Machine Learning for NLP Assignment 2

Theoretical Component

Melis Nur Verir (2726079)

Vrije Universiteit Amsterdam

1 Assignment II

This article will examine the task chosen for one of the key tasks discussed in class, Sentiment Analysis, and previous work techniques used for it. The sentiment refers to a person's perspective, emotion, judgment, or evaluation of a given good or service. Sentiment analysis is contextual text mining that recognizes and extracts subjective information from the source material. It assists businesses in understanding the social sentiment surrounding their brands, products, or services while keeping an eye on online discussions. It has also been used in some studies to detect hate speech.

Feature extraction plays a key role when assessing sentiments from subjective literature using machine learning approaches. Text is used as the input for the feature extraction process, and the generated extracted features can be stylistic, syntactic, or discourse-based, or they can be lexico-syntactic. Therefore, features used in most studies are divided into three categories: discourse-level features, syntactic features, and lexical syntactic features. Discourse level properties are Title of First Paragraph of Document, Term Distribution, and Collocation. Dictionary-syntactic features are POS, SentiWordNet, Frequency and Stemming. The syntactic properties are Chunk Label Dependency Parsing Depth [4]. The dependence depth feature is said to be particularly helpful for constructing theme expressions when these aspects are analyzed. In a dependency tree, a certain Theme word is often found between a specified depth range. Theme expressions might be Named Entities, common names, or words from other POS categories. The frequency of a word in a document is always important in determining its significance. After function words are eliminated, the algorithm provides four distinct high-frequency word lists for the four POS categories of adjective, adverb, verb, and noun. Values for word frequency are then successfully employed as a key characteristic. A stemmer examines all of the word forms found in a certain document's prefixes and suffixes. A stemming cluster approach could have been employed when some languages are not available. A limited number of clusters are formed with the detected root word serving as the center among words that are determined to have the same root form. According to early 2000s studies, the most popular collection of features for sentiment analysis is syntactic properties. These consist of punctuation, part-of-speech (POS) tags, and word n-grams. Phrase patterns that employ POS tag n-gram patterns are another set of syntactic properties. According to Hatzivassiloglou et al., [11] adjectives, adverbs, nouns, and verbs are often the words in sentences that convey opinion, and many tasks for extracting that opinion rely heavily on adjective terms. Previous research showed that certain word patterns, such as nouns followed by positive adjectives frequently expressing positive emotion and nouns followed by negative adjectives frequently expressing negative sentiment, generally imply positive sentiment orientation [1].

In the 2014 Review of Feature Extraction in Sensitivity Analysis study [2], various techniques such as POS tagging, stemming, and stop word removal were applied to the dataset to reduce noise and facilitate feature extraction. Stemming and lemmatization were considered the two main morphological processes of the preprocessing module during feature extraction. The result is that stemming performs faster in applications where accuracy is not a major concern. Studies on feature-based opinion mining have used a range of feature extraction and enhancement techniques, including as NLP and rule-based methods, statistical techniques, and ontology-based techniques. In one of these investigations, a rule-based feature extraction method was suggested by Ding et al., [5].

In comparison to the volume of review data, our technique extracts a sizable number of characteristics. The primary justification for the removal of a significant number of attributes is that words with identical or nearly identical meanings are not regarded as sharing the same qualities. For instance, although the terms "photo," "picture," and "image" all refer to the same object, they are each given various attributes because they are words [8]. A hybrid method that combines characteristics including unigrams, bigrams, POS features, expressions, character n-grams, and preprocessing was provided by Habernal et al. (2014)[7] after they studied and assessed different preprocessing, feature extraction, and classifier strategies.

Based on speech segment analysis, one of the studies [10] presented the random subspace approach (POS-RS) for sentiment analysis. By introducing two crucial factors, content lexicon subspace rate and function lexicon subspace rate, POS-RS preserves the harmony between accuracy and the diversity of base learners. Both bias and variance may be reduced concurrently via POS-RS. A hybrid strategy based on word2vec and SVMperf was proposed by another previous study [6]. To begin with, the study utilizes word2vec to group related characteristics in the chosen domain. After that, creates the training file using feature selection techniques based on lexicon and part of speech. Finally, it uses Word2vec and SVMperf to categorize the comment strings.[9] A recent study [3] performed sentiment analysis by applying feature extraction techniques such as TF-IDF and Doc2vec, effectively and state that classifying the text into positive and negative polarities by using various preprocessing methods like eliminating stop words and tokenization which increases the performance of sentiment analysis in terms of accuracy and time taken by the classifier [3].

References

1. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. **26**(3) (jun 2008). https://doi.org/10.1145/1361684.1361685, https://doi.org/10.1145/1361684.1361685
2. Asghar, D.M., Khan, A., Ahmad, S., Kundi, F.: A review of feature extraction in sentiment analysis. Journal of Basic and Applied Research International **4**, 181–186 (01 2014)
3. Avinash, M., Sivasankar, E.: A study of feature extraction techniques for sentiment analysis. In: Emerging Technologies in Data Mining and Information Security, pp. 475–486. Springer (2019)
4. Das, A., Bandyopadhyay, S.: Topic-based Bengali opinion summarization. In: Coling 2010: Posters. pp. 232–240. Coling 2010 Organizing Committee, Beijing, China (Aug 2010), https://aclanthology.org/C10-2027
5. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. p. 231–240. WSDM '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1341531.1341561, https://doi.org/10.1145/1341531.1341561
6. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications **40**(16), 6266–6282 (2013)
7. Habernal, I., Ptáček, T., Steinberger, J.: Supervised sentiment analysis in czech social media. Information Processing & Management **50**(5), 693–707 (2014)
8. Jeong, H., Shin, D., Choi, J.: Ferom: Feature extraction and refinement for opinion mining. ETRI Journal **33** (2011)
9. Singh, N.K., Tomar, D.S., Sangaiah, A.K.: Sentiment analysis: a review and comparative analysis over social media. Journal of Ambient Intelligence and Humanized Computing **11**(1), 97–117 (2020)
10. Wang, G., Zhang, Z., Sun, J., Yang, S., Larson, C.A.: Pos-rs: A random subspace method for sentiment classification based on part-of-speech analysis. Information Processing & Management **51**(4), 458–479 (2015)
11. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: EMNLP (2003)