A Supervised Approach to Analyse and Simplify Micro-texts

Vaibhav Chaturvedi, Arunangshu Pramanik, Sheersendu Ghosh, Priyanka Bhadury and Anupam Mondal

Abstract Analysis of micro-text presents a challenging task due to the incompleteness of its corpus in the domain of Natural Language Processing (NLP). Primarily, micro-text refers to the limited textual content in the form of letters and words, collected from various web-based resources. In the present paper, we are motivated to build a supervised model for analysing micro-text. The model assists in simplifying the texts and extracting the important knowledge from unstructured corpora. Additionally, we have prepared an experimental dataset to validate the proposed model. The validation process offers 94% accuracy to identify the micro-text from the unstructured corpus like Twitter. The proposed model helps to design various applications such as annotation system and prediction system for micro-texts in the future.

Keywords Micro-text · Machine learning · Natural language processing · Dependency matching

V. Chaturvedi · A. Pramanik (✉) · S. Ghosh · P. Bhadury · A. Mondal
Department of Computer Science & Engineering, Institute of Engineering & Management, Y-12, Salt Lake Electronics Complex, Sector–V, Kolkata 700091, India
e-mail: arunangshu01@gmail.com

V. Chaturvedi
e-mail: vaibhav1579@hotmail.com

S. Ghosh
e-mail: shibrata15@gmail.com

P. Bhadury
e-mail: priyankabhadury5@gmail.com

A. Mondal
e-mail: anupam.mondal@iemcal.com

J. K. Mandal and D. Bhattacharya (eds.), *Emerging Technology in Modelling and Graphics*, Advances in Intelligent Systems and Computing 937,
https://doi.org/10.1007/978-981-13-7403-6_8

1 Introduction

Micro-text refers to a microfilmed or micro-photographed[1] textual information. This information is widely availed by several users in various social media platforms to express their emotions [1, 2]. The micro-text analysis task is challenging due to the presentation of texts in an impromptu manner. This presentation also introduced other difficulties such as identification of knowledge-based information and proper meaning recognition from unstructured corpora. Primarily, the users often make use of a various form of micro-texts at the time of communication through social media platforms like Facebook, Twitter, YouTube, Google+, WhatsApp, Instagram, and LinkedIn [3]. Also, they are preferred to use abbreviation of texts and other SMS languages patterns to convey their sentiments or emotions very fluently to other [4, 5]. We have observed that the micro-texts are generated by digital short messages between the range of 2–700 characters along with unconventional grammar and style [3]. It is also hybridisation between informal and traditional expressions with threading characteristics. Micro-text is sufficiently divergent for a progenitor to necessitate unique study. The researchers have observed that the traditional long-text techniques do not translate well to micro-text due to unstructured and semi-structured nature of the corpus. Additionally, the micro-text corpora contain a minute-level time-stamp and a source attribution.

In this paper, we have developed an analysis system for simplifying the micro-text. The simplified form assists in converting the structural corpus from the unstructured and semi-structured corpora. In order to build this system, we have employed various well-known machine learning classifiers such as Logistic Regression, Decision Tree, and Support Vector Classifier (SVC) on the top of the prepared experimental dataset. Finally, we have validated the proposed system using test dataset as a part of experimental dataset, which provides an adequate output. The output may assist in designing various domain-specific applications such as annotation, relationship extraction, and concept-network systems in the future [6, 7].

The rest of the paper presents a detail background study related to micro-text analysis in Sect. 2. Sections 3 and 4 describe the proposed system and the result analysis of the mentioned system in details. Finally, Sect. 5 illustrates the concluding remarks and future scopes in this research.

2 Related Work

Universal in the present world, micro-text has refined the way of communication with an effortless technique. It redefines new defiances for Natural Language Processing (NLP) tools, which are usually designed for well-written text. In some cases, authors present a novel supervised method to translate Chinese abbreviations, which extract the relation between a full-form phrase and its abbreviations from monolingual cor-

[1] https://www.merriam-webstar.com/.

pora, and induce translation entries for the abbreviation by using its full-form as a bridge. Other works based on this are Topic Detection in IRC chat-rooms which follows the approach of TF/IDF with temporal augmentation which impacted to the achievement of 71.5% accuracy on a system modelled to detect the speaker's intent to flirt using a spoken corpus of speed dates [1].

In order to develop an analysis system for micro-texts, the researchers have applied primarily two different types of approaches, namely lexicon-based and machine learning-oriented approaches [8, 9]. Additionally, they have introduced sentiment analysis and phonetic-based approach to analyse the micro-texts [8, 10, 11].

Generally, it follows a dictionary-based approach where acronyms and emoticons are found and extracted from various online sources which ended up forming a table of 1727+ acronyms and 512+ emoticons. The task performed for this proposed model follows the same polarity classification as used for sentiment analysis. The use of POS-tags is more frequent in subjective texts which can be hypothesised by dependency types [8].

The use of SO-dictionaries was challenged to be dogmatic. Hence, the researchers have employed supervised and unsupervised learning modules after extracting various features like parts of speech (POS) and sentiment words, etc. to build and validate the micro-text analysis systems [4].

Due to the origin of human life on Earth, people have been considered as social animals exploitable to opinions as practically all vows and conducts are influenced by them. Generally, when decisions are to be made, individuals and organisations frequently go for other's perspections. Perspections, opinions and its associated concepts like sentiments, emotions, attitudes, etiquettes, evaluations comprise of the sentiment analysis. Gradual upsurge of Web 2.0, people express their views upon various matters and on certain issues. The economic benefits from this can be derived from the knowledge are pretty decent that the market has proposed solutions for analysis of these views. Sentiment analysis is a branch of effecting computed research that tends to assimilate the text into either positive, negative, neutral or mixed expressions. The task required in this field is polarity classification which determines the above objective. No standardisation is followed about polarity categories, but the results give analysis of binary or ternary classification. The task has been incorporated from two different perspectives supervised machine learning (ML) approaches and non-supervised semantic-based methods. Statistical approaches have proved to be subdued as statistical text classifiers only work with adequate precision when given a satisfactorily vast input text. Concept-level sentiment analysis deals with large semantic text analysis in scientific community as well as the business world [12].

Common textual languages in phones generally SMS languages have made more significance on the monotony of a common man's life. Re-imagining and reconstructing a large word into consequential formatted small words have made work and communication more likeable and lovable. The phonetic-based approach follows a simple but tactful algorithm to manipulate micro-text. Soundex is the most famous algorithm which is used effectively to group similar sounding letters together, and

each group unit is assigned to a letter of the numeric characters. The main objective is to use homophones for encoding [8].

The provided background assists in developing the proposed system for micro-texts, which is described in the following section.

3 Methodology

In order to build the proposed system for micro-text analysis, we have initially prepared an experimental dataset which has been collected from Twitter.[2] This experimental dataset has been processed through three different types of classifiers such as Logistics Regression, Decision Tree, and Support Vector Classifier (SVC) for improving the accuracy of the proposed simplification system [13, 14]. In the following subsections, we describe, (a) how we have prepared the dataset? (b) selection process of machine learning classifiers, and (c) the design steps of the proposed model in details.

3.1 Dataset Preparation

Initially, we have collected a dataset from a Twitter Repository,[3] contains around 2500 textual tweets. Besides, we have preprocessed the crawled tweets and identified a number of 1000 tweets as an experimental dataset. All the tweets of the experimental dataset have been satisfied the length of the micro-text as 2–120 characters. These tweets have been labelled manually by a group of Internet users to assign the tweets as a micro-text or general tweets. Thereafter, we have split the experimental dataset into two parts such as training and test dataset. The training dataset helps to learn the classifiers whereas test dataset assists in validating the proposed system. Training dataset contains 800 number of tweets where rest of 200 tweets presented as a test dataset. In the following subsection, we have discussed the classifier selection approach.

3.2 Classifier Selection

Logistic regression is a classified algorithm mainly used for Machine Learning Statistical Analysis. It constructs a statistical model to apply a binary dependent variable. The variable contains a coded data as 1 which indicates a success or 0 for failure. The

[2] https://twitter.github.io/.

[3] https://twitter.github.io/.

algorithm is generally used for the prediction of the probability of the variable. In simple words, the model generates a variable say *P* as a function of another dependent variable *X*.

Decision tree is a one-dimensional regression analysis which is used to place a sine curve in accordance with and addition of a noisy observation. If the maximum depth of the tree is plotted to be high enough, the regression impacts to learn fine details of the trained data and from the incurred noise. It breaks the dataset into tiny datasets whereas at the same time, the decision tree is hierarchically formed.

Support vector classifiers, commonly known as support vector machines or networks, are trained supervised learning models with algorithms in association that examines data used for classification as well as regression. However, to use SVM for analysis of sparse data predictions, it must fit the dimensions properly.

3.3 Proposed Algorithm

Thereafter, we have applied the following algorithm to identify the micro-text from the unstructured corpus.

Step-1: Initially, we have collected a dataset from the Twitter repository and preprocessed them.

Step-2: Prepared an experimental dataset after manually labeled (L) the crawled micro-texts tweets.

Step-3: Extract various features like capital words, alphanumeric characters, etc for the experimental dataset in the form (X) = X_1, X_2, ... , X_n.

Step-4: The prepared experimental dataset is split into two parts as training and testing datasets.

Step-5: The extracted features (X) and their corresponding label (L) of the training dataset have been processed through three different classifiers namely Logistic Regression (M_{LR}), Decision Tree (M_{DT}), and SVC (M_{SVC}) to build the model.

Step-6: Thereafter, we have merged these classifiers M_{LR}, M_{DT}, and M_{SVC} with the help of Equation 1 to build another approach (M_{Merged}) under the proposed module.

$$M_{Merged} = M_{LR} \cup M_{DT} \cup M_{SVC} \quad (1)$$

Step-7: Finally, the test dataset has been applied on the above classifiers to validate the proposed micro-text analysis system.

In the following section, we have discussed the validation process and obtained output for the proposed system.

4 Estimated Results

Besides, to validate the proposed system, we have employed the test dataset and processed through all the classifiers individually. The accuracy of these classifiers has been measured using standard evaluation matrices like precision, recall, and F-Measure. Table 1 presents a comparative analysis between all the mentioned classifiers to the process of designing micro-text analysis system.

Additionally, we have generated the confusion matrix for all classifiers. Table 2 shows a confusion matrix for Logistic Regression Classifier.

The result shows that the combined classifier provides an adequate output for the proposed system. We have also observed that the combined classifier offers 94% accuracy to identify the micro-text from the corpus.

Table 1 An evaluation of the proposed system using precision, recall, and F-measure for all classifiers

Classifiers	Precision	Recall	F-measure
M_{LR}	0.948	0.929	0.939
M_{DT}	0.921	0.909	0.915
M_{SVC}	0.933	0.921	0.927
M_{Merged}	0.954	0.942	0.948

Table 2 A confusion matrix representation for identifying micro-text using Logistic Regression Classifier

Samples		Predicted		Total
		Micro-text	General-texts	
Original	Micro-text	146	8	154
	General-texts	11	35	46
Total		157	43	

5 Conclusion and Future Scopes

This paper aims towards deciphering information from the micro-text, widely used in different social media platforms. Here, we have adopted the technique of Machine Learning and applied it to the abbreviated texts and the various small phrases containing alphanumeric characters to interpret them correctly. The concepts of Logistic Regression, Decision Tree, and SVC have been applied to achieve adequate accuracy for simplifying the micro-texts. The simplified micro-texts may assist in designing various social media applications as a platform to interact and thus it is gaining its importance in market understanding where skilful and strategic planning is required. It can thus be predicted that studies and enhances development on micro-text analysis on big data platform for customer relation management and various other aspects are going to get increasing attention in near future.

References

1. J. Ellen, All about microtext-a working definition and a survey of current micro-text research within artificial intelligence and natural language processing, in *ICAART 2011*, vol. 1 (2011), pp. 329–336
2. J. Young, C.H. Martell, P. Anand, P. Ortiz, H.T. Gilbert IV, A microtext corpus for persuasion detection in dialog, in *Analyzing Microtext* (2011)
3. J.C. Mallery, Semantic content analysis: a new methodology for the RELATUS natural language environment, in *Artificial Intelligence and International Politics* (1991), pp. 347–385
4. D. Vilares, Sentiment analysis for reviews and microtexts based on lexico syntactic knowledge, in Proceedings of 5th BCS-IRSG symposium on future directions in information access, pp. 38–43. 2013
5. P. Smith, M.G. Lee, A CCG-based approach to fine-grained sentiment analysis in microtext, in *AAAI Spring Symposium: Analyzing Microtext*, vol. 13 (2013), p. 1
6. A. Mondal, E. Cambria, D. Das, S. Bandyopadhyay, Mediconceptnet: an affinity score based medical concept network, in *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017* (pp. 22–24)
7. A. Mondal, E. Cambria, D. Das, A. Hussain, S. Bandyopadhyay, Relation extraction of medical concepts using categorization and sentiment analysis. Cogn. Comput. 1–16 (2018)
8. R. Satapathy, C. Guerreiro, I. Chaturvedi, E. Cambria, Phonetic-based microtext normalization for twitter sentiment analysis, in *2017 IEEE International Conference on Data Mining Worshops (ICDMW)* (IEEE, 2017), pp. 407–413
9. E. Bejek, P. Strank, P. Pecina. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures, in *Proceedings of the 9th Workshop on Multiword Expressions* (2013), pp. 106–115
10. A. Mondal, R. Satapathy, D. Das, S. Bandyopadhyay, A hybrid approach based sentiment extraction from medical context, in *SAAIP@ IJCAI*, vol. 1619 (2016), pp. 35–40
11. A. Mondal, E. Cambria, D. Das S. Bandyopadhyay, Employing sentiment-based affinity and gravity scores to identify relations of medical concepts, in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Honolulu, HI (2017), pp. 1–7
12. G. Garzone, Textual analysis and interpreting research (2000)
13. Matthew Shardlow, A survey of automated text simplification. Int. J. Adv. Comput. Sci. Appl. **4**(1), 58–70 (2014)
14. A. Siddharthan, Syntactic simplification and text cohesion. Res. Lang. Comput. **4**(1), 77–109 (2006)