

Multi-entity sentiment analysis using entity-level feature extraction and word embeddings approach

Colm Sweeney
Queen's University,
Belfast
csweeney17@qub.ac.uk

Deepak Padmanabhan
Queen's University,
Belfast
D.Padmanabhan@qub.ac.uk

Abstract

The sentiment analysis task has been traditionally divided into lexicon or machine learning approaches, but recently the use of word embeddings methods have emerged, that provide powerful algorithms to allow semantic understanding without the task of creating large amounts of annotated test data. One problem with this type of binary classification, is that the sentiment output will be in the form of '1' (positive) or '0' (negative) for the string of text in the tweet, regardless if there are one or more entities referred to in the text. This paper plans to enhance the word embeddings approach with the deployment of a sentiment lexicon-based technique to appoint a total score that indicates the polarity of opinion in relation to a particular entity or entities. This type of sentiment classification is a way of associating a given entity with the adjectives, adverbs, and verbs describing it, and extracting the associated sentiment to try and infer if the text is positive or negative in relation to the entity or entities.

1. INTRODUCTION

Sentiment Analysis incorporates the use of natural language processing, statistics and text analysis to identify and extract subjective information in source materials that can refer to a named entity. A Named Entity (NE) is considered: a person, an organization, a location, an expression of time, a quantity and others, referred to by name. Once the entities have been identified using POS tagging or entity extraction processing, they can be used to identify the main subjects of a sentence, and once identified, the descriptive words, for example, adjectives, verbs or adverbs can be used to identify sentiment towards the entity.

The advent of social media has allowed users write large numbers of short text messages to comment about current events, politics and products. As a form of communication, in some situations, it provides information that can be more up-to-date than conventional news sources, and this has encouraged the research community to analyze ways to extract information from this data source. Microblogging sites have established themselves as a main communication tool, for example Twitter has attracted "over 500 million registered users and publish 340 million tweets per day" (Lunden, 2012). These tweets often contain mentions of numerous entities and additional information, like an opinion, that is being viewed more and more, as valuable currency, that can make or break a product in the marketplace.

Some of the problems encountered when processing social media content include: the length of texts (texts are typically very short, for example Twitter's limit is 140 characters); noisy texts (informal text snippets include misspellings and do not contain grammatically correct sentences); and the reliability of information in this type of text messages is uncertain, compared to news media texts. The short message length limits the amount of context available to allow an understanding of the text content (Lim, et al., 2013), but these short snippets of text can be used to provide real-time insights into the aggregated sentiments of people and capture public opinion about product preferences.

While machine learning can identify the expressed sentiment, they ignore any implicit sentiment that relates to one specific entity or another. This paper will attempt to analyze sentiment at the entity level, to designate the sentiment associated with the individual entity and not to the text snippet. For example, the following comment "Applause! Insurelife recognized for best practices unlike Busibank, worst ever", indicates two sentiments for

two different financial companies, Insurelife and Busibank. While the sentiment towards Insurelife would be read as positive, the sentiment relating to Busibank is negative. It would be ineffective to label the overall opinion of this post, as it does not solely relate to one entity. With this in mind, the key aim of this project is to investigate how entities and their descriptor words can be used to identify the sentiment of the tweet in relation to the entity or entities, where more than one entity exists.

This can be approached through the use of one of the most popular sentiment lexicons – SentiWordNet 3.0, which evolved from annotating WordNet synonym entries in relation to their polarity scoring (Baccianella, et al., 2010). The scores given for positive, negative and neutral classes range between zero and one, and the summation of all three scores is 1. SentiWordNet has been shown to perform better than other lexicon dictionaries (Taboada, et al., 2011).

2. RELATED WORK

The main objective of Sentiment Analysis is to associate a given entity with a word or phrase describing the entity and extract the associated sentiment identifiers to infer if the text is positive or negative in relation to the main entity. This can also be called opinion mining/extraction and is now a topic of active research interest in many different communities. Until recently, the main body of research in the area of Sentiment Analysis has been conducted on news feeds and in particular types of domain, for example movie review websites, and most of the existing Twitter-specific Sentiment Analysis research seems to be term-based (see (Pak, et al., 2010), (Go, et al., 2009), (Barbosa, et al., 2010)), where certain words are extracted from Twitter posts that contain a certain term, and the sentiments of these terms are analysed from the tweets.

There are two traditional approaches to this type of sentiment classification. The first is the knowledge based approach, where predefined dictionaries of opinion words are used to search the input words to find the appropriate instances. The publicly available corpus SentiWordNet (Baccianella, et al., 2010) is an extension of WordNet where each

synset is annotated with labels indicating how objective, positive, and negative the terms in the synset are. The sentiment is determined by comparing tweets against the pre-defined entry in the dictionary, which makes it easy to determine the polarity of a specific sentence. As the core objective is to identify sentiments with respect to entities (and their attributes) from natural language, following on from the last stage, the attributes (adjectives, adverbs and verbs) that have been identified as being associated with a main entity or topic ((Hatzivassiloglou, et al., 2000) (Taboada, et al., 2011) (Volkova, et al., 2013)) can be checked to see if they have a corresponding sentiment polarity in the sentiment dictionaries to indicate a positive or negative scoring for the overall tweet.

The second (language based) approach involves supervised machine learning, where trained classifiers can be used for sentiment classification. The most common algorithms for machine learning approaches include: Naive Bayes; Support Vector Machines; (SVM) and Maximum Entropy, and different algorithms achieved different results for the different research parties - SVM being preferred by (Pang, et al., 2002) and Naive Bayes by (Pak, et al., 2010) and (Parikh, et al., 2009) who found that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

As opinions are usually targeted at an entity, an appropriate approach to Sentiment Analysis is the use of Part-Of-Speech (POS) tagging to label the entity types and any other words, or phrases that are relevant to the entity. POS tagging is also useful as noun and verb forms of the same word may have different sentiments. Early work by (Hatzivassiloglou, et al., 2000) found that the presence and type of certain terms like adjectives can be used to indicate whether a sentence is being subjective or objective. Other parts of speech have been found to be useful in Sentiment Analysis, such as adverbs ((Benamara, et al., 2007), (Taboada, et al., 2011)) nouns (Nasukawa, et al., 2003), and verbs (Wiebe, et al., 2004), or all three, ((Volkova, et al., 2013), (Jmal, et al., 2013), (Subrahmanian, et al., 2008)) as these types of words have been found to play an important role in Sentiment Analysis.

While the majority of sentiment analysis relates to text where the opinions are focused on a central

topic, this may not always be the case, as opinions can be expressed about a number of entities. According to Liu (Liu, 2012), there can be two types of opinions, regular opinions and comparative opinions (Jindal, et al., 2006). Regular opinions relate to one entity or an aspect of the entity, for example “The iPhone screen is small,” which expresses a negative sentiment on the aspect - screen of the iPhone. A comparative opinion, on the other hand compares more than one entity based on any number of shared aspects, for example “The Samsung Galaxy S8 looks better than iPhone 7” compares a Samsung phone and an iPhone based on their appearance (an aspect) and expresses a preference for the Samsung phone.

Most sentiment analysis research considers the whole document or text snippet as having a single polarity, in relation to a movie, for example. But in many cases a segment of text can have opinions on more than one entity with differing polarities. This suggests that the polarity classification should be performed at an entity level. One of the earliest approaches to sentiment analysis for multiple entities was proposed by (Moilanen, et al., 2009) in 2009. Their approach was to use a dependency parser to establish at entity level, the polarity of each entity in the text. A later approach by (Gryc, et al., 2010) applied the same techniques to infer the polarity in the political domain on blog messages.

Recently, word embedding-based approaches (Mikolov, et al., 2013), (Pennington, et al., 2014)) have become popular, as they are able to capture the semantics and context of words, by using the machine learning approach of Neural Networks (Socher, et al., 2011). The Word2Vec method (Mikolov, et al., 2013) can be a supervised or unsupervised word embedding-based approach that aims to detect the meaning and the relationships between words by learning how the words co-occur in a corpus. It produces a vector space where each unique word in the corpus is allocated a corresponding vector in the space.

These vectorized representations of words, learned through the Word2Vec algorithm, have proven to efficiently manage semantic meanings of those words, where words that have a similar context in

the training corpus are located in close proximity in the word vector. This type of model has proven very useful in the field of Natural Language Processing (NLP) applied to Sentiment Analysis.

This research proposes a dual classification approach for the task of polarity analysis. An initial classification of tweets identifies those that relate to a single entity, and the word embedding-based approach will be used to determine the opinion polarity of these tweets. Only the subjective tweets are being used as (Pang, et al., 2002) have found that polarity classification is improved by the removal of objective sentences from the training set. The next step will analyse and classify the emotion relating to multiple entities in the Twitter text by using a sentiment lexicon-based technique that appoints a score to indicate the polarity of opinion in relation to each particular entity in text that contains multiple entities.

3. METHODOLOGIES

The goal is to build a classification system that can decide whether a tweet is positive or negative towards each given entity. The training dataset that is used contains only positive or negative tweets, as they denote subjective opinion.

For those tweets that are identified as containing more than one, different entity, they are removed for processing as described in section 3.2 and 3.3. These tweets are analyzed, where particular descriptor words are extracted as features from the text, and the sentiment of these words are identified using the SentiWordNet lexicon. The remaining tweets are processed using the Word2Vec algorithm, as below. The dataset used for training and testing contains 1,578,627 classified tweets, where each row is marked as 1 for positive sentiment and 0 for negative sentiment. Approximately 1/10 of the corpus was used for testing the algorithm, while the rest was used for training.

3.1. Training

Before training the Word2Vec model, the data was pre-processed, where all uppercase letters and words were changed to lower case letters to ensure

uniformity. In addition, the sentences were split into individual tokens by using NLTK's **punkt** tokenizer. Additional pre-processing steps clean the data by removing any HTML markup and punctuation. Stop words which refer to the most common words in a language, such as “a” and “the”, are usually removed for NLP tasks. In the case of training the Word2Vec model, stop words are not removed, as the algorithm relies on the broader context of the sentence in order to produce high-quality word vectors. With the pre-processing complete, the Word2Vec model is trained on the given training dataset that was derived from a total of over a million tweets (dataset sourced from Kaggle¹), with the sentiment tag removed to provide unlabelled training, as Word2Vec does not need labels in order to create meaningful representations (Ye, et al., 2015).

3.2. Word2Vec algorithm

The Word2Vec algorithm uses the genism method in Python to create the word vectors. Word2Vec can employ one of two training algorithms – Continuous bag- of-words (CBOW) or skip-gram. As the skip-gram algorithm approach has proved to be more efficient for big datasets, as is the case of the corpus adopted, it is selected for this work. Word2Vec works in a way that is comparable to deep approaches like deep neural networks, where it allows for a representation of semantically similar words with neighboring points in the same vector space.

3.3. Dependency Parser

The Twitter texts that contain more than one entity will be processed further by taking the descriptor features and using SentiWordNet for an overall score per entity. SentiWordNet has been chosen as the sentiment lexicon for this task, as it has been shown to perform better than other lexicon dictionaries (Taboada, et al., 2011). First the Twitter specific dependency parser, TweepoParser from (Kong, et al., 2014), is used to generate a syntactic structure of each tweet. The tokenized tweet is

created with the associated part-of-speech tags and syntactic dependencies. The parsed file is structured according to the CoNLL format showing different information in different columns.

The generated syntactic structure for the following tweet is presented in Table 1:

“Applause! Insurelife recognized for best practices unlike Busibank, worst ever”

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD
1	Applause	-	N	N	-	0
2	!	-	,	,	-	-
3	Insurelife	-	^	^	-	4
4	recognized	-	V	V	-	0
5	for	-	P	P	-	4
6	best	-	A	A	-	7
7	practices	-	N	N	-	5
8	unlike	-	P	P	-	4
9	Busibank	-	^	^	-	9
10	,	-	,	,	-	-
11	worst	-	A	A	-	6
12	ever	-	R	R	-	6

Table 1. TweepoParser output

The fields in Table 1 include: ID showing token position; FORM showing the form of the word (or punctuation symbol); CPOSTAG and POSTAG showing the coarse grained and fine grained part-of-speech tags respectively; the HEAD denotes the head of the current token; and DEPREL shows any dependencies and the type of dependency relation (CNL-2006).

For the example tweet above, two separate entities – Insurelife and Busibank have been identified by the ‘^’ symbol in the POSTAG field. The 2 entities do not match and are therefore different. With the entities identified, any neighboring words showing

¹ <http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/>

part-of-speech tags for Adverbs, Adjectives and Verbs will be analysed for sentiment, using entity-level feature extraction.

3.4. Entity-level feature extraction

This involves the identification of the dependent words in relation to the entities, where these words are used to identify the polarity towards a given entity. The tweet level and the entity level polarity (towards a given topic) can be different, when relating to different topics or entities, resulting in contrasting opinions in the one tweet. The entity-level feature extraction is applied using the SentiWordNet lexicon against the parsed text to identify the polarity of the descriptor words, to better identify the sentiment of a tweet towards a given entity or entities.

Continuing on from feature extraction research based on associated token words on either side of the topic (Sweeney, et al., 2017), the dependency of descriptor words on a topic or entity, can be considered more intuitive as an effective method, as neighboring tokens (2 words on either side of the entity) can emphasize the correct sentiment value. With only descriptor words being used as neighboring tokens, this research aims to improve on this by identifying the descriptor words and using them to identify the sentiment directed towards individual entities.

Figure 1 shows the architectural components of the proposed hybrid supervised solution. Tweets are processed using the TweepoParser application. The parsed output is used to differentiate tweets that relate to more than one entities and the remaining, that relate to one entity only.

3.5. Scoring tweets

The descriptor words of the multi-entity tweets are scored using SentiWordNet sentiment lexicon. The overall scoring per entity is printed out as output for the multi-entity tweets. The remaining tweets are classified using a Random Forest classifier that uses feature vectors created by a Word2Vec model that was trained on a large classified Twitter corpus. The output includes the tweet number and the sentiment

scoring per tweet – 1 for a positive sentiment and 0 for a negative sentiment.

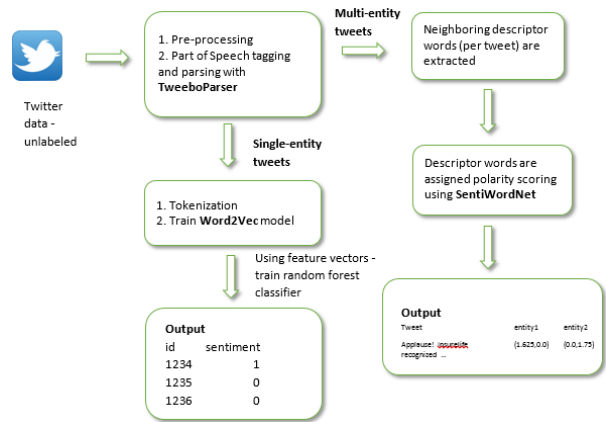


Figure 1. Flowchart of application

4. Results

As a baseline, the texts were parsed with the Tweet specific parser - TweepoParser and by extracting descriptor words (Adverbs, Adjectives and Verbs) only per tweet and using SentiWordNet to score each of the descriptor words, the scoring for this system is providing **69%** accuracy. To improve on this system, the subjective tweets that relate to only one entity are processed by the Word2vec modelling technique integrating a neural network implementation to learn the distributed representations for words to allocate a scoring per tweet. The tweets that relate to more than one entity are processed separately using TweepoParser, and the descriptor words neighboring each entity are used to score sentiment relating to that entity. With all entities in multi-entity tweet being correctly scored, this provided an improvement in accuracy to **71%**. So, by integrating a word embeddings approach for single entity tweets, accompanied by a sentiment lexicon approach on parsed text for multi-entity tweets, this has improved the accuracy of sentiment scoring on Twitter texts.

To evaluate the performance of sentiment classification systems, the following three performance measures are used: Precision; Recall and F1-Measure. Precision and Recall are calculated and combined to produce the averaged F1-measure. Positive or negative reference values were used to calculate the precision, recall, and F-

measure of the lexicon-based classification approach (see table 2).

	Precision	Recall	F1-score
Classification using SentiWordNet + POS	0.68	0.68	0.68
Classification using Word2Vec +SentiWordNet + POS	0.72	0.69	0.71

Table 2. Application results

5. CONCLUSIONS AND FUTURE WORK

Most research in the field of sentiment analysis usually integrates either machine learning algorithms or lexicon based approaches. With the

increased popularity of word embeddings methods, this research aims to highlight how a hybrid word embeddings and lexicon based approach can be used to tackle the problem of sentiment analysis on multiple entities. Using a Twitter-specific parser to identify the descriptor words that relate to a specific entity for text that contains multiple entities, this research is able to achieve very good results.

Future work could involve other factors that can have a high impact on sentiment for a tweet containing more than one entity, for example identifying negation to improve the accuracy of the lexicon based part of the classification. Further investigation of entity and descriptor word relationship could also be investigated further. For example, if the word is a noun preceded by an adjective, or if the preceding word is an adverb, this could be indicators of a possible descriptor/entity relationship.

REFERENCES

Baccianella S., Esuli A. and Sebastiani F.

SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining [Conference] // Proceedings in Language Resources and Evaluation Conference (LREC-10). - 2010. - pp. 2200-2204.

Baccianella S., Esuli A. and Sebastiani F.

SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining [Conference] // International Conference on Language Resources and Evaluation (LREC - 2010). - 2010.

Barbosa L. and Feng J. Robust sentiment detection on Twitter from biased and noisy data [Conference] // In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages . - [s.l.] : Association for Computational Linguistics, 2010.

Benamara F. [et al.] Sentiment analysis: Adjectives and adverbs are better than adjectives alone [Conference] // International Conference on Weblogs and Social Media (ICWSM '07). - 2007.

CNNL-2006 CoNLL-X Shared Task: Multi-lingual Dependency Parsing [Online] // <http://ilk.uvt.nl/>. - ILK Research Group. - <http://ilk.uvt.nl/conll/#dataformat>.

Go A. and Bhayani R. Huang, L. Twitter sentiment classification using distant supervision - CS224N Project Report [Report]. - [s.l.] : Stanford, 2009.

Gryc W. and Moilanen K. Leveraging Textual Sentiment Analysis with Social Network Modelling [Conference] // In Proceedings of the "From Text to Political Positions" Workshop. - Amsterdam : [s.n.], 2010.

Hatzivassiloglou V. and Wiebe J.M. Effects of adjective orientation and gradability on sentence subjectivity [Conference] // In Proceedings of the 18th conference on Computational linguistics. - [s.l.] : Association for Computational Linguistics, 2000.

Jindal N. and L. Bing Mining comparative sentences and relations [Conference] // Proceedings in National Conference on Artificial Intelligence (AAAI-2006). - 2006.

Jmal J. and Faiz R. Customer review summarization approach using Twitter and SentiWordNet [Conference] // Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics. - [s.l.] : ACM, 2013. - p. 33.

Khan F. H., Bashir S. and Qamar U. TOM: Twitter opinion mining framework using hybrid classification scheme [Journal] // Decision Support Systems. - 2014. - pp. 245-257.

Kiritchenko S., Zhu X. and Mohammad S. M. Sentiment analysis of short informal texts [Journal] // Journal of Artificial Intelligence Research. - 2014. - pp. 723-762.

Kong L. [et al.] A dependency parser for tweets [Conference] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-14). - [s.l.] : Association for Computational Linguistics, 2014. - pp. 1001-1012.

Kouloumpis E., Wilson T. and Moore J. D. Twitter sentiment analysis: The good the bad and the omg! [Conference] // The International AAAI Conference on Web and Social Media (ICWSM-2011) . - [s.l.] : Association for the Advancement of Artificial Intelligence (AAAI), 2011.

Lim E-P., Chen H, and Chen G. Business Intelligence and Analytics: Research Directions [News Article] // Transactions on Management Information Systems (TMIS). - [s.l.] : ACM, 2013 йил. - 4 : Vol. 3.

Liu B. Sentiment analysis and opinion mining [Journal] // Synthesis lectures on human language technologies. - 2012. - 1 : Vol. 5. - pp. 1-167.

Lunden I. Analyst: Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City [Online] // <http://techcrunch.com>. - 2012 йил. - 2014 йил 12-10. - <http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/>.

Mikolov T. [et al.] Efficient estimation of word representations in vector space. - [s.l.] : arXiv preprint arXiv:1301.3781, 2013.

Mohammad S. and Turney P.D. Crowdsourcing a Word-Emotion Association Lexicon [Article] // Computational Intelligence. - 2013.

Moilanen K. and Pulman S. Multi-entity Sentiment Scoring [Conference] // Proceedings in RANLP (Recent Advances in Natural Language Processing) 2009. - 2009.

Nasukawa T. and Yi J. Sentiment analysis: capturing favorability using natural language processing [Conference] // Proceedings of the Conference on Knowledge Capture. - 2003.

Owoputi O. [et al.] Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters [Conference] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. - [s.l.] : Association for Computational Linguistics, 2013.

Pak A. and Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining [Conference] // Proceedings of the International Conference on Language Resources and Evaluation (LREC '10). - 2010.

Pang B., Lee L. and Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques [Conference] // In Proceedings of the ACL-02 conference on Empirical methods in natural language processing. - [s.l.] : Association for Computational Linguistics, 2002. - Vol. 10.

Parikh R. and Movassate M. Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques - CS224N Final Report [Report]. - 2009.

Pennington J., Socher R. and Manning C.D. GloVe: Global vectors for word representation [Conference] // In Proceedings of the conference on empirical methods on natural language processing (EMNLP '14). - 2014.

Socher R. [et al.] Semi-supervised recursive autoencoders for predicting sentiment distributions [Conference] // In Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP '11). - 2011.

Subrahmanian V. S. and Reforgiato D. AVA: Adjective-verb-adverb combinations for sentiment analysis [Journal] // IEEE Intelligent Systems. - 2008. - pp. 43-50.

Sweeney C.J., Hong J and Liu W Sentiment Analysis using entity-level feature extraction [Conference] // In Proceeding of Conference on Computational Linguistics and Intelligent Text Processing (CICLing '17). - Budapest : [s.n.], 2017.

Taboada M. [et al.] Lexicon-based methods for sentiment analysis [Journal] // Computational linguistics. - 2011. - pp. 267-307.

Volkova S., Wilson T. and Yarowsky D. Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams [Conference] // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. - [s.l.] : Association for Computational Linguistics, 2013. - pp. 505-510.

Wiebe J. M. [et al.] Learning subjective language [Journal] // Computational Linguistics. - 2004. - Vol. 30.

Ye K. [et al.] Summarizing Product Aspects from Massive Online Review with Word Representation [Conference] // In International Conference on Knowledge Science, Engineering and Management. - [s.l.] : Springer, Cham, 2015.