Sentiment Analysis using RNN

CONTENTS

Abstract:

In this paper, we present an algorithm to tackle the problem of classification of sentiment in social media texts at large, i.e. movie reviews, product reviews in e-commerce websites and social media analysis, each consisting of single or multiple sentence(s) that most of the time include pop culture texts. In our experiment, we use some combination of quantitative and qualitative methods. We first generate and empirically cross-validate a gold-standard array of lexical features (with their corresponding sentiment intensities) which are precisely synced with sentiment in microblog-like pieces. Subsequently, we combine the constructed lexical features in accordance with five general rules that represent grammatical and syntactical conventions for emphasizing and expressing sentiment intensity index. We present an architecture that derives vector representations (like word2vec) of the lexical sentiment features. We leverage a new technique that expands upon previous works on sentence-level lexical sentiment classification, using Recurrent Neural Network (where every time-step corresponds to sentence in a micro-blog after stripping down the emoticons and punctuation) and use it jointly with a Recursive Neural Networks.

Keywords: Sentiment Analysis, Word2Vec, Machine Learning, Deep Learning, Recurrent Neural Network, Neural Network, Social Media Analysis, Fuzzy Logic, Cognition.

1. Introduction:

Language is a powerful tool to communicate and convey information. It is also a means to express emotion and sentiment. Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes and emotions towards entities such as products, services, organizations, individuals, issues, events, movies and topics. In simple words, it is used to track the mood of the public. It uses natural language processing and data mining techniques to the problem of extracting opinions from text. In data mining research field, machine learning techniques have been applied to automatically identify the information content in text. In recent years, the exponential increase in the Internet usage and exchange of public opinion is the driving force behind sentiment analysis. The web is a huge repository of structured and unstructured data. The analysis of this data to extract latent public opinion and sentiment is a challenging task. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to the overall document. Sentiment analysis is performed at different levels of granularity with different levels [1]. The most useful application of sentiment analysis is the sentiment classification of product reviews. This sentiment classification can be categorized into positive and negative. Positive rating returns are good, negative returns are bad review. Users express opinions through review sites such as Amazon, Internet Movie Database (IMDB) and Epinions, as well as through blogs, discussion forums, peer-to-peer networks, user feedback, comments and various types of social network sites. These kinds of online media have resulted in large quantities of textual data containing opinion and facts. Over the years, there has been extensive research aimed at analyzing and classifying text and data, where the objective is to assign predefined category labels to documents based upon learned models. This has led to the development of sentiment analysis and classification systems. Sentiment analysis and classification [2] [3] are technically challenging. Bo Pang et al. [4] used movie reviews to train an algorithm that detects sentiment in text. Movie reviews are a good source for this kind of work because authors clearly express an opinion and authors are accompanied by rating that make it easier to train learning algorithms on this data. Most text classification methods that classify a given document into one of the predefined classes are based on bag of words [5], where each document is represented by set of words. The rapid growth of the World Wide Web has facilitated increased online communication and opened up newer avenues for the general public to post their opinions online. This has led to generation of large amounts of online content rich in user opinions, sentiments, emotions, and evaluations. Recently, many websites have offered reviews of items like books, cars, snow tires, vacation destinations, movies etc. Informers describe the items in some detail and evaluate them as good/bad, preferred/not preferred and positive/negative. That is, whether people recommend or do not recommend a particular item. For example, people express their views for movies, as, “I like movies and the story is fascinating”. In sentimental

classification problem, movie review mining is a challenge. The inspiration for this work has come from studies in classification. Research area of sentiment analysis is motivated for doing sentiment classification using new combination of classifier for improving accuracy. A modern approach towards sentiment classification is to use machine learning techniques which inductively build a classification model of a given set of categories by training several sets of labeled documents. Popular machine learning methods include Naive Bayes, K-Nearest Neighbor, Support Vector Machines and Neural Network. A broad range of high quality lexicon is often trivial for fast and accurate sentiment classification on large scales. Linguistic Inquiry and Word Count or LIWC is an ex-ample of such a lexicon [4], [5]. It has been widely used in the social media domain, as its straightforward dictionary and simple word lists are easily inspected, under-stood, and extended if desired.

Our approach seeks to construct a sentiment analysis engine that is 1) well equipped to work on social media style sentences, yet readily normalize to several areas, 2) independent of any training data, but is constructed from a valence-based, generalizable, human-curated gold-standard sentiment lexicon 3) sufficiently fast for using online with streaming data, and 4) unaffected from a performance-speed tradeoff.

2. Review of Literature

Much work has recently been undertaken in sentiment analysis over the last few years. Pang and Lee [4] gives an excellent review. Work has been done specifically on sentiment analysis and even more recently work has been carried out on mining tweets from Twitter. DENG et al. (2011) uses sentiment analysis for Stock Price Prediction in [6]. The classifier tries to classify the review into positive or negative categories. The classification result will be the basis of the rating. With the proportion of positive and negative reviews, the system could provide the rating information to end users. Bo Pang et al., in [4], presented sentiment classification using machine learning techniques. For the effectiveness of classification of documents by overall sentiment used learning methods Naive Bayes, Maximum Entropy classification and Support Vector Machines. The unigrams and bigrams features were used for classification. The movie-review corpus with randomly selected positive sentiment and negative sentiment reviews were used for experimental setup. Result of the machine learning algorithms clearly surpasses the random-choice baseline of 50%. Authors also handily beat two human-selected-unigram baselines of 58% and 64% and performed well in comparison to the 69% baseline achieved via limited access to the test-data statistics. Whereas the accuracy achieved in sentiment classification is much lower when compared to topic based categorization. Yaying Qiu et al., in [7], constructed extend Bayes model with assigning weights to important features. Authors have researched problems in how to classify Chinese text efficiently and effectively. For that purpose authors used the approaches with Naive Bayes and CF methods are used to measure the relevance between a feature and a category to make up the deficiency of Chi-Square statistic method. Authors select best features based on a proposed method called CHCFW to reinforce the distribution of key features in the document and remove the disturbed features. Experiment results had shown that how the size of the best feature set by chosen influence the accuracy using the CHCFW method and the ratio of training set is 80%. Authors effectively classify Chinese text but some problems need to be improved to calculate the weight of each feature which needs a relatively long time, therefore, the whole process is somehow time consuming. Duan et al., in [8], presented mining online user reviews for both quantitative aspects and textual content from multi-dimensional perspectives. In recent years, online user generated content exploded which revolutionized the hotel industry.

Previously, linear classification methods, e.g. Supper Vector Machines (SVM) and logistic regression [1] have been used to solve sentiment Classification problems. Maximum Entropy and Naïve Based Classifiers have also been studied [3] in this field. In [1], Maas et. al. tackle the problem of sentiment classification by learning word vectors, using an unsupervised model to capture nuanced sentimental information. It achieves very impressive results as their semantic and sentiment model capture nuances within the similar sentiments very well.

Richard et. al. [2] semantic word spaces are not ill-equipped to parse meaning of longer phrases. They, instead, propose a Recursive Neural Tensor Network for more rigorous classification. 'Stanford Sentiment Treebank', also introduced in [2], is a dataset that consists of over 2015,154 phrases with fine-tuned sentiment values which spans over a parse tree containing 11,855 sentences. It shows significant improvement over bag-of-words models.

Most of the applied researchers in this field heavily rely on preexisting manually created lists of lexicons. There are some popular lexicons like LIWC 1 [5], GI2[7], Hu-Liu 043[8] etc. where words are divided into binary classes, and on the other hand, lexicons like ANEW 4 [9], SentiWordNet5 [10], and SenticNet6 [11] associate words with valence-based scores for sentence-level sentiment intensity.

3. Objective of the Project

The objective of this project is to show how sentimental analysis can help improve the user experience over a social network or system interface. The learning algorithm will learn what our emotions are from statistical data then determine the mood of the experience. After that it will change our social interactions accordingly on our social network sites or movie datasets and other interfaces like desktop or system services or web-pages. Suppose you are bored or sad, in the case of social networks one thing the computer could do is to be more suggestive of things that lighten your mood and change interactions like backgrounds color's, icons, and services. The movie review dataset websites could automatically try suggesting cinema of our taste with people and their opinion that would help improve the choices, while hiding others that might make it worse. The project aims to implement these in the social network community as well as movie review services and interfaces of our systems, while making our lives better and our experience richer and efficient.

4. System Design

The experiments were performed on an Intel core i5 1.70 GHz processor and the algorithm was implemented using Python 3.4 development tool. The processing time is strictly dependent on the quantity of moving points, train datasets and on the dimension of the test datasets. Tests on the Movie reviews database ran smoothly for positive and negative sentiments for a Bayesian classifier.

- Python 3.4, Natural Language Toolkit 3.0.
- Basic NLP tasks are performed using NLTK 3.0 such as: Stemming, Tokenization, Corpus Reading, Stop Words Removal etc.
- LIBSVM v3.20 (A library for Support Vector Machines). Used Bag of Words Representation of Review for Feature Vector with frequency omitted.

5. Implementation Details

There has been to date no uniform terminology established for this relatively young field. In this section, we attempt to explain some of the terms that are currently in vogue, and what these terms tend to mean in research papers. To see that the distinctions in common usage can be subtle, consider how interrelated the following set of definitions given in Merriam Webster's Online Dictionary are: Synonyms: opinion, view, belief, conviction, persuasion, sentiment

- Opinion implies a conclusion thought out, yet open to dispute (each expert seemed to have a different opinion).
- View suggests a subjective opinion (very assertive in stating his views).
- Belief implies often deliberate acceptance and intellectual assent (a firm belief in her party's platform).
- Conviction applies to a firmly and seriously held belief (the conviction that animal life is as sacred as human).
- Persuasion suggests a belief grounded on assurance (as by evidence) of its truth (was of the persuasion that everything changes).
- Sentiment suggests a settled opinion reflective of one's feelings (her feminist sentiments are well-known).

According to Dave et al. [5], the ideal opinion-mining tool would "process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)." According to Pang et al. [4], the term Sentiment Analysis parallels Opining mining in certain aspects. A sizable number of papers mentioning "sentiment analysis" focus on the specific application of classifying reviews as to their polarity (either positive or negative. Bag of Words is a simple representation of a document/text used in Natural Language Processing and Information. Decision Planes are affine Hyperplanes which can be described by a single linear equation in Cartesian Coordinates. Margin is distance of the object/Point from the decision hyper plane. Smoothing [9] refers to the idea of avoiding noise in data. Tokenization [10] is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. Stemming [11] is the process for reducing inflected (or sometimes derived) words to their stem, base or root form– generally a written word form. Stop Word [12] are the words which don't carry much lexical information". For ex – "the'', "r" etc.

6. Experimental Results

Proposed algorithm is implemented on python 3.6.1 with NLTK package, with an Intel core 5th Gen i7, 2.5 GHz processor with 6 Gigabytes of RAM.

7.1 Experimental Setup

For fair comparison of our results against the broader body of literature, we assess the correlation of calculated sentiment intensity rating to the mean sentiment rating from ten prescreened human raters, along with the 3-class (positive, negative, neutral) classification accuracy parameters of precision, recall, and F1 score. In our analysis, precision is defined as the ration between the numbers of true classification, to the total number of elements labeled as belonging to that class (both correct and incor-rect labeling count). Recall is the ratio of number of correct classification to the total number of classifications that are known to be in the said class. The F1 score signifies the overall accuracy, it is defined as the harmonic mean of precision and recall.

We compared our results to eight state-of-the art sentiment analysis lexicons: VADER (Valence Aware Dictionary for sEntiment Reasoning) [15], Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD) using WordNet, and the Hu- Liu04 opinion lexicon.

7.2 Experimental Analysis

As seen in Table 1-4, in most scenarios, our approach outperforms all the other well-established lexicons for sentiment analysis. In case of social media posts our approach provides better overall precision, recall and F1 score than human raters. And in two other cases we get better recall scores than individual human raters.

Table 1. Comparison of 3-class performance classification performance on Social Media Posts

	Correlation to ground truth (mean of 10 human raters)	Classification accuracy metrics		
		Overall Precision	Overall Recall	Overall F1 Score
Social Media Posts (5,000 Tweets)				
Ind. Human	0.909	0.88	0.81	0.83
Ours	0.813	0.99	0.93	0.96
VADER	0.799	0.99	0.92	0.96
Hu-Liu '04	0.713	0.89	0.67	0.74
SCN	0.542	0.79	0.70	0.72
GI	0.512	0.79	0.51	0.67
SWN	0.441	0.74	0.60	0.61
LIWC	0.606	0.91	0.49	0.60
ANEW	0.451	0.79	0.46	0.57
WSD	0.401	0.69	0.44	0.52

Table 2. Comparison of 3-class performance classification performance on Movie Reviews

	Correlation to ground truth (mean of 10 human raters)	Classification accuracy metrics		
		Overall Precision	Overall Recall	Overall F1 Score
Metacritic Movie Reviews (8,500 review snippets)				
Ind. Human	0.898	0.96	0.92	0.90
Ours	0.691	0.81	0.70	0.69
VADER	0.441	0.74	0.58	0.61
Hu-Liu '04	0.359	0.64	0.48	0.66
SCN	0.255	0.61	0.61	0.5
GI	0.351	0.69	0.61	0.44
SWN	0.245	0.66	0.59	0.60
LIWC	0.168	0.65	0.31	0.44
ANEW	0.164	0.55	0.40	0.44
WSD	0.339	0.60	0.49	0.59

Table 3. Comparison of 3-class performance classification performance on product reviews on Amazon.com

	Correlation to ground truth (mean of 10 human raters)	Classification accuracy metrics		
		Overall Precision	Overall Recall	Overall F1 Score
Amazon.com Product reviews (11,000 review snippets)				
Ind. Human	0.925	0.95	0.81	0.89
Ours	0.765	0.92	0.85	0.85
VADER	0.578	0.78	0.61	0.71
Hu-Liu '04	0.555	0.77	0.66	0.66
SCN	0.368	0.59	0.69	0.59
GI	0.421	0.69	0.52	0.56
SWN	0.365	0.62	0.53	0.58
LIWC	0.319	0.74	0.39	0.33
ANEW	0.277	0.68	0.39	0.41
WSD	0.305	0.60	0.59	0.56

Data used for training and validation of RNN model is demonstrated in Table 5. The data used for the pre-training phase and data used to create the word embedding do not necessarily have to overlap. The neural network was trained on a large number of tweets (30 million) for one epoch, and then it was finally trained on the supervised data (20k tweets) for about 16 epochs.

Table 4. Comparison of 3-class performance classification performance on Social Media Posts

	Correlation to ground truth (mean of 10 human raters)	Classification accuracy metrics		
		Overall Precision	Overall Recall	Overall F1 Score
NY Times Editorials (3,500 Article review snippets)				
Ind. Human	0.792	0.82	0.60	0.71
Ours	0.611	0.78	0.61	0.66
VADER	0.492	0.67	0.51	0.55
Hu-Liu '04	0.441	0.74	0.49	0.59
SCN	0.259	0.60	0.50	0.41
GI	0.339	0.64	0.48	0.51
SWN	0.289	0.59	0.51	0.59
LIWC	0.228	0.60	0.21	0.29
ANEW	0.212	0.61	0.41	0.45
WSD	0.235	0.60	0.51	0.52

Table 5. Data used for training the RNN model

Stages	Twee	Neutral	Negative	Positive
Word embedding	100M	-	-	-
Pre-training	30M	-	21M	19M
Training	20104	7996	7114	4994
Validation	3000	1098	817	1085
Test	21132	8091	6613	6424

The accuracy of the recursive-recurrent approach for training and dev phase are plotted in Fig 1. We used a learning rate of 0.001 and a regularization strength of 0.0001.

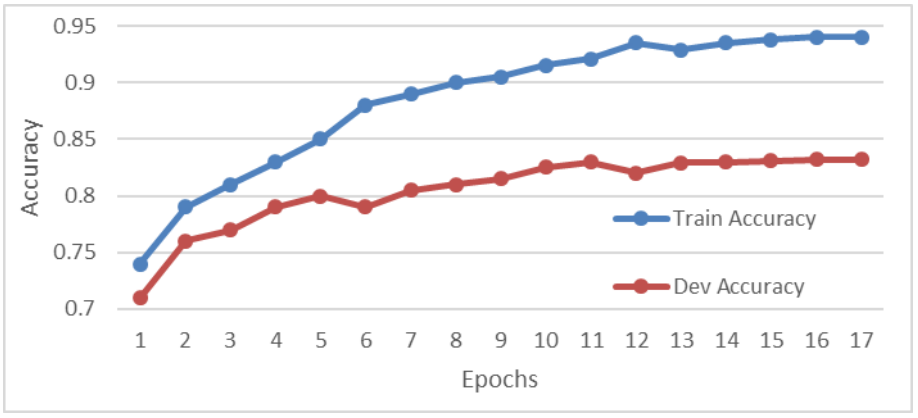

Fig. 1. Transfer Learning RNN: Training and Dev Accuracy vs. Number of epochs

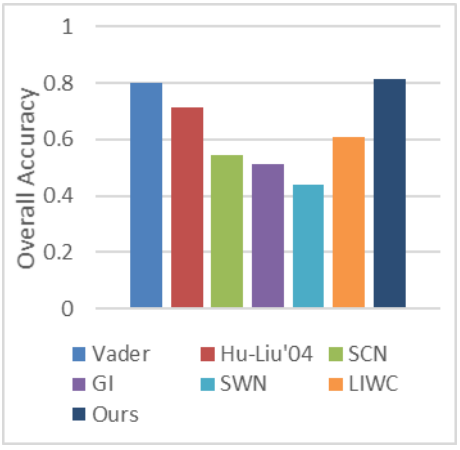

Fig. 2. Performance Comparison on sentiment analysis of tweets.

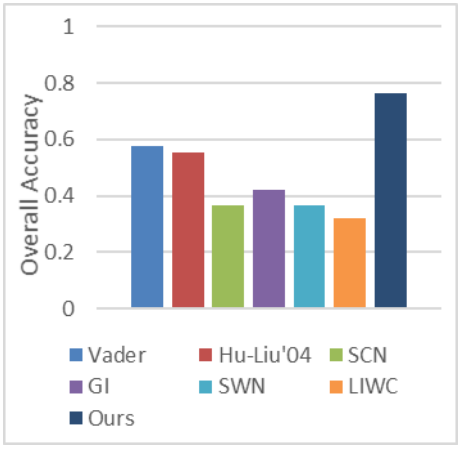

Fig. 3. Performance Comparison on sentiment analysis of movie reviews.

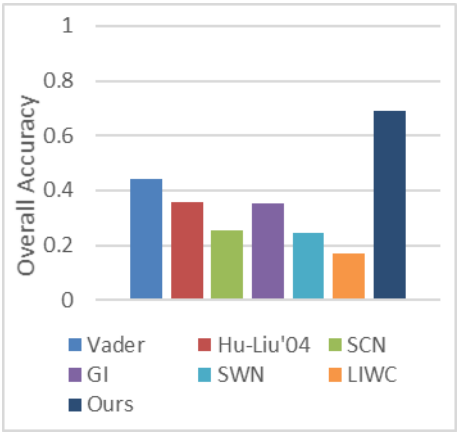

Fig. 4. Performance Comparison on sentiment analysis of product reviews.

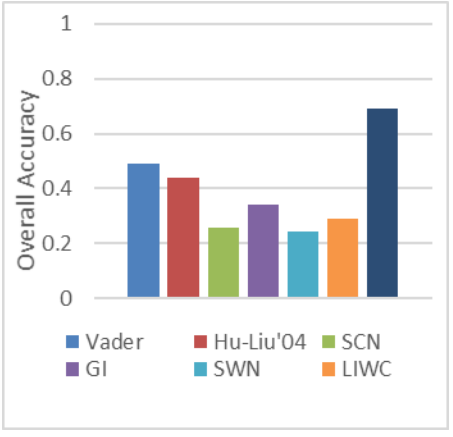

Fig. 5. Performance Comparison on sentiment Analysis of newspaper editorials

7. Conclusion

In this experiment, we present an algorithm to tackle the problem of classification of sentiment in social media texts at large, i.e. movie reviews, product reviews in e-commerce websites and social media analysis, each consisting of single or multiple sentence(s) that most of the time include pop culture texts. In our experiment, we use some combination of quantitative and qualitative methods. We then combine these lexical features with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. The results are not only encouraging – they are indeed quite remarkable; our algorithm performed as well as (and in most cases, better than) other highly regarded sentiment analysis tools. Our results highlight the gains to be made in computer science when the human is incorporated as a central part of the development process. Accuracy of sentiment analysis is increased by the proposed system from dependence and independence assumptions among features. In future, apply this work on clustering domain for movie review dataset for opinion mining applications where the cluster based features are used to address the problem of scarcity of opinion annotated data in a language.

8. References

1. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 142-150). Association for Computational Linguistics.
2. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (Vol. 1631, p. 1642).
3. Pang, B., Lee, L., and Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.
4. Pennebaker, J. W., Francis, M., & Booth, R. (2001). Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Erlbaum.
5. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net.
6. Liu, B. (2010). Sentiment Analysis and Subjectivity. In N. Indurkhya & F. Damerau (Eds.), Handbook of Natural Language Processing (2nd ed.). Boca Raton, FL: Chapman & Hall.
7. Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). General Inquirer. Cambridge, MA: MIT Press.
8. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proc. SIGKDD KDM-04.
9. Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings.
10. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
11. Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2. In Proc. AAAI IFAI RSC-12.
12. Morin, Frederic and Bengio, Yoshua. Hierarchical probabilistic neural network language model. In Proceedings of the International Workshop on Artificial Intelligence and Statistics, pp. 246–252, 2005.
13. Mnih, Andriy and Hinton, Geoffrey E. A scalable hierarchical distributed language model. In Advances in Neural Information Processing Systems, pp. 1081–1088, 2008.
14. Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Distributed representations of phrases and their compositionality. In Advances on Neural Information Processing Systems, 2013c.
15. Gilbert, CJ Hutto Eric. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text." Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf. 2014.