A Three Layer Stacked Auto Encoders for Semantic Hashing

Sadi Mohan Krishna
PG Scholar, Department of CSE
Godavari Institute of Engineering & Technology (A)
Rajahmundry, Andhra Pradesh, India.

Radha Mohan Pattaanayak
Associate Professor, Department of CSE
Godavari Institute of Engineering & Technology (A)
Rajahmundry, Andhra Pradesh, India

Abstract— ***Suggesting an original method for sympathetic short texts is to present a device to augment short texts with thoughts and co-occurring terms that are mined from a probabilistic semantic network. To bunch short texts by their senses, we suggest to add more semantic signals to short texts. Precisely, for each term in a short text, we get its concepts and co-occurring terms from a probabilistic information base to augment the short text. Also, we present a basic bottomless learning network entailing of a 3-layer stacked auto-encoders for semantic hashing.***

Keywords: semantic enrichment, semantic hashing, deep neural network

I. Introduction

News endorsement scheme need to procedure the news titles which may be not severely syntactical; in web search, queries contain of an actual minor number of keywords. Short texts present new tests to numerous text related tasks counting information retrieval (IR), classification, and clustering's. Different long documents, two short texts that have alike sense do not unavoidably share many words. For instance, the meanings of "upcoming apple products" and "new iphone and ipad" are carefully connected, but they share no shared words. The lack of adequate statistical info leads to problems in successfully gaging comparison, and as a consequence, countless existing text analytics algorithms do not smear to short texts right. More highly, the deficiency of statistical information also means problems that can be securely ignored when we knob long documents develop dangerous for short texts. Due to the shortage of background information, these vague words make short texts firm to comprehend by machines.

II. RELATED WORK

Salakhutdinov and Hinton recommend a original information recovery mechanism called semantic hashing. The prototypical is stacked by RBMs and acquires to map a document semantic to a compressed binary code. Compared with traditional methods, such as TF-IDF and LSA, their semantic hashing model accomplishes similar repossession presentation.

III. Literature Survey

[1] High quality feature collection is essential to uphold high precision, but now we do not have the branded training data for assessing features that we have in oversaw learning. We present a new feature selection technique that is enthused by pseudo relevance feedback. We usage the top-ranked and bottom ranked documents recovered by the bag-of- words method as typical sets of applicable and non-relevant documents. The produced features are then appraised and drinkable on the foundation of these sets.

[2] We have offered a new kernel function for computing the semantic similarity amongst pairs of short text snippets. We have revealed, both anecdotally and in a human-evaluated enquiry proposal system that this kernel is an real amount of similarity for short texts, and everything well even when the short texts existence careful have no shared terms. Furthermore, we have also if a theoretic examination of the kernel function that shows that it is well-suited for usage with the web. There are numerous lines of upcoming work that this kernel lays the foundation for. The first is development in the cohort of query expansions with the goalmouth of refining the bout score for the kernel function. The additional is the combination of this kernel into other kernel-based machine learning methods to control its aptitude to deliver development in tasks such as organisation and gathering of transcript.

IV. Problem Definition

Several present text analytics algorithms do not smear to short texts straight. Added prominently, the absence of arithmetical information al so means problems that can be carefully unnoticed when we lever long documents convert serious for short texts.

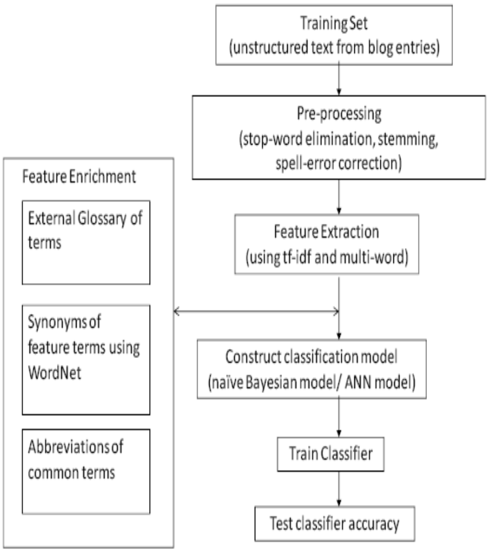
Search-based methods may work well for so-called head queries, but for appendage or detested queries, it is actual expected that some of the top search results are extraneous, which means the augmented short text is probable to cover a lot of blare.

V. PROPOSED APPROACH

Our tactic is a semantic network centered approach for elevating a short text. We contemporary an original apparatus to semantically augment short texts with thoughts and co-occurring terms, such exterior knowledge's are indirect from a large scale probabilistic knowledge improper using our projected detailed methods.

Concepts and co-occurring terms efficiently enrich short texts, and permit heals their sympathetic of them. Our auto-encoder based DNN model is intelligent to detention the intellectual features and compound correlations from the input text such that the scholarly compressed binary codes can be castoff to embody the sense of that text.

VI. SYSTEM ARCHITECTURE


Training Set
(unstructured text from blog entries)
Pre-processing
(stop-word elimination, stemming, spell-error correction)
Feature Enrichment
External Glossary of terms
Synonyms of feature terms using WordNet
Abbreviations of common terms
Feature Extraction
(using tf-idf and multi-word)
Construct classification model
(naïve Bayesian model/ ANN model)
Train Classifier
Test classifier accuracy

VII. PROPOSED METHODOLOGY

7.1 PREPROCESSING:

Stop-words are practical words which befall regularly in the terminology of a language and are not telling of any precise class of documents. Words like "the", "is", "in", "or", "it", "for" etc. are stop-words in English.

Removing stop- words decreases the scope of the text to be treated by the classification algorithm. Stemming reduces a word to its root or base form and thus lessens the number of word features to be treated.

Stemming is based on the remark that words with shared stems typically have comparable meanings. Formless text, particularly blog posts often comprises spell-errors, improper punctuations, abbreviations, special characters which are non-text etc. Spell-errors can be situated and corrected using text processing tools.

7.2 FEATURE EXTRACTION:

The maximum important words or the words with upper most biased power need to be recognized as features for classification. This is typically did using statistical and semi-semantic techniques. Well-known feature ex-traction techniques contain TF-IDF and Multi-words.

TF-IDF is an abbreviation for term frequency-inverse document frequency. It is a probability-based arithmetical amount to control the implication of a word feature in a text document amount.

It is created on the experiential that a term is a good discriminator if it arises regularly in a document but does not happen in countless separate documents of the corpus

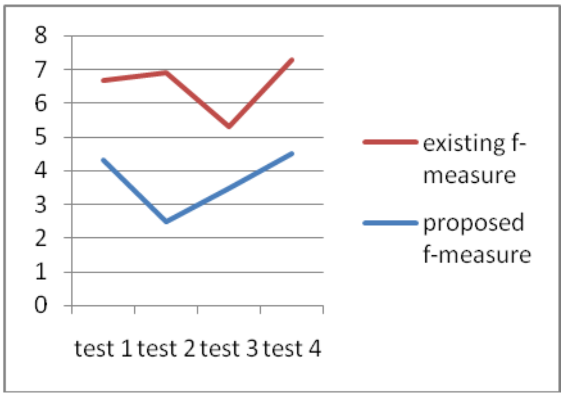
7.3 NAÏVE BAYESIAN:

It is created on the shortening supposition of provisional objectivity between attributes. Given a training set covering attribute values and corresponding target values (classes), the naïve Bayesian classifier envisages the class of an unseen (new) instance, built on beforehand experiential likelihoods of the feature terms up in that instance

VIII. ALGORITHM

INPUT: unstructured text
STEP1: pre process the text includes stop-word elimination, stemming, spell-error correction.
STEP2: implementation of feature extraction.
STEP3: feature enrichment with synonyms of feature terms and abbreviation of common terms.
STEP4: construct classification model by using naïve bayesian model.
STEP5: enrichment of short text by using step4.

IX. RESULTS



In the result designate the proposed approach presentation in terms of f-measure specify the Enrichment correctness associated to previous method.

X. EXTENSION WORK

Regularly this assignment comprises some subtasks in ordinary language dispensation like tokenization, stop-word removal, stemming and spell-error correction shadowed by feature set construction, modeling using an suitable machine learning technique and lastly, classification using the skilled model.

XI. CONCLUSIONS

We bring out inclusive experiments on short text centered responsibilities with information repossession and cataloguing. The note worthy enhancements on both tasks display that our upgrading mechanism could successfully augment short text representations and the projected auto-encoder based deep learning model is clever to scramble compound features from input into the solid binary codes.

References

[1] M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 377–386.

[2] W. tau Yih and C. Meek, "Improving similarity measures for short segments of text," in Proc. 22nd Nat. Conf. Artif. Intell., 2007, pp. 1489–1494.

[3] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang, "Query enrichment for web-query classification," ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 320–352, 2006.

[4] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA, USA: MIT Press, 1998.

[5] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in Proc. 18th ACM Conf. Inf. Knowl. Manage., 2009, pp. 919–928.

[6] S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using wikipedia," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 787–788.

[7] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, pp. 1606–1611.

[8] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in Proc. 19th Int. Joint Conf. Artif. Intell., 2005, pp. 1048–1053.

[9] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in Proc. Int. Conf. Manage. Data, 2012, pp. 481–492.

[10] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledge base," in Proc. 22nd Int. Joint Conf. Artif. Intell., 2011, pp. 2330–2336.

[11] D. Kim, H. Wang, and A. H. Oh, "Context-dependent conceptualization," in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 2654–2661.

[12] B. Stein, "Principles of hash-based text retrieval," in Proc. ACM 30th Annu. Int. Conf. Res. Develop. Inf. Retrieval, 2007, pp. 527–534.

[13] R. Salakhutdinov and G. E. Hinton, "Semantic hashing," Int. J. Approx. Reasoning, vol. 50, no. 7, pp. 969–978, 2009.

[14] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Comput., vol. 14, no. 8, pp. 1771–1800, 2002.

[15] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," J. Amer. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990.