countries like India, where there is much diversity in language and culture, it is very challenging to make the micro-texts consistent when there is a frequent use of multiple languages. India includes Hindi as the national language and 29 regional languages (speaking and writing both). This diversity in Indian languages is also reflected in the tweets. Examples in Table. 2 demonstrate this diversity.

Table 2. Use of multiple languages in Tweets

Type	Tweets	Language and remarks
Type 1	Modi wave is every where	English
Type 2	मोदी wave हर जगहं है|	Hindi English mixed (Hinglish)
Type 3	Modi lahar har jagahn hai	Hindi written using English script (Transliteration)
Type 4	मोदी लहर हर जगहं है|	Hindi
Type 5	মোদি ঢেউ সর্বত্র হয়	Bengali (All other regional languages in India)

3) Data Sparsity

Data sparsity is always being a challenge in the training of sentiment classifier. Micro-texts particularly Tweets produce highly sparse data because of the use of evolutionary terminologies and noises. Despite increased memory requirements by the sparse data, it compromises with the information content of the dataset. Sparse data increases the number of unnecessary features in the training dataset. Textual data sparsity can be easily eliminated by using NLP approaches [14] [37], such as utilization of contextual information, use of syntactic information and semantic smoothening [37]. However, due to the excessive noise in the data, it is challenging to make the data dense. The frequency chart in Fig. 4 shows the statistics of term frequency of the complete data collected by us. From the Fig. 4, it can be easily concluded that the micro-texts dataset is a highly sparse data.

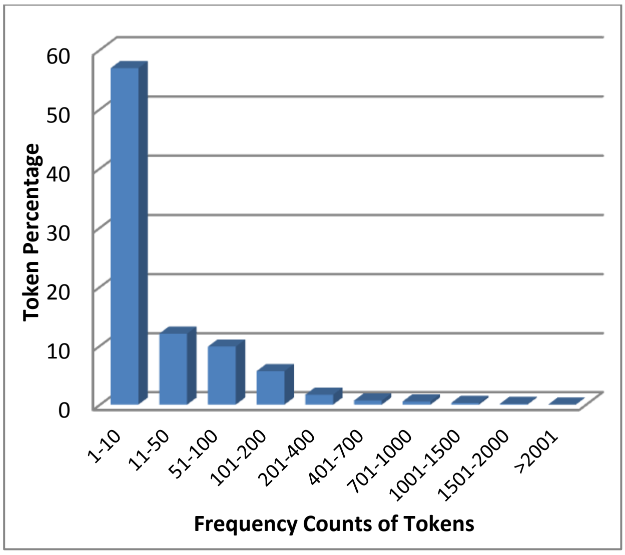


Fig.4. Frequency count of tokens

4) Data stream nature of on-line micro-texts

The on-line micro-texts such as tweets exhibit the characteristics of text data stream. An on-line micro-text can be observed as a vector of words, which are generated continuously with the time. For simplicity, we can denote an on-line micro-text by$(M^P)_t$, which is a P dimensional vector of words $[W_1, W_2, W_3, \dots \dots W_{P-1}, W_p]_t$ generated at time t as shown in Equation (2).

$$\left(M^P\right)_t = \left[W_1, W_2, W_3 \dots\dots W_{P-1}, W_P\right]_t \qquad (2)$$

As compared to the stationary data like stories, the sentiment analysis on text data stream is difficult for various reasons, for example, unlike to static data the data stream is unable to store entirely in memory, due to which it is needed to perform SA in real time. Furthermore, the strength of sentiment for any event or topic may changes with time in text data stream, whereas, the sentiment strength in static data remains persistent.

The necessity of real-time sentiment analysis of on-line micro-texts discourses the use of classical time-consuming NLP methods used so far for sentiment analysis of the static textual data. However, as mentioned earlier, the machine learning approaches are now becoming a dominating approach for sentiment analysis of on-line micro-texts.

E. Application based challenges

Recently, the on-line micro-texts, especially the Twitter data have been utilized in many applications, such as Named Entity Recognition (NER) [38] and Sentiment Analysis (SA) [39]. The conventional NLP tools that were trained on consistent text such as news articles are inferior for the analysis of tweets. The named entity classification in tweets is difficult because of various reasons, for example, there is an excess of distinctively named entity types in tweets, and the infrequent occurrence of related tweets makes the data insufficient for exploiting the context. Furthermore, the capitalizations in tweets are less reliable. Both NRE and SA have their specific necessities; however, the noise in the data is the primary concern for both the applications.

V. Results

The main intention our experiment is to investigate the challenges associated with the sentiment analysis of on-line micro-texts. Since we have considered the tweets as an example of on-line micro-texts, we performed certain observations on the Twitter data to verify the inconsistency and noises present in the tweets, which degrades the performance of conventional NLP tools for the analysis of on-line micro-texts. Table. 3, provides the statistics of the each type of tokens present in the collected data. Column 2 and Column 3 of the table shows the total number of tweets having such tokens and the total number of each type of tokens present in the complete dataset respectively.

Table.3. Types of tokens and their frequency count

Types of Tokens	No. of Tweets having types token as mention in 1st column	Total No. of tokens in all tokens
@	3001	4662
Hashtag (#)	4012	6371
URLs	1066	1135
Misspelled	2984	7823
Abbreviations	1103	2978
Stop words	4816	16891
Slangs & Lingoes	894	1093
Emoticons	1231	1387
Numbers	1134	2334
Other Symbols	2734	1845
Other Languages	611	832
Total	**5000**	**79502**

A. Tokens and Tweets Wise Result Analysis

The percentage of tweets having various types of tokens is depicted in Fig. 5. This section presents tokens and tweets wise description of our empirical observations.

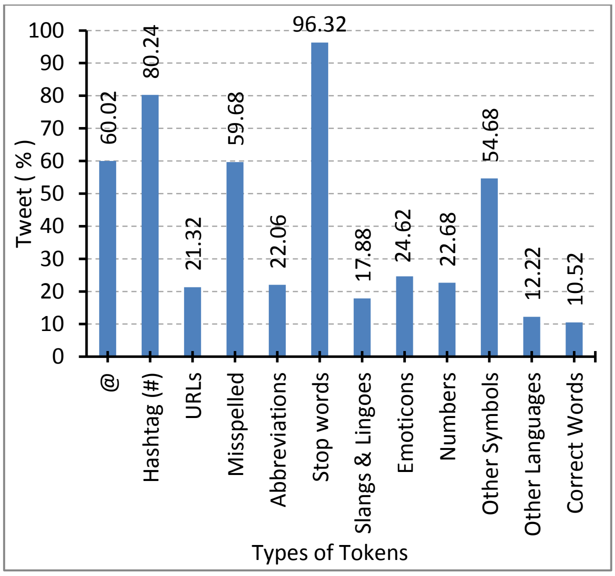


Fig.5. Percentage of tweets having various types of tokens

1) Spelling Errors

In our collected dataset, we identified that 1408 tokens belong to unintentional spelling errors collectively, which is about 1.77% of total tokens and 18.06% of tweets are having this error. Furthermore, we identified 1121 tokens with intentional spelling errors, which is about 1.41% of total tokens and 15.84% of tweets are having this mistake. We also added those errors in spelling errors, which are not identified as slangs, lingoes, abbreviations and phrase shorter forms. By adding them all, we observed that about 59.68 % of total tweets are having spelling errors of these forms as shown in Fig 5. Finally, we observed that 10% of total tokens are spelled wrong as shown in Fig 6.

2) Lingoes and Slangs

In our data set, we identified 1093 tokens as lingoes and slangs, which is about 1.12% of total tokens as shown in Fig. 5. About 17.88% of tweets are having such errors as shown in Fig. 6. Although there may be more such errors but the identified errors are limited to our methods and list to identify these errors. In our experiment, most of unidentified lingoes and slangs were added to the group of misspelled words.

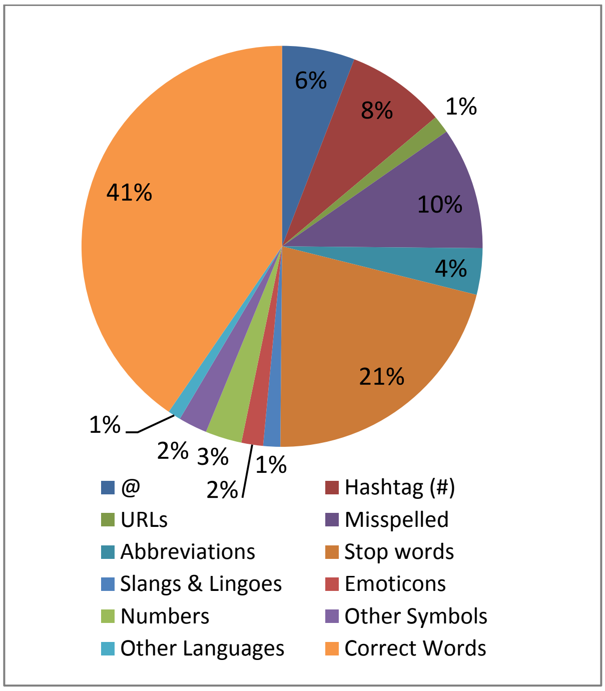


Fig.6. Percentage of various noisy tokens with total tokens

3) Abbreviations and Shortening of Phrases

We observed 435 distinct tokens of abbreviation and shortenings types including our search terms e.g. BJP, AAP and some very popular acronym e.g. MP (Member of Parliament), LS (Lok Sabha), PM (Prime Minister), etc. The repeated occurrences of these distinct tokens summed up to 2978 tokens, which is about 4.0% of total tokens and 22.06% tweets were having abbreviation(s) as shown in Fig 5 and Fig. 6 respectively.

4) Emoticons

In our data set, we identified 1387 emoticons, which is about 2.0% of total tokens as shown in Fig. 6 and 1231 tweets were having emoticons in them, which about 24.62% of total tweets as shown in Fig. 5.

5) Twitter Specific Tokens

In our data, we get 453 distinct hashtags out of total 6371 hashtags. We observed that about 80% of tweets are having hashtags. Furthermore, we get 4662 ‘@’s with 3001 tweets having tokens starting with ‘@’. We get 1066 tweets having URL in them. These observations are depicted in Fig 5 and Fig 6.

6) Stopwords

The excessive presence of stop-words in any textual data makes the data highly sparse, the same case exists with the on-line micro-texts (e.g. tweets). In our collected tweets, we observed the excessive presence of stop-words, which makes the data highly sparse and reduces its credibility in the analysis. In our dataset, we find that about 96.32 % percent of tweets are having stop-words that share the 21% of total tokens.

B. Overall Analysis of Collected Data

As shown in Fig.7, only about 11.0 percent of tweets are not having any noise and hence does not require any cleaning. We have observed that 40 percent of tokens are standard words, and rest tokens are out of vocabulary. A high percentage of erroneous tokens are observed as compared to error free tokens as shown in Fig. 8.

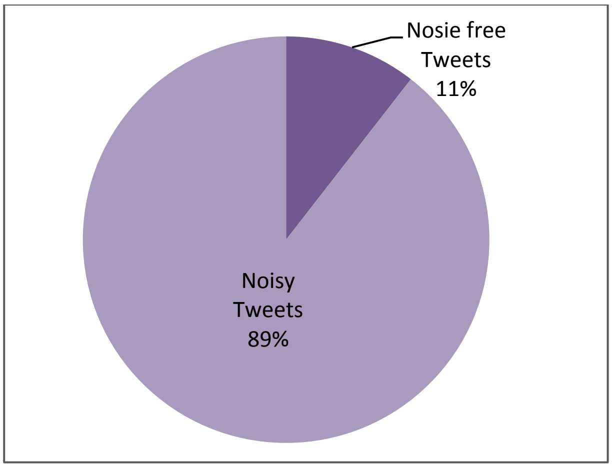


Fig.7. Noisy V/s Noise Free Tweets

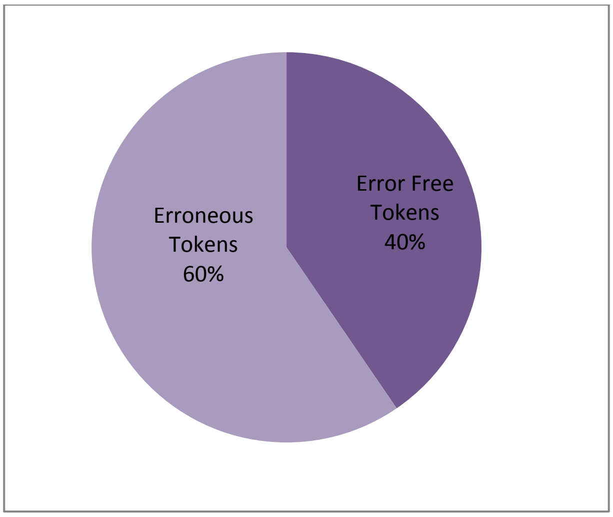


Fig.8. Percentage of erroneous and error free tokens

VI. Conclusion

In this paper, we have investigated various kind of noises associated with the on-line micro-texts, which introduce many challenges in the sentiment analysis of on-line micro-texts. We classified these challenges broadly in five different categories. The results of our observations demonstrate that the on-line micro-texts are highly noisy due to which is difficult to apply NLP tools, which are trained by noise-free training datasets. The classical NLP methods are also not applicable to these data directly without having some specific modifications. Furthermore, classical NLP methods are very time consuming that is why they can be used in the real-time analysis of on-line micro-texts. Recently, many researches have shown a great success of ML approaches over classical NLP approaches for the sentiment analysis of on-line micro-texts.

References

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168-177: ACM.

[2] R. Srivastava, M. Bhatia, H. K. Srivastava, and C. Sahu, "Exploiting grammatical dependencies for fine-grained opinion mining," in *Computer and Communication Technology (ICCCT), 2010 International Conference on*, 2010, pp. 768-775: IEEE.

[3] A. Kumar and M. S. Teeja, "Sentiment analysis: A perspective on its past, present and future," *International Journal of Intelligent Systems and Applications*, vol. 4, no. 10, p. 1, 2012.

[4] B. Narendra, K. U. Sai, G. Rajesh, K. Hemanth, M. C. Teja, and K. D. Kumar, "Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 8, no. 8, p. 66, 2016.

[5] (28/09/2016). *Tweets*. Available: https://dev.twitter.com/overview/api/tweets

[6] A. Fahrni and M. Klenner, "Old wine or warm beer: Target-specific sentiment analysis of adjectives," in *Proc. of the Symposium on Affective Language in Human and Machine, AISB*, 2008, pp. 60-63.

[7] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *AAAI*, 2004, vol. 4, no. 4, pp. 755-760.

[8] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1-167, 2012.

[9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[10] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural language processing and text mining*: Springer, 2007, pp. 9-28.

[11] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 417-424: Association for Computational Linguistics.

[12] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

[13] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 151-160: Association for Computational Linguistics.

[14] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," 2012: CEUR Workshop Proceedings (CEUR-WS. org).

[15] A. Bifet, G. Holmes, and B. Pfahringer, "Moa-tweetreader: real-time analysis in twitter streaming data," in *International Conference on Discovery Science*, 2011, pp. 46-60: Springer.

[16] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1601-1604, 2010.

[17] (28/09/2016). *REST APIs*. Available: https://dev.twitter.com/rest/public

[18] S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*, 2006, pp. 69-72: Association for Computational Linguistics.

[19] G. van Rossum and F. L. Drake, "Python Reference Manual, PythonLabs, Virginia, USA, 2001," *Available online at:(accessed 1 December 2012)*, 2001.

[20] R. C. Team, "R: A language and environment for statistical computing," 2013.

[21] B. Han, P. Cook, and T. Baldwin, "Automatically constructing a normalisation dictionary for microblogs," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 421-432: Association for Computational Linguistics.

[22] B. Han, P. Cook, and T. Baldwin, "Lexical normalization for social media text," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 1, p. 5, 2013.

[23] R. Khoury, R. Khoury, and A. Hamou-Lhadj, "Microtext Processing," in *Encyclopedia of Social Network Analysis and Mining*: Springer, 2014, pp. 894-904.

[24] Z. Xue, D. Yin, B. D. Davison, and B. Davison, "Normalizing Microtext," *Analyzing Microtext*, vol. 11, p. 05, 2011.

[25] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data," in *RANLP*, 2013, pp. 198-206.

[26] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, 1966, vol. 10, p. 707.

[27] K. Toutanova and R. C. Moore, "Pronunciation modeling for improved spelling correction," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 144-151: Association for Computational Linguistics.

[28] G. Kothari, S. Negi, T. A. Faruquie, V. T. Chakaravarthy, and L. V. Subramaniam, "SMS based interface for FAQ retrieval," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 2009, pp. 852-860: Association for Computational Linguistics.

[29] S. Jiampojamarn, C. Cherry, and G. Kondrak, "Joint Processing and Discriminative Training for Letter-to-Phoneme Conversion," in *ACL*, 2008, pp. 905-913.

[30] T. Rama, A. K. Singh, and S. Kolachina, "Modeling letter-to-phoneme conversion as a phrase based statistical machine translation problem with minimum error rate training," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, 2009, pp. 90-95: Association for Computational Linguistics.

[31] A. Van Den Bosch and S. Canisius, "Improved morpho-phonological sequence processing with constraint satisfaction inference," in *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology*, 2006, pp. 41-49: Association for Computational Linguistics.

[32] M. S. Stinson, S. Eisenberg, C. Horn, J. Larson, H. Levitt, and R. Stuckless, "Real-time speech-to-text services," *Reports of the National Task Force on Quality Services in Postsecondary Education of Deaf and Hard of Hearing Students. Rochester, NY: Northeast Technical Assistance Center, Rochester Institute of Technology,* 1999.

[33] K. Taghva and J. Gilbreth, "Recognizing acronyms and their definitions," *International Journal on Document Analysis and Recognition,* vol. 1, no. 4, pp. 191-198, 1999.

[34] Y. Park and R. J. Byrd, "Hybrid text mining for finding abbreviations and their definitions," in *Proceedings of the 2001 conference on empirical methods in natural language processing,* 2001, pp. 126-133.

[35] K. Gimpel *et al.*, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 2011, pp. 42-47: Association for Computational Linguistics.

[36] *The Twitter glossary*. Available: https://support.twitter.com/articles/166337

[37] H. Saif, Y. He, and H. Alani, "Semantic smoothing for twitter sentiment analysis," 2011.

[38] A. Ritter, S. Clark, and O. Etzioni, "Named entity recognition in tweets: an experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1524-1534: Association for Computational Linguistics.

[39] R. Srivastava and M. Bhatia, "Quantifying modified opinion strength: A fuzzy inference system for Sentiment Analysis," in *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, 2013, pp. 1512-1519: IEEE.

Authors' Profiles

Mr. Ritesh Srivastava obtained his B.E. degree from Agra University, Agra, India and M.Tech degree from Netaji Subhas Institute of Technology (NSIT), University of Delhi (DU), New Delhi in Computer Science & Engineering. Currently, he is pursuing Ph.D. from NSIT, (University of Delhi), New Delhi. He has about ten years of teaching & research experience. His research areas include Machine Learning (ML), Data Stream Mining, Text Mining, Information Retrieval (IR) and Natural Language Processing (NLP). He is a member of various professional bodies and SIGs like IEEE, ACM, ACM-SIGKDD, Big Data Community.

Dr. M.P.S. Bhatia received his Ph.D. in Computer Science from the University of Delhi. Dr. Bhatia is a Professor in the Division of COE at the Netaji Subhas Institute of Technology, affiliated to the University of Delhi. He is also serving the Institute as Dean, Student Welfare and Head, Placement Cell. He has guided many M.Tech and Ph.D. students in their research work. His research interests include data mining, cyber security, semantic web, machine learning, social network analysis and sentiment analysis. He is an author or co-author of many research papers in international journals and conferences. Dr. Bhatia is a member of IEEE (Institute of Electrical and Electronics Engineers) and CSI (Computer Society of India).

How to cite this paper: Ritesh Srivastava, M.P.S. Bhatia,"Challenges with Sentiment Analysis of On-line Micro-texts", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.7, pp.31-40, 2017. DOI: 10.5815/ijisa.2017.07.04