JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Semantic and sentiment analysis of short text data

Mr. Mithun M. Patil, Student, Department of Computer Engineering, TEC, Nerul
Mr. Vishwajit B. Gaikwad Professor, Department of Computer Engineering, TEC, Nerul

***Abstract*- Short textual data available online in the form of social media posts, product reviews, customer queries, search engine queries contain huge source of information, and extracting required knowledge out of it is valuable to business. However there are various challenges in analyzing such unstructured data. Shorts text such as social networking comments, customer reviews are usually grammatically incorrect, lacks sufficient statistical information to support many state-of-the-art approaches for text mining such as topic models. Further, they are more ambiguous with misspelled words and are generated in an enormous volume, which further increases the difficulty to handle them. In order to infer the actual meaning of short text it is essential to have semantic knowledge. After studying multiple methods proposed recently in the field of text analysis, this work has proposed a prototype system that uses sequential neural network for text processing. It divides the task into three subtasks as text segmentation, type detection and sentiment analysis. The proposed system has offline and online modules. In offline module, KERAS sequential model is build using twitter data to calculate affinity score and detect POS tags. Online model initially cleanse the data by filtering English stop words, followed by stemming using Snowball Stemmer and finally correct misspelled words using SpellChecker. The preprocessed data is then segmented using bi-gram and feed to KERAS sequential model to derive the semantic and sentimental knowledge of input text. Proposed method is tested on airline data and results are compared with some state-of-art methods. The analysis shows proposed method is better or equally effective compared to other methods.**

***Keywords*- POS tagging, bi-gram, KERAS Sequential model.**

I. INTRODUCTION

In today's world everyone is connected through internet as the smart phones and internet have become so chip that everyone could afford it. With the increasing use of internet, the amount of data generated in the form of short text is enormous. The sources are chatting applications (Whatsapp, Telegram etc), social networking sites (Face book, Twitter, Instagram), online customer reviews, internet blogs, search engine queries, news titles etc. Analyzing and accurately deriving the useful information is very crucial to many business applications such as e-commerce, spam filtering, web search engines, chatbots etc. However analyzing short texts is very difficult and challenging task. Primary reason is short texts usually do not follow language grammar. It may be ambiguous, noisy with random word placement. Further in today's fast-changing world it is generated in huge volume which makes it complex and time consuming to perform tasks such as information extraction, clustering and classification. These challenges give rise to significant amount of ambiguity and make it extremely difficult to handle and process available data. Many existing text analytics algorithms are inefficient for analyzing short texts mainly due to lack of sufficient statistical information. Consider polysemy of word "apple". It has different meanings such as a fruit, a tree, a company or a brand. Due to the lack of contextual information, these ambiguous words make it extremely hard for computer to understand short texts.

Typically, there are three phases in understanding short text: segmentation, type detection and sentiment analysis. Text segmentation is to break input text into smaller terms which can a word or a phrase. Type detection is to attach meaningful type to each term within input text. Generally POS taggers determine lexical types based on grammatical rules. These approaches are inapplicable as the short texts usually don't follow grammar and lacks sufficient contextual information. Further traditional POS tagging methods cannot distinguish semantic types which, however, are very important for sentiment analysis. In instance disambiguation meaningful labels or types are assigned to each term. These concepts are derived from domain ontology. Sentiment analysis, also referred to as opinion mining, is an approach to NLP that identifies the emotional tone behind a body of text. It helps organizations to determine and categorize opinions about their products, services, and ideas. Organization can use Sentiment analysis to gather insights from complex and unstructured data that comes from online sources such as customer reviews, emails, blog posts, support tickets, web chats, social media channels, forums and comments. In addition to identifying sentiment, opinion mining can extract the polarity or the amount of positivity and negativity within the text. Furthermore it can be applied to varying scopes such as document, paragraph, sentence and sub-sentence levels.

Although the three steps for short text understanding looks straight forward there are many challenges and new approaches must be introduced to tackle these challenges. Short texts are usually noisy, informal and error-prone. It contains abbreviations, nicknames, misspellings etc. For example, "New York city" is sometimes referred as "nyc". This calls for the vocabulary to incorporate as much information about abbreviations and nicknames as possible. Meanwhile, extracting approximate terms is also required to handle spelling mistakes in short texts. Next challenge is ambiguous type where a term can belong to several types, and its best type in a short text depends on context semantics. For example, "watch" in "watch price" refers to wrist watch and should be labeled as instance, whereas in text "watch movie", it is a verb. Short texts are generated in a large volume as compared to whole documents. For example, latest statistics indicate Google now processes over 40,000 search queries every second on average, which translates to over 8.5 billion searches per day and around 3 trillion searches per year worldwide. Twitter is generating around 6000 tweets every second which corresponds to 500 million tweets per day. Therefore, a feasible method for short text processing should be able to handle short texts more effectively and efficiently. However, a short text can have multiple possible segmentations, a term can be labeled with multiple types, and an instance can refer to hundreds of concepts. Hence, it is extremely difficult and time consuming to eliminate these complexities and achieve the best semantic interpretation for a short text.

II. LITERATURE SURVEY

The present work in text analysis is mainly focused on tokenization which spits input text into set of terms or tokens and assigns part-of-speech tags [1][2][3][4][5][6]. For text segmentation various vocabulary based approaches [2][3][4] are proposed which are using online knowledge bases

and dictionaries to extract terms. Longest cover method is one of the vocabulary based method which searches for longest matching term in dictionary to segment input text. Chen et al [2] proposed sentiment analysis of twitter data using an unsupervised method of named entity recognition (NER) which utilizes Wikipedia and web corpus for segmentation. Hua et al. proposed trie-based framework [1] which uses graph to represent terms which are candidate for segmentation and their relationship. One of the common drawbacks of existing methods for text segmentation is they only consider lexical features and ignore the semantics within the segmentation. Statistical methods for segmentation calculate occurrences of two terms together in corpus. N-gram [2][3] is one of the statistical model which calculates frequencies of two or more words occurring together in corpus to decide whether those words can be treated as a term. Semantic hashing [4] is another approach which represents text into binary code which is then used for clustering. However for short texts such approach can yield incorrect information sometimes because of noisy nature.

In Part-of-speech tagging appropriate lexical types are assigned to individual terms based on their meaning and context. It can be done using grammatical methods which uses predefined rules or statistical approaches [1] which uses models trained on large corpus. Rule based approach incurs high cost of constructing production rules however gives stable results. Whereas statistical models use learned statistics instead of tagging rules to assign tags, here the results are unstable. Both the approaches assume that terms are correctly arranged in given input which may not be the case in short text. Song et al. [5] proposed a method of short text understanding by using a popular knowledge base Probase [5] for getting real world concepts and uses Bayesian inference for building words concept vector. However most of the existing knowledge bases are limited in scale and scope. Further most of them do not consider content semantics.

Ming et al [7] proposed a long short term (LSTM) based recurrent neural network model which recognize text emotion by deriving two word vectors semantic and emotional. It can detect seven distinct emotions categories as anger, anxiety, boredom, happiness, sadness, disgust and surprise. LSTM models overcome the drawback of traditional recurrent neural networks that is can't learn long distance dependent information. Jin et al [8] proposed bag of words model to process short texts for duplicate detection. It has used Word2vec to derive word vectors and Simhash algorithm is used to compare sequences using hamming distance.

TABLE 1: Comparison table of various methods proposed for short text understanding

SR No.	Author	Method	Outcome	Limitations	Year
1	Y. Song, H. Wang, Z. Wang, H. Li, W. Chen [5]	Short text conceptualization using Probase	It finds named entities in input text and assigns it meaningful labels.	Applicability is limited. Does not focus on word semantics.	July 2011
2	C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, [2]	Named entity recognition in targeted Twitter stream	It uses Wikipedia and web corpus to segment tweets using n-gram method and does NER using unsupervised model	Applicability is restricted to tweets only.	August 2012
3	Z. Yu, H. Wang, X. Lin and M. Wang [4]	Short text clustering using Deep Neural Network	It uses DNN to convert input text into binary code and then two texts are compared using their binary code to cluster similar texts.	Heavy computation is required. Model output totally depends upon quality of testing set.	Feb 2016
4	W. Hua, Z. Wang, H. Wang, K. Zheng and X. Zhou [1]	Semantic analysis of short text using online knowledgebase and web corpus	Segmentation is done by applying Monte Carlo algorithm on term graph, followed by POS tagging using Standford tagger	Totally dependent upon online knowledge base. Computation of co-occurrence network is very complex and need much time and space.	March 2017
5	M. Su, C. Wu, K. Huang, Q. Hong [7]	Text emotion recognition based on word vector	It first extract semantic word vector and emotional word vector using word2vec model and auto encoder respectively. The concatenated vector is analyzed using LSTM to recognize text emotion.	Results are derived from limited amount data. Need to improve accuracy.	May 2018
6	J.Yang, G. Huang, B. Cai [10]	Short text clustering using TRTD	Topic representative terms are discovered by individual occurrence frequency and co-occurrence frequency	Simple yet effective method of text clustering. Discover topic terms based on high frequency count which may not be the case always.	July 2019
7	R.Man, K.Lin [11]	Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network	Article feature extraction using BERL and Convolutional neural network	Heavy computation is required. Model output totally depends upon quality of testing set.	April 2021

III. PROBLEM STATEMENT

Analyzing textual data available online in multiple forms is crucial to many e-commerce and other business processes. It is important for organizations to accurately analyze their customer reviews, social networking posts, news and chatbots queries which are in the form of short texts, in order to better understand customer needs and gain competitive advantage. However processing this unstructured, complex and huge data is highly challenging task as it lacks contextual information. Further short texts are noisy, may contain abbreviations and ambiguous words which makes it extremely difficult to infer semantic meaning out of it. As a result, traditional natural language processing tools, such as part-of-speech tagging, dependency parsing cannot be applied efficiently to short texts.

Given a short text s written in a natural language, we generate a semantic interpretation of s represented as a sequence of typed-terms namely

$\overline{s} = \{\overline{t_i} | i = 1, \dots, n\}$

And from semantic knowledge we determine the sentiment which can be positive, negative or neutral of input text.

E.g.

Input sentence: "went bank to deposit money"

Output: Went [verb], bank [noun –financial institution], deposit [verb], money [noun]

Sentiment: Neutral

IV. PROPOSED SYSTEM

Many approaches have been proposed recently to enable short text understanding. These methods, however, have their own limitations due to limited context available. Without knowing the word semantics and distribution of the senses, it is difficult to build a model to choose appropriate semantic tag for a word in a context. The proposed method is based on the work conducted by Wen et al. [1] and Zhongyuan et al. [6], which represent input text as a combination of concepts and instances. The system captures semantic relatedness between instances using a probabilistic topic model LDA, and disambiguates instances based on related instances. In this work, it is observe that other terms, such as verbs, adjectives, and attributes, can also help with instance disambiguation. It has also taken into consideration the challenges of misspelled words and short forms that mainly exists in short textual data. By resolving existing challenges efforts are made in proposed system to reduce errors and yield best possible results.

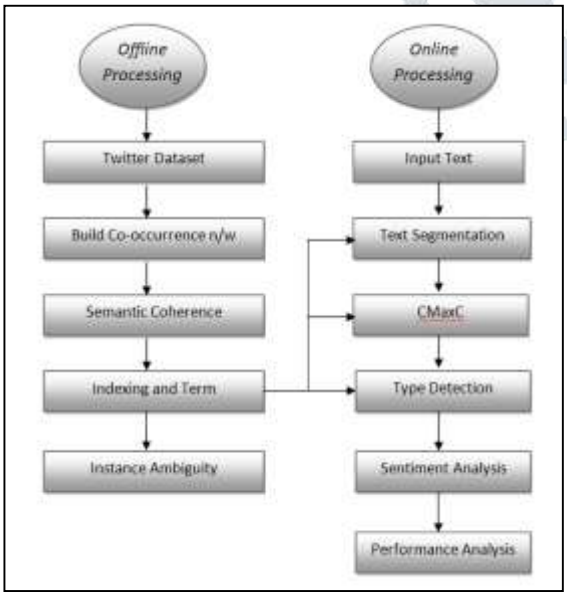

Fig 1: Flow diagram of short text analyzing method

Fig. 1 depicts the working model of proposed system. In offline processing, twitter dataset with 400,000 tweets is used. Initially twitter data is pre-processed by removing English stop words and applying Snowball stemmer. Bigrams are derived from the cleansed data and most frequently words are grouped together forming co-occurrence network. In semantic coherence, dataset is break into training set containing 320,000 tweets and testing set containing 80,000 tweets. The vocabulary is built based on distinct words in dataset using Word2Vec model. In indexing and term extraction, keras tokenizer is used to build vocabulary index based on word frequency which is used subsequently to transform each text in training and testing set into a sequence of integers. In final step the Label Encoder is used to encode labels in training set into number and the KERAS sequential model is built.

In online part, input text is first processed using python Spell Checker to correct misspelled words in text. For text segmentation, python tokenizer is used which first breaks the text into set of terms and converts each term into numeric sequence for further processing. Offline KERAS model is then used to calculate the affinity score among terms by using the sequence vector. The sentiment of input text is then determined using affinity score. Word sense disambiguation is done using Simplified Lesk algorithm. Here the idea is words in a given input text will have a similar meaning, the correct meaning of each word context is found by getting the sense which overlaps the most among the given context and its dictionary meaning.

V. Result Analysis

The performance of proposed method is evaluated via Accuracy, Precision, Recall & F1 Score metrics. Below results analysis shows the effectiveness of sentiment analysis by multiple methods namely Textblob Analysis, VADAR analysis, SentiwordNet Analysis and proposed method based on KERAS sequential model using LSTM.

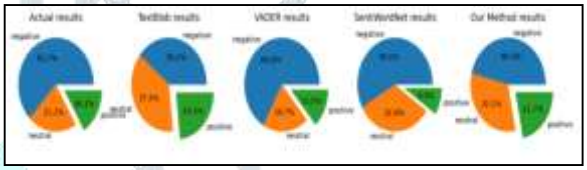

Fig 2: Comparison of Textblob, VADAR analysis, SWN analysis & proposed method based on KERAS model

Airline dataset with 14640 customer reviews has been taken for results analysis. It contains reviews along with labeled sentiment for each customer review. Fig. 2 shows the sentiment count predicted by various methods on twitter dataset. We have compared our method with Textblob, VADAR and SentiwordNet methods. Three labels namely positive, negative and neutral are used to denote the sentiment of input text. Results are compared using metrics Precision, Recall, F1-score and Accuracy. Below are comparison graphs of each metric.

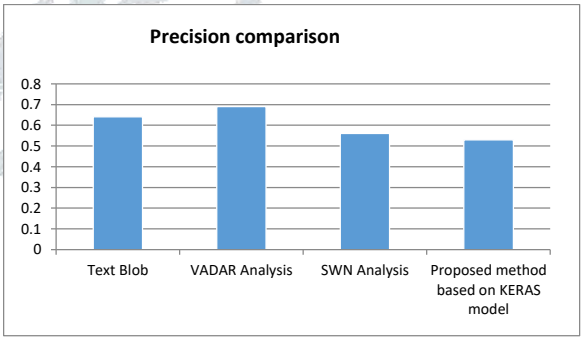

Fig 3: Precision comparison of Textblob, VADAR analysis, SWN analysis & proposed method based on KERAS model

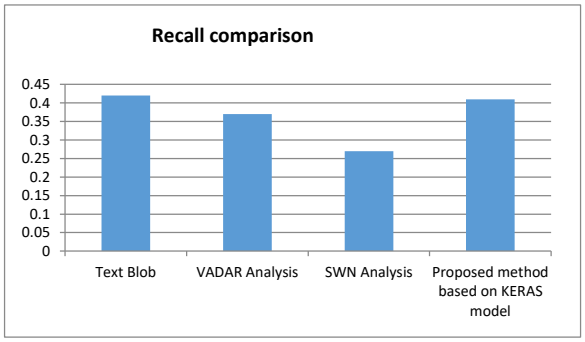

Fig 4: Recall comparison of Textblob, VADAR analysis, SWN analysis & proposed method based on KERAS model

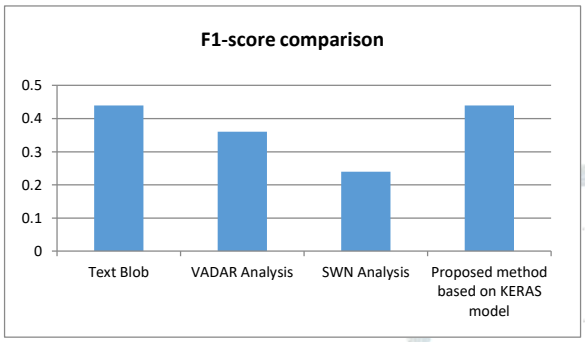

Fig 5: F1-score comparison of Textblob, VADAR analysis, SWN analysis & proposed method based on KERAS model

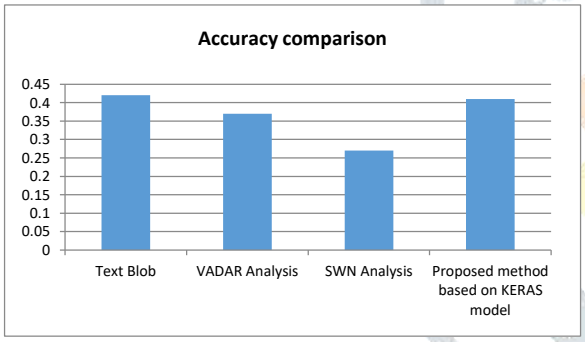

Fig 6: Accuracy comparison of Textblob, VADAR analysis, SWN analysis & proposed method based on KERAS model

VADAR analysis has highest precision (69%). It tells of all the tweets that labeled as positive, how many are actually positive. Textblob method has slightly good recall (42%) compared to our method (41%). It tells of all the tweets that are actually positive, how many we labeled positive. F1-scores of Textblob and our method are same (44%) which takes both false positives and false negatives into account. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. The results show our method is better than VADAR analysis & SWN analysis and equally effective as Text blob analysis for determining the sentiment of the short text.

VI. CONCLUSION

This paper presents existing work in the field of semantic and sentiment analysis of the short, textual data. It has highlighted multiple challenges in text processing such as noisy form, misspelled words, lack of sufficient information and enormous volume. This paper studies various methods published recently in the field of text analysis and has proposed a prototype model to tackle some of the challenges faced by existing approaches.

The main task of understanding semantics of short text is divided into three steps: Segmentation, Type detection and word sense disambiguation. Text segmentation can be done using statistical methods such n-gram or using rule-based approach such as longest cover method. Rule-based approach assigns POS tags to unknown or ambiguous words based on defined language rules. Whereas Statistical approaches build a statistical model automatically from a corpora and labeling untagged texts based on those learned statistical information. However both rule-based and statistical approaches to POS tagging rely on the assumption that texts are correctly structured which is mostly absent in short texts.

Semantic Labeling discovers hidden semantics of input text by assigning appropriate labels to individual terms which provides maximum semantic coherence. Various methods for analyzing semantics of text include Named entity recognition, Topic modeling and Entity linking. Named entity recognition locates named entities in a text and classifies them into predefined categories such as persons, organizations, using linguistic grammar-based techniques as well as statistical models like CRF and HMM. Topic models attempt to recognize latent topics based on observable statistical relations between texts and words. Entity linking method employs existing knowledge base and focuses on retrieving explicit topics expressed as probabilistic distributions on the entire knowledgebase. It recognizes entities and links them to the corresponding entities in a knowledge base.

In this work a generalized framework is proposed to analyze sentiment of short textual data. An effort is made in anticipation of getting an alternative method to understand short texts efficiently and effectively, which exploits semantic knowledge. The proposed method is divided in two parts: offline and online. In offline part a KERAS sequential model is built and trained on twitter dataset. It uses bi-gram for text segmentation and builds co-occurrence network based on occurrence of terms in the dataset. Vocabulary index based on word frequency is built using tokenizer. LSTM-based KERAS sequential model which is a deep learning model is used for determining the similarity and relatedness among terms of input text. The model is trained on twitter dataset of 4,00,000 tweets using multiple epochs. The affinity score is then used to determine the sentiment of the sentence.

To test the effectiveness of the proposed method, airline dataset of 14640 customer reviews with positive, negative or neutral labeled sentiments is taken. Results of proposed knowledge-intensive approach are compared with existing methods namely: Text Blob sentiment analysis, VADAR sentiment analysis and Sentiwordnet analysis. Results are compared using metrics Precision, Recall, F1-score and Accuracy. VADAR analysis has highest precision while Textblob method has almost equal recall as proposed method. F1-score is usually more useful than accuracy, especially if you have an uneven class distribution. F1-score comparison shows our method is better than VADAR analysis & SWN analysis and equally effective as Text blob analysis for determining the sentiment of the short text.

REFERENCES

[1] W. Hua, Z. Wang, H. Wang, K. Zheng and X. Zhou, “Understand Short Texts by Harvesting and Analyzing Semantic Knowledge”, IEEE Transactions on Knowledge and Data Engineering, vol. 29(3), March 2017, pp. 499-512.

[2] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B. Lee, "Twiner: Named entity recognition in targeted twitter stream,” in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, Aug 2012, pp.721–730.

[3] M. Hagen, M. Potthast, B. Stein, and C. Brautigam, "Query segmentation revisited", in Proceedings of the 20th International Conference on World Wide Web, New York, USA, 2011, pp. 97–106.

[4] Z. Yu, H. Wang, X. Lin and M. Wang, "Understanding Short Texts through Semantic Enrichment and Hashing", IEEE Transactions on Knowledge and Data Engineering, vol. 28(2), Feb 2016, pp.566 - 579

[5] Y. Song, H. Wang, Z. Wang, H. Li, W. Chen, "Short Text Conceptualization using a Probabilistic Knowledgebase", IJCAI Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, vol.(3), July 2011, pp.2330-2336.

[6] Z. Wang, K. Zhao, H. Wang, X. Meng, and J. Wen, “Query Understanding through Knowledge-Based Conceptualization”, IJCAI, July 2015.

[7] M. Su, C. Wu, K. Huang, Q. Hong, "LSTM-based Text Emotion Recognition Using Semantic and Emotional Word Vectors", IEEE First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), May 2018

[8] J. Gao, Y.He, X. Zhang, Y. Xia, "Duplicate Short Text Detection Based on Word2vec", IEEE International Conference on Software Engineering and Service Science, Nov 2017

[9] E. Brill, "A Simple Rule-Based Part of Speech Tagger", Proceedings of the third conference on Applied natural language processing, Pages 152–155, March 1992.

[10] J.Yang, G. Huang, B. Cai, "Discovering Topic Representative Terms for Short Text Clustering", IEEE access, vol .7, July 2019, pp. 92037 – 92047

[11] R.Man, K.Lin, "Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network", IEEE Conference, April 2021