International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 946-951 www.ijaem.net ISSN: 2395-5252

Opinion Mining of Hinglish Data Using Transformer Model

Kanchan Chouksey, Prof. Rajendra Arakh
[1]Student, Global Nature Care Sanghthan Group of Institutions, Jabalpur, MP
[2] Prof, Global Nature Care Sanghthan Group of Institutions, Jabalpur, MP

Submitted: 01-08-2021 Revised: 14-08-2021 Accepted: 17-08-2021

ABSTRACT: Text sentiment classification has occupied a pivotal position in sentiment analysis research; it offers important opinion mining functions. Nowadays, with explosion of information, many researchers are focusing on sentiment classification research on massive amounts of data. However, the traditional machine learning methods cannot acquire text semantic information and most research achievements are about single language, in this paper, a multi-lingual based method for opinion mining which integrates the multiple language is being studied.

KEYWORDS: Opinion Mining, Multi-Language, NLP, LSTM, BERT and Transformer.

I. INTRODUCTION

Language is an essential tool for humanity. It is a means of transmitting thoughts, opinions, emotions, persuasion, requests for information, and so on. It draws particular attention to the aim of the intelligent machine, especially in the context of work in the field of artificial intelligence. With the Turing test which is one of the first tests developed to determine whether a machine is intelligent or not [1]. This means that an intelligent machine must have the capacity for comprehension and generation, in the broadest sense of the word, hence the interest in the processing of natural language (NLP) at the dawn of the computer age. Its role is to process information contained in written documents or audio recordings of speeches. Thus, we distinguish two distinct but complementary lines of research in the field of NLP. We have on the one hand, the one that aims to solve the problem by cognitive and linguistic aspects. And secondly, that of optimizing and adapting existing NLP techniques for various application areas such as opinion mining. Thus, discourse marking is today one of the most applied techniques in opinion mining, the first of which was proposed in [2]. The author proposes a model that uses the orientation of two opinion words and co-occurrence statistics obtained from a search engine. First, a speech portion marking (POS) is applied to all words in the document. POS marking automatically identifies the linguistic category to which a word belongs in a sentence. The most used grammatical classes are nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions and interjections. The hypothesis of this model is that sentences containing a sequence of adjectives or adverbs followed by an adverb probably express an opinion.

As a result, all sentences corresponding to one of the Penn-Treebank models [3] described above are then extracted. The sentiment of each selected sentence is calculated using Point wise Mutual Information (PMI) [4]. The probabilities of the PMI values are estimated using the frequencies of the words. They are calculated from the number of times a search engine returns them by responding to a query composed of a sentence and these words. The PMI-SO is calculated as the average PMI-SO of the sentences it contains. If this value is positive, the document orientation is labeled as "positive", otherwise it is labeled as "negative". This method presents a great variability of performance when applied to different domains. It reaches an accuracy of 84% for auto reviews and 66% for movie reviews. This difference in precision can be explained by the fact that the criticisms of the films are subjective while those of the cars are more objective.

1.2 Multilingual Sentiment Analysis

Sentiment analysis aims to automatically identify the sentiment polarity of given texts, which has broad applications, including recommendation systems, sentiment summarization, opinion retrieval, and so on. Given the explosively growing number of online reviews in different languages, multilingual sentiment analysis has recently attracted a great deal of attention from both academia and industries. According to the resources employed, existing methods for multilingual sentiment analysis can basically be categorized into two types, namely, machine-translation-based methods and bilingual-dictionary-based methods.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 946-951 www.ijaem.net ISSN: 2395-5252

Machine translation (MT) has been widely employed in cross-language related work. For example, it is often used to translate the labelled data in a source language into a target language. However, such machine-translation based methods are confronted with three problems: First, they are inefficient when dealing with massive data; Second, current MT systems are not powerful to achieve accurate results. Particularly, they usually generate one best translation, which may not be suitable for the situation at hand; Third, the models used in statistical MT rely on a set of characteristics observed on training examples, but large-scale bilingual parallel corpora for a specific domain are not available in some cases.

Utilizing bilingual dictionaries [5] in multilingual sentiment analysis could be effective as the methods using a high quality MT system. Bilingual dictionaries cannot only reduce workload for labelling data, but also allow one integrating various term weighting and selection methods. However, comprehensive bilingual dictionaries may not be always available, especially for minority language pairs, while generating a bilingual dictionary is difficult and laborious.

In addition to the above issue of resource dependency, another grand challenge of multilingual sentiment analysis is sentiment analysis itself. Sentiment analysis is a hard problem, because many reviews are sentimentally ambiguous for many reasons. For instance, objective statements interleaved with subjective statements can be confusing for learning methods, and subjective statements with conflictive sentiments further make sentiment analysis more complicated.
In the case of multilingual sentiment analysis where the different expression styles in different languages and cultures are considered, the conflictive sentiments problem becomes more difficult.

II. LITERATURE REVIEW

A. Text Summarization

Text summarization focuses on identifying and extracting the main entities and facts from a raw text document. The extraction framework detects discrete portions of the text that are most representative of the document's content. Most existing experiments on text summarization focus on an individual document. Recently, researchers [6] have considered text summarization of multiple documents about related information. The summaries are generated by selecting sentences that address the most specific word associations within the documents. Those approaches rely on the strength of word associations in the set of documents to be summarized.

The work, aspect-based summaries of customer reviews, is related to but different from ordinary text summarization in several ways. First, we do not summarize documents by picking or rewording a subset of the original sentences to obtain their main information, as in common text summarization. Instead, we identify and extract certain product aspects and the corresponding opinions about them from online reviews. Second, we do not focus on facts; we observe and extract subjective information and opinions based on facts. Third, a summary in our system is structured using opinion_aspect relationships, whereas most text summarization systems produce another unstructured text document.

B. Sentiment Classification and Subjective Classification

Sentiment analysis can be categorized into three subtasks: sentiment classification, subjective/objective identification, and aspect-based sentiment analysis. Sentiment classification, also known as document-level sentiment analysis, is the most broadly researched topic. It classifies a review as conveying a positive or negative feeling. In this task, the whole document is considered as the elemental information unit. Researchers have shown that adjectives are good indicators of subjective and evaluative sentences [7]. Turney's group applied an unsupervised learning technique based on point-wise mutual information, and Pang et al. used supervised machine learning methods (support vector machines (SVMs), naive Bayes) to classify movie reviews. Whitelaw et al. [8] applied WordNet to construct a lexicon. To automatically determine whether a term is indeed a marker of opinion content, Esuli and Sebastiani introduced SentiWordNet1 as an enhanced lexical resource for sentiment analysis, and Ohana and Tierney applied SentiWordNet to document-level sentiment classification. Recently, SentiWordNet 3.0 was released, with better accuracy than previous versions. SentiStrength determines sentiment strength from informal English documents using a new method to exploit the de facto grammar and spelling styles of cyberspace. Several researchers have used a lexicon-based approach to extract sentiments from text. A semantic orientation calculator performs the sentiment classification task using a dictionary of words tagged with their semantic orientations and incorporating intensification and negation. Vo and Ock [9] proposed an unsupervised approach that classifies

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 946-951 www.ijaem.net ISSN: 2395-5252

the polarity of a review with a combination of methods, including point-wise mutual information and SentiWordNet, and adjusts the phrase score in the case of modification. However, sentiment analysis at the document-level is too weak for most current applications. The next level of sentiment analysis is subjective classification, which identifies subjective sentences. Sentence sentiment analysis usually includes two steps: identifying subjective sentences and classifying the opinions they express as positive or negative. Rilloff and Wiebe presented a bootstrapping approach that learns linguistic extraction patterns for sentiment expressions. A training set is automatically created by using a high-accuracy classifier to label a dataset, which is then taken as input by an extraction pattern-learning algorithm. The extracted patterns are then used to identify more and more subjective sentences. The pattern-learning algorithm learns many subjective patterns and progressively increases recall while retaining high precision. Yu and Hatzivassiloglou [10] proposed several methods to identify sentence similarity, including the naïve Bayes classification. Mukund et al. used a modified SVM-based approach to distinguish subjective sentences from objective sentences.

C. Aspect-Based Sentiment Analysis

Opinion mining is valuable at both the document and sentence levels, but it does not determine precisely what people liked and disliked. Thus, algorithms are needed to digest a massive amount of information and extract product aspects and their corresponding opinions. In this research, we focus on identifying and extracting the product features that reviewers mention in their reviews. We considered several potential approaches to meet our goal.

1) Extraction Based On Frequent Nouns

Liu et al. used a data mining method to generate feature based customer reviews [11]. This algorithm detects explicit expressions (nouns and noun phrases) from a large review dataset. A part-of-speech tagger is applied to extract nouns and noun phrases. Their occurrence is calculated, and only frequently used ones are kept. This algorithm works because when reviewers comment on different features of a product, their words converge. The precision of this algorithm was improved in the Opine system, which uses relaxation labeling to identify the opinion orientation of words in context; this method accurately identifies sentiment phrases and their corresponding polarities. Moghaddam et al. improved the frequency-based method by adding a filter to delete non-aspect nouns. Zhu et al. [12] introduced a technique that uses the C-value measure from to identify multi-word aspects. Long et al. [13] proposed an aspect extraction method based on frequency and information distance. Their system first identifies main feature words using the frequency-based method and then identifies other words related to the aspect using the information distance measure in.

2) Extraction Using Topic Modeling

Topic modeling is an unsupervised machine learning approach used to summarize documents by considering each document as a mixture of topics and each topic as a probability distribution. Probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) are the two main techniques used for topic modeling. In the opinion mining field, researchers have proposed a joint model to represent both sentiment words and topics simultaneously, which is possible because every opinion has a target. Mei et al. introduced an aspect sentiment mixture model using pLSA and a positive and negative sentiment training dataset. However, some researchers proved that topic modeling is unsuitable for identifying aspects. Later et al. introduced an approach that first identifies aspects using topic models and then detects opinion words by considering only adjectives. In [14], a semi-supervised joint model enables the user to customize some seed feature terms for specific topics to generate aspect distributions that meet a specific requirement.

3) Extraction Using Supervised Learning

Researchers have used many supervised learning approaches for subjective information extraction. The most dominant approaches are based on sequential learning: hidden Markov models and conditional random fields (CRFs). Yu et al. [15] introduced a supervised learning method called One-lass SVM to identify aspects from the pros and cons of review format-2, as in [16]. The aspects are classified and ranked according to their frequency and their contributions to the overall rating of the reviews. Ghani et al. applied both semi-supervised.

learning and supervised learning for aspect identification.

In today's world, the developing nations form a large part of the online users. The exhibition of opinions and sentiments need always be restricted to a single language, say, English. In India, user's tweet using Hindi. Many international companies looking to establish 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 „AurDikhao" hashtag on Twitter. Similarly,

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 946-951 www.ijaem.net ISSN: 2395-5252

Dominos, a fast food chain also has a famous tagline – „Hungry Kya?

Our task is related to but quite different from previous publications because we aim to build an automatic language-based system by using NLP tools to extract coarse syntactic knowledge and infer opinion aspect relationships from the statistics accumulated while obtaining the coarse knowledge. Our application will use opinion aspect relationship knowledge, including product aspect inferences and sentiment extraction.

The taxonomy of sentiment analysis methods are classified into different categories shown in fig. 1.1.

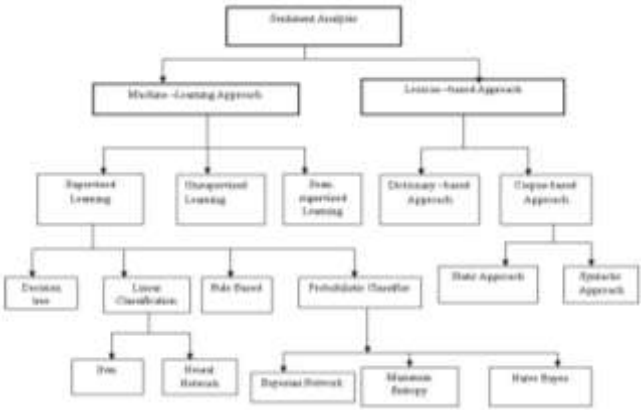


Fig. 1.1: Taxonomy of Sentiment Analysis Methods.

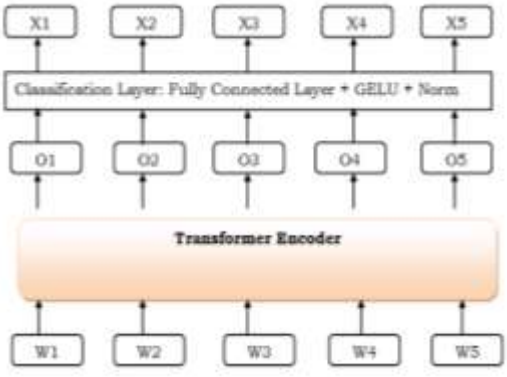
III. PROPOSED WORK

Bidirectional Encoder Representations from Transformers (BERT) is a technique for NLP pre-training developed by Google. BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. The attention mechanism was born to help memorize long source sentences in neural machine translation (NMT).The secret sauce invented by attention is to create shortcuts between the context vector and the entire source input. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Statistics show that half of the messages on Twitter are in a language other than English.

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering (SQuAD), Natural Language Inference (MNLI), and others. BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling.

BERT Working BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. The detailed workings of Transformer are described in a paper by Google. As opposed to directional models, which read the text input sequentially (left-to-right or right to-left), the Transformer encoder reads the entire sequence of words at once. Therefore, it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

The chart below is a high-level description of the Transformer encoder. The input is a sequence of tokens, which are first embedded into vectors and then processed in the neural network.



International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 946-951 www.ijaem.net ISSN: 2395-5252

The output is a sequence of vectors of size H, in which each vector corresponds to an input token with the same index. When training language models, there is a challenge of defining a prediction goal. Many models predict the next word in a sequence (e.g. "The child came home from ___"), a directional approach which inherently limits context learning.

To overcome the limitations of existing work, we proposed a attention based transformer mechanism. It makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text.

Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task.

As opposed to Existing models, which read the text input sequentially (left-to-right or right to-left), the Proposed model will reads the entire sequence of words at once.

3.3 Proposed Algorithm

Step 1: Dataset created of multilingual tweets.
Step 2: Split the datasets into tokens.
Step 3: Map the tokens to their index in the tokenizer vocabulary.
Step 4: Add special tokens to start and end of each sentence.
Step 5: Pad and truncate all sentences to a single constant length.
Step 6: Explicitly differentiate real tokens from padding tokens with the attention mask.
Step 7: Get the final results into three classes, "Positive", "Negative", "Neutral".
Step 8: Exit process.
Finally, this simple fine tuning procedure was shown to achieve state of the art results with minimal task adjustment, for a wide variety of task classification.

IV. RESULTS AND EVALUATION

In the proposed method we work for multilingual opinion mining. Most of the research in the field of opinion mining focuses on single language. But due to demand of opinion mining, here in this work we basically focuses on multilingual opinion mining so that people can get more benefits from user reviews in multiple languages. In the proposed work we have used BERT transformer along with encoding techniques. The result shows that proposed method outperforms the existing model.

Classifier	Accuracy (in %)
Multi-Layer Feed Forward Network	64.70
LSTM	68.92
Proposed	**70.06**

Table 5.1: Performance evaluation.

V. CONCLUSION

In this paper we are using pre-trained BERT- model with transformation mechanism. First, the pre-trained BERT model weights already encode a lot of information about our language. As a result, it takes much less time to train our fine-tuned model - it is as if we have already trained the bottom layers of our network extensively and only need to gently tune them while using their output as features for our classification task. In fact, the authors recommend only 2-4 epochs of training for fine-tuning BERT on a specific NLP task (compared to the hundreds of GPU hours needed to train the original BERT model or a LSTM from scratch!).

REFERENCES

[1]. B. Liu, Sentiment analysis and opinion mining, Synthesis lectures on human language technologies 5 (1) (2012) 1–167.
[2]. K. S. Hasan, V. Ng, Extra-linguistic constraints on stance recognition in ideological debates, in: ACL (2), 2013, pp. 816–821.
[3]. S. Moghaddam, M. Ester, Aspect-based opinion mining from product reviews, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2012, pp.1184–1184.
[4]. B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: Mining text data, Springer, 2012, pp. 415–463.
[5]. W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and

applications: A survey, Ain Shams Engineering Journal 5 (4) (2014) 1093–1113.

[6]. H. Saif, Y. He, H. Alani, Semantic sentiment analysis of twitter, in: International semantic web conference, Springer, 2012, pp. 508–524.

[7]. D. Zhou, Y. Yang, Y. He, Relevant emotion ranking from text constrained with emotion relationships, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Vol. 1, 2018, pp. 561–571.

[8]. A. Balahur, Z. Kozareva, A. Montoyo, Determining the polarity and source of opinions expressed in political debates, Computational Linguistics and Intelligent Text Processing (2009) 468–480.

[9]. O. Biran, O. Rambow, Identifying justifications in written dialogs, in: Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on, IEEE, 2011, pp. 162–168.

[10]. A. Murakami, R. Raymond, Support or oppose?: classifying positions in online debates from reply activities and opinion expressions, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 869–875.

[11]. R. Agrawal, S. Rajagopalan, R. Srikant, Y. Xu, Mining newsgroups using networks arising from social behavior, in: Proceedings of the 12th international conference on World Wide Web, ACM, 2003, pp. 529–535.

[12]. L. Li, Z. Wu, M. Xu, H. Meng, L. Cai, Recognizing stances in mandarin social ideological debates with text and acoustic features, in: Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on, IEEE, 2016, pp. 1–6.

[13]. T. Kyaw, S. S. Aung, Stance mining for online debate posts using part-of- speech (pos) tags frequency, in: 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA), IEEE, 2018, pp. 102–107.

[14]. S. Ghosh, K. Anand, S. Rajanala, A. B. Reddy, M. Singh, Unsupervised stance classification in online debates, in: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, ACM, 2018, pp. 30–36.

[15]. P.Wei,W. Mao, D. Zeng, A target-guided neural memory model for stance detection in twitter, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.

[16]. B. Liu, "Sentiment analysis and opinion mining," Synth. Lectures Human Lang. Technol., vol. 5, no. 1, pp. 1_167, 2012.