Conventional and Structure Based Sentiment Analysis: A Survey

Omar Ali
School of Computing
University of Portsmouth
Portsmouth, UK
omar.ali1@port.ac.uk

Alexander Gegov
School of Computing
University of Portsmouth
Portsmouth, UK
alexander.gegov@port.ac.uk

Ella Haig
School of Computing
University of Portsmouth
Portsmouth, UK
ella.haig@port.ac.uk

Rinat Khusainov
School of Computing
University of Portsmouth
Portsmouth, UK
rinat.khusainov@port.ac.uk

Abstract—Sentiment Analysis is a strand of Natural Language Processing that deals with the emotional polarity a given piece of text has. To gain this understanding from just a string of words, we must first consider a suitable way to break down the text to further classify what each part means. This can be done in a plethora of ways, which mostly stem from the understanding of a classifier. We believe that there is a large amount of information stored in structure-based features within the text, for instance, where the writer may place negation-terms *not, neither* and how this affects the overall polarity. Similarly, how words in sentences or sections, remote to the current-analysed section, may affect the polarity of said section. A combination of features from both a conventional and a structure-based understanding may also provide us with a larger accuracy in polarity. Therefore, this paper aims to explain both conventional sentiment analysis methods with structure-based methods as well as their practices, advantages and disadvantages concluding with how sentiment analysis can move forward with the appropriation of *hybrid methods* (methods involving motifs, practices and understandings) from conventional and structure-based methods, for classification.

Index Terms—machine learning, sentiment analysis, naïve bayes, support vector machines, maximum entropy, rhetorical structure theory, discourse analysis

I. INTRODUCTION

Sentiment analysis is the process of defining the polarity of a word, phrase, sentence or block of text. The polarity may be defined as either positive, negative or in some cases neutral. Such information can be useful when deciphering a person's opinion or emotional state from a string of words. The common approaches to sentiment analysis can involve: dictionary-based methods whereby sentiment scores are aggregated for a set of words defining that word's polarity; Dictionary-less approaches involving machine-learning implementations, such as Naïve Bayes, Support Vector Machines, and Maximum Entropy, where classifiers (systems for categorising input data) are trained with manually labelled inputs (Supervised learning) [1]; Neural-Network based approaches such as BERT whereby a model can be trained without the need for any labelled data (Unsupervised) and lastly, methods involving both machine-learning and dictionaries where machine-learning methods are used for the generation of dictionaries to be used when denoting a word's polarity. This paper intends to look into the current, respected, methods of sentiment analysis as well as explore the current implementations of structure-based analy-

sis, namely Rhetorical-Structure-Theory with text discourse-analysis, concluding with a proposed combination of both word and structure-based approaches we intend to research further in the future.

II. THE SURVEY

This work aims to explore what structure-based sentiment analysis methods can provide that conventional methods do not. Furthermore, we intend to expand the understanding of conventional methods that use structure-based features (ambiguous methods) and assess their impacts, both good and bad, on sentiment analysis, concluding lastly with a discussion on hybrid-methods more explicitly combining both conventional and structure-based methods.

III. BACKGROUND AND MOTIVATION

A. Background

Automated text classification is a large field and has been thoroughly explored in [2], as well as more specifically in sentiment analysis through the use of neural networks in [3]. Word embedding methods used for classifications in sentiment analysis, using different types of neural network architecture have been explored further in [4] more specifically in terms of deep learning. Naive Bayes and Support Vector Machines have both been explored in [5] for text classification uses. Rhetorical Structure Theory and Discourse Structure have been presented in [6] and their uses in text classification have been explored in [7]. The use of discourse structure and its implications on sentiment analysis, have also been covered in [8].

B. Motivation

Subjectivity in text plays a huge part in the way other people understand written text. However, it is not as trivial for computer/set of algorithms to convey such information. Considering the main approaches to sentiment analysis, it is clear that structure-based features present in the text can be lost as words are broken down to be classified and used. The structure can play a substantial part when deciding polarity of an input-text. For instance, from a given input-text with known word-polarities, positive$^+$ and negative$^-$:

"I believe running to be a great$^+$ past-time. Swimming is not as enjoyable$^+$."

We can see that on a word-level, there is a directly polar word that may sway each sentence to be positive, however, due to the placement of a negation-term *not as*, we can deduce, from our understanding, that the second sentence is not a positive one.

Similarly,

"Swimming is not as enjoyable$^+$ a past-time as running."

The change in sentence structure slightly changes how we interpret the overall polarity of that sentence. The scoping of the negation-term *not as* changes as it affects more of the words in the sentence overall. We must be able to consider a way in which we can effectively scope the negation-term that takes into account similar-meaning sentences. For instance:

"Swimming is [not$^-$ as enjoyable$^+$]$^-$ a past-time as running."

This example shows how negation can be considered in the overall polarity of the given sentence. The next, less trivial, issue to solve is how we can scope said negation to make sure that it does not affect the polarity of a term later on in the sentence.

On a larger scale, given the input text:

Example 1:

- *My least$^-$ favorite things about swimming are:*
  - *The great$^+$ depth of water;*
  - *The bad$^-$ service at the swimming centre.*

Example 2:
The colour green is not$^-$ my favorite$^+$ colour, but I like$^+$ green trees.

In Example 1, we can intuitively see that the polar-word *least* has a knock-on effect in the latter statements within the list, namely *"The great$^+$ depth of water"*. This means that the scoping of the negation-term needs to be closely considered when evaluating the other points as well at the negation term's own sentence. When it comes to the evaluation of input-text we could either see this as a situation where a) the overarching word can be seen as a multiplier [9] to the other words – increasing their overall polarities (Expressed in Example 1) or b) the polarities must either cancel-out one another – slightly neutralising or balancing the overall polarity (expressed in Example 2). Lastly, some sentences or statements within a text-input can have uneven weighting based on their position relative to others. Sentences encased in brackets can be seen as less important/extra information, however, they may also contain potentially extreme polar terms that should not have the same weight as those outside brackets. Such an approach to weighing words in this circumstance is explained in section VI.

IV. CONVENTIONAL SENTIMENT ANALYSIS

Broadly speaking, sentiment analysis is "the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes" [10]. Conventional sentiment analysis, from a text point-of-view, involves a broad range of techniques already widely used in sentiment analysis. These can be broken down into lexicon-based, machine learning-based and a hybrid of both. Methods involving lexicons (either generated with training data through machine-learning or manually annotated) tend to assess the polarity of documents on a per-word basis. The structure of these lexicons are usually split into positive and negative strands and in some cases, such as SocialSent [11], will weight both their positive and negative ratings for each entry with negatively-polar words being (< 0) and positively-polar words being (> 0) (show in Table I). In naïve approaches, words are assigned equal *"mass"* meaning that each word has an equal amount of overall effect on the polarity of the entire input. This method of sentiment analysis, on a small scale, (i.e. audience-film-reviews), can prove to be effective as they are usually not as large as, say, newspaper articles. Thus the equal *mass* that a word can have on the overall text is not as significant.

TABLE I
EXCERPT FROM SOCIALSENT'S [11] LEXICON

Word	Polarity	Standard Deviation
monster	-2.82	0.87
...	...	...
wonderful	2.76	0.71

A. Machine Learning Approaches for Conventional Methods

This section aims to cover the conventional machine learning concepts used in conventional sentiment analysis as well as to further assess their effectiveness, advantages and dissadvantages.

1) Bag-of-Words: The bag-of-words model, in a text classification sense, simplifies a given text into a bag or multi-set such that for a given text:

"Neill likes the colour red, Alex likes red too."

A bag-of-words representation can look like (Using JSON notation):

{Neill:1, likes:2, the:1, colour:1, red:2, Alex:1, too:1}

Whereby the multiplicity of the words/features *red* and *likes* is 2 and the multiplicity of the other words are 1. Common words or stop-words, such as *it*, *and*, *if*, or *a* may have much larger multiplicities and frequencies. These, therefore, be normalized using Term Frequency-Inverse Document Frequency (tf-idf). Weights generated in tf-idf take the bag-of-words model output and invert their influence. This allows for less-frequent words (that are potentially more relevant) to stand out, rather than the more common words taking control of the classification.

2) Naïve Bayes: Naïve Bayes is another method that is commonly used in text classification, whereby we classify data in a way that allows us to look at conditional probability in such a way that enables us to flip the condition in a convenient way. This can also be expressed as "event X will

happen given evidence Y" [12]. In essence, the Naïve Bayes method allows us to estimate the probability of something based on its occurrence. This classifier, as the title suggests, is a particularly innocent approach as the classifier makes a large assumption on the polarity of the input. Furthermore, the Naïve Bayesian implies that each word has an equal effect on the polarity of the input - therefore, each word is seen independently of one another, and in turn, no link can be established between them. The authors of [13], also employ the use of Naïve Bayes when analysing sentiment in tweets. However, the authors filter the tweets using common emoticons, e.g. ":-), :) and =)" for positive tweet data, and ":- (, :(, =(and ;(" for negative tweet data – to which the authors of [12] claim to be, in of itself, "a naïve approach to collect and classify 300000 tweets into three categories". In order to get objective tweets, and further, neutral data, the authors of [13] filters the tweets for newspaper publisher accounts, in this case, The New York Times. From this, they can train their Naïve Bayes classifier. The authors firstly pair the Naïve Bayes approach with classical n-grams for their sentiment analysis model, however, later claim that to increase accuracy, "they should discard common n-grams, as n-grams do not strongly indicate any sentiment nor indicate objectivity of a sentence". The problem expressed in the Naïve Bayes classifier, where words are independent to other words, is part of the main focus of the intended further work to be carried out after this survey, where we will explore more into how discourse structure (explained in Section 2) has an overall effect on the polarity of the input, when further combined with current conventional approaches, to produce a novel method to sentiment analysis. (explained more in the Further Work section)

3) Support Vector Machines (SVM): Support Vector Machines are used in machine learning for classifying data in n-dimensional space. An SVM model produces a plane across such space - dividing the dataset into each class. SVM models are considered to be fairly intensive across large datasets and extra variables need to be considered to optimise and improve the accuracy of the model. Firstly, a gamma variable is introduced. With a low Gamma variable, large amounts of data in the set are considered when plotting the plane, whereas, with a high Gamma, fewer plots are considered. Another key variable, when considering the accuracy is the Margin. This is the measurement of distance from the plane to the points on either side. Ideally, this measurement needs to be similar across each side of the plane. Too much to one or the other side can produce an unwanted bias towards one of the classifications. In terms of sentiment analysis, SVMs are considered to be an optimal solution to sentiment analysis, however as we find later, the study [14] disproves this in their experiment when comparing SVMs, with Naïve Bayes' and Maximum Entropy. The authors of [15] use SVMs to classify their dataset for sentiment analysis across review data from [16], [17] corpus further stating that it is capable of classifying a large amount of features in the text. The authors then combine this with n-grams for the input of the classifier

– finally stating that unigrams (n=1) produced the best result. SVMs are also appropriated in [14] in their analysis regarding tweets, whereby they compare it with a Naïve Bayes and Maximum Entropy to solve the classification problem, as well as train their model. Once they have finished their training and classification, they then employ semantic analysis. This is derived from the WordNet database which defines words that are linked together, whereby words that are closer together are considered "semantically similar". The authors further add that they can determine synonym-like similarities – claiming that sentences like "I am happy", where the adjective, "happy" is selected and "compared with stored feature vectors for synonyms" of that word. The authors further by stating that the words, "glad" and "satisfied" can be selected producing a similar, positive polarity. Finally adding that the accuracy could be improved with a larger dataset.

4) Maximum Entropy: Maximum entropy is the method of classifying text whereby the "probability that a document belongs to a particular class given a context must maximize the entropy of the classification system" [18]. It is also stated in [19], more clearly, "Maximum Entropy models offer a clean way to combine diverse pieces of contextual evidence to estimate the probability of a certain linguistic class occurring with a certain linguistic context". The use of Maximum Entropy in [20] showed promising results when compared to the aforementioned methods (Naïve Bayes and SVM). Their results, displayed in table III, showed that Maximum Entropy - Implemented in their experiment using Maxent (an implementation of Maximum Entropy) out-performed the other methods above, when compared to a movie review corpus, however, we are presented with a contrast from [14] as they show in their results (table II) when comparing the approaches (using unigrams) that Maximum Entropy has the lowest accuracy of the 3 methods discussed (Naïve Bayes and SVMs). This may be due to the smaller corpus available to the study in [14] (the Twitter corpus) as the authors go on to confirm that, "The training data set can be increased to improve the feature vector related sentence identification process". Also, the dataset had been pre-processed differently, for example, sentences like *"that painting is Beauuuutifull"* are processed and converted to *"painting Beautiful"*. Perhaps if the focus was skewed to account for exaggerations in words – a more polar output would be obtained from that particular instance, as well as potentially taking faster as less time would be needed to pre-process the corpus. The dataset in [20] however, is taken from [16] movie review corpus and remains unprocessed throughout the experiment, perhaps leading to a slight change in accuracy across both experiments.

In [14] the methods described above are compared further, concluding that the Naïve Bayesian approach produced the most accurate results for classification in comparison with Maximum Entropy and Support Vector Machines (using a unigram model). However, semantic analysis using WordNet produced the best results overall, shown in Table II.

TABLE II
("Performance Comparison of Techniques in Terms of Accuracy" [14])

Method	Accuracy (Using Unigrams)
Naïve Bayes	88.20
Support Vector Machines	85.50
Maximum Entropy	83.80
Semantic Analysis (WordNet)	89.90

TABLE III
("Results in Terms of Accuracy on Move Review Corpus" [20])

Method	Unigrams	Unigrams and Subj. Anal.	Bigrams
NB	81.45	83.94	83.15
SVM	85.45	86.35	85.25
Max. Ent.	84.80	87.40	85.40

V. Evaluation of Conventional Sentiment Analysis

The above methods on sentiment analysis and text classifications offer many options when it comes to our later study on a combination of structure-based methods with conventional methods. Firstly, Support Vector Machines, when compared to Naïve Bayes and Maximum Entropy, performed exceptionally well in terms of accuracy when subject to unigrams. The results of [14] and [20] results (shown in Table II and III) reflected this, both agreeing that unigrams performed the best in this circumstance. Secondly, Support Vector Machines performed quite satisfactory, however, a problem that seems to plague a lot of the research was the size of the dataset being too small. For this method to be entirely ideal, the speed in which SVMs process the data would need to be faster than it currently is (as of 2020). Lastly, the Naïve Bayes method seems to be the best when it comes to defining a benchmark to test novel methods of sentiment analysis due to its aptly-named, naïve, method whereby, when combined with n-grams, the assumption on the polarity can be made quite quickly.

A. Advantages

- Widely available implementations for each machine-learning concept described;
- Large corpora and lexicons are freely-available for machine-learning training;
- Unsophisticated methods (involving the aggregation of words and their; predetermined polarity) can still be effective for small input-text such as *tweets* or user film-reviews.

B. Disadvantages

- for larger input-texts, unartful methods can make wide assumptions about polarity due to the amount of information in the text that can get disregarded, such as sentence structuring or scoping of particular keywords.

VI. Structure-Based Analysis

Another approach that is considered for sentiment analysis is the use of structure rather than per-words methods to decipher sentiment. Broadly speaking, we are retrieving the sentimental polarity of a given input-text from structure-based features. When analysing sentiment per-word, we tend to disregard the structure to streamline the data/text for sentiment mining later, however, a person's emotional state can have a large effect on the way someone constructs a sentence or paragraph. The subsequent sections aim to explore the methods of structure-based analysis as well as applications of such methods.

A. Methods of Structure-Based Analysis

1) Analysis of Sentence Structure-Based Features on Scientific-Paper Citations: The study in [21] firstly proposes their method of structure-based analysis on a newly made corpus built up from scientific paper citations. They intend to use a novel collaboration of features including n-grams, specialised science-specific lexical features, dependency relations sentence splitting and negation features, claiming that 3-grams and dependencies performed the best for this task. The authors represent their citations as a feature-set in an SVM. These features are as follows:

a) Word Level: Word Level Features where they use unigrams and bigrams as well as 3-grams as new features to capture longer technical terms. The authors also include a "science-specific sentiment lexicon" to the feature-set - consisting of 83 polar phrases extracted from 726 citations. These include phrases such as *efficient*, *popular* and *state-of-the-art*.

b) Contextual Polarity: Contextual Polarity Features are appropriated whereby sentence-based-features such as "subjectivity clues... along with the number of adjectives, adverbs pronouns modals and cardinals" are gathered. The authors then appropriate type-dependency structures [22], stating that this method describes the grammatical relationships between words. The authors propose capturing the long-distance relationships between words further stating that in the sentence *"<CIT> showed that the results for French-English were competitive to start-of-the-art alignment systems"*, the relationship between the words "results" and "competitive", will be captured with such dependency structures as opposed to trigrams, which the authors claim, will miss this relation.

This is done, in their case, by trimming the sentences' parse tree down – walking from the citation node (named <CIT>) to the root of the tree. Where the <CIT> token is defined using a regular expression that looks for authors' names. This is done to "remove any lexical bias associated with proper names of researchers". The final method that they use is negation (explained further below) through the attaching of negation notes to the clause subtree. The authors solve the negation-scoping issue (also explained further below) through the use a window-based method where any "words within a k-word range of any negation-term are suffixed with a token *_neg* to distinguish them from their non-polar versions". The study in [21] presents the example: "Turney's method did not work*_neg* well*_neg* although they reported 80% accuracy in <CIT>". Stating in their results, that "contextual polarity features do not
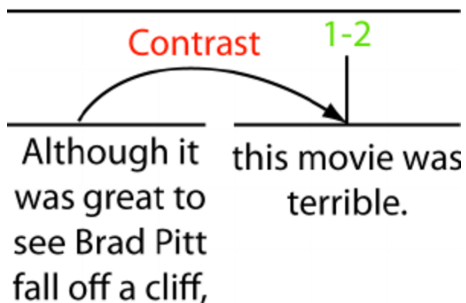
Fig. 1. Example RST structure

work well on citation text" and that also "adding a science-specific lexicon does not help either" - further claiming that n-grams are sufficient to capture discriminating lexical structures as well as stating that "word level and contextual features are surpassed by dependency features". Lastly, the use of the k-word window (k=15) for negation improved the performance, however, was not statistically significant, perhaps due to the skewed class distribution; the authors lastly proposing a larger dataset to amend this.

TABLE IV
([21], Results using science lexicon (scilex), contextual polarity (cpol), dependencies (dep), negation (neg), sentence splitting (split) and word-level (wlev) features.)

Feature	Macro-F	Micro-F
1 grams	0.581	0.863
1-2 grams	0.952	0.864
1-3 grams	0.597	0.859
" + POS	0.535	0.859
" + POS (tokenized)	0.596	0.859
" + scilex	0.597	0.860
" + wlev	0.535	0.859
" + cpol	0.418	0.859
" + dep	0.760	0.897
" + del + split + neg	0.683	0.872
" + dep + split	0.642	0.866
" + dep + neg	0.764	0.898

2) Discourse Structure and Rhetorical Structure Theory:
Discourse structure can be defined as a formal representation of a document or piece of text. In terms of Rhetorical-Structure-Theory [6], we appropriate discourse structure in such a way that allows us to produce a tree-like structure. Such structures allow us to further analyse the text.

a) Discourse Structure: Discourse structure In layman's terms, the discourse can be seen as a "sentence of sentences" [23]. The authors of [24] use this understanding to split documents into important and not important spans of text that are then used to define the polarity of the entire document. The document is split through the use of a discourse theory called Rhetorical Structure Theory (RST) whereby the document can be processed in such a way that the sentences can be grouped by rhetorical relativity to one another. The authors suggest that this will lead to "determining the parts of the text that are most relevant to the overall document sentiment". The text is split and ordered into segments or units with different

levels of importance. The authors state that this will allow for sentiment analysis algorithms to work more reliably when "weighting sentiment of parts of a text in accordance with their associated impact on the document's sentiment". They explain further that the theory works by twofold, hypotactic and paratactic, relations. Hypotactic, where one span of text is "classified as the nucleus, whereas the other spans are classified as a satellite". The authors state that RST claims the nuclei to be more significant to the overall understanding of the text. In paratactic relation, spans are considered to be equally significant meaning that all spans are classified as the nuclei. The author continues by explaining that the smallest text spans can hold rhetorical relations as Elementary Discourse Units (EDU's). These can be joined together to create new text spans which can also hold rhetorical relations to other spans creating a hierarchical structure, continuing to say that RST also defines several types of relations "(e,g., elaboration, attribution, contrast, etc.)." which in their framework carry weights to determine the polarity the overall document. Figure 1 explains, further, the structure mentioned above whereby the sentences "Although it was great to see Brad Pitt fall off a cliff," and "the movie was terrible", where the former can be seen as the nucleus and the latter the satellite, provide a contrasting review (one of the several relations defined by RST) as the satellite contrasts with the nucleus. Contrasting with the classical sentiment analysis technique, where all words can have an equal effect on the polarity of the sentence. The author further explains that the word "great" would cancel out with the word "terrible" in classical conventional sentiment analysis – creating a possible overall neutral review. In further contrast with the proposed method whereby, based on the structure of the sentence, we can weigh each span by how relevant it is to the final polarity of the document. The authors of [25] also appropriate the study of Discourse Structures in their study where they intend to teach Deep Neural Networks from Discourse Trees to develop a discourse-aware method for sentiment analysis "that can recognize differences in salience between individual subordinate clauses". Their method involves the representation of the semantic structure of a document in the form of a binary tree, or hierarchical discourse tree. The scores for the polarity are then gathered for each leaf of the binary tree through the use of polarity dictionaries and word embeddings. This tree is then traversed by their specific neural network, called a Discourse-LSTM. From this traversal, they are then able to aggregate the score based on the discourse structure, to finally compute a score for the entire document. The authors compares their model with the naïve bag-of-Words approach, which is widely considered the general baseline for most sentiment analysis. Their baseline model counts term-frequencies in the document, to produce a document-term-matrix where the term frequencies are scaled using tf-idf – placing stronger weights on characteristic terms. As explained before, and further reiterated in [25], the benefit of this particular, proposed, model allows for clauses to be weighted based on their overall significance to the entire document - finally going on to say, that compared to their

chosen, unsophisticated, benchmark, Discourse-LSTM outperforms when distinguishing polarity in text.

VII. EVALUATION OF STRUCTURE-BASED SENTIMENT ANALYSIS

Overall, the use of discourse structures provides a fresh perspective, from a sentiment analysis point of a view, on how we can elicit emotional polarity from the text. More specifically allowing one to see sentiment beyond a per-word-polarity (Dictionary). Traditional methods such as the ones mentioned above all tend to disregard structure when the text is processed and instead only look at words directly next to them to draw some link between one another. The use of discourse analysis could aid us in varying the weights that words have on entire input-texts. It is innocent to assume that words within potentially irrelevant parts of text should be weighted in the same way as other potentially more important parts of the text.

A. Advantages

- As stated in [26], "a person's speaking or writing style is akin to personality", further reiterated in [27], who has shown that the mental state, *more specifically* a person's depression, neuroticism and narcissism had a direct impact on their writing style. From a sentiment analysis point-of-view, this kind of clarification on sentiment-driven writing style and structure, can aid our understanding and development in the structure-based analysis as well as enriching the current understanding of sentiment analysis on a whole.

B. Disadvantages

- More involved structure-based methods, though they provide a larger amount of detail, can impose higher computation times. This can result in less-efficient processing of corpora of *tweets* as they are only ever going to be less than 280 characters;
- Writing structure is still quite ambiguous in terms of its link between emotion and writing style. Opinions may sometimes be included in the satellites of appropriate nuclei, rather than the core sentence itself, usually in the form of an elaboration relation [28]. This type of structural-ambiguity can pose a large problem to solve in discourse-parsers as not everyone may structure their opinions in such away.

VIII. AMBIGUOUS SENTIMENT ANALYSIS METHODS

It is not always clear-cut whether a method of defining sentiment is directly structure-based or conventional. Furthermore, defining whether the method is sentence or document-based is also shrouded in ambiguity. Some methods, namely, negation have some structural elements present in their core functionality (word-scoping) when they are used to classify the polarity.

A. Negation

Negation is an element of linguistics whereby prefixes such as *not, non, etc.* directly change the polarity of the next-occurring word. For sentiment analysis, negation must be taken into account as largely polar words such as *Amazing* and *Terrible* can be prefixed with negations such as *not, wasn't, etc*, which directly affect the overall polarity. It is explained in [29] that it is important we know what words are affected (scope) by the negation-term as these words will then affect the overall polarity. The author continues by stating that on a sentence-level analysis, the scope in which the negation-term covers is a key part of the analysis. Small clauses can be completely affected by single negation-terms, however, with larger sentences, containing >2 clauses, negation-terms may only affect some of the words. In this case, negation is implemented through syntactic-negation. When it comes to negation and how we can appropriate its rules into text classification, we must take into account both conjunction analysis as well as punctuation.

a) Conjunction Analysis: This involves the analysis of clause conjunctions such as *and, also, etc.*. These are useful when it comes to scoping in order to "identify the scope of negation in compound sentence". Furthermore, negation-terms in clauses joined by *but* prevent the negation-term from affecting the next clause in the sentence - only negating the first. However negation-terms joined by *and* may allow for the negation-term to affect the second clause also. This type of analysis can be seen as a structure-depended method as we are directly interacting with the placement of the clauses and how they are joined and affected by a word inside previous clauses.

b) Punctuation Marks: are used to limit the scope also. *Full-stops, exclamation marks, question marks, etc.* may restrict the scope of the negation-terms, however, if a comma is to come after an *and* or *or* then, they can be treated the same way as their pair conjunction [29]. Effective scoping and polarity-inversions are reiterated by the authors of [30] with the example:

"Not only is this phone expensive but it is also heavy and difficult to use."

The authors state that the negation-term should not affect the next part of this sentence (split with the conjunction *but*) and should also, in this case, not invert any other opinions or polar-words later on in this clause. They [30] Go to explain that negations are not always as explicit as *not, neither*, but in fact, can come in the form of a determiner such as *less* or *more*. For example:

"I believe the team performed less well^{+} to the other team"

This type of negation can be considered context-based, in this case, as we need to understand or quantify what we are comparing the subject to before we can make a definitive decision on the polarity.

B. Contrast Transition

As described in [31], contrast transitions can be seen as a type of negation-technique where words such as *however* and *but* can be seen to also invert upcoming clauses. For example:

*"I love cherries **however** bad they may seem"*

The example provided in [31],

"It could have been a great product. I dislike it, however."

demonstrates the complications that contrasting terms may have as well as the consequences of handling such a sentence with regular negation techniques. Furthermore, it is explained that the use of contrast transitions, along with negation methods, whereby any clauses containing contrast-transitions are removed to *aid* in negation [31]. The authors find, however, that this is clearly, ineffective. Therefore, a hybrid approach is considered between negation and contrast-transitions, where both transition and negations keywords, *however* and *but*, are used to aid in negation.

C. Intensifiers

Intensifiers, as described in [32] are a set of words used to amplify the polarity of an item. The position of these words, much like negation, is key to the overall polarity of its partnered word. These words are appropriated in the work of [9] on sentiment analysis whereby words such as *somewhat*, *pretty* (as an adverb) and *really* have an additive percentage increase on the word's polarity. For instance, given the intensifier *"really"* with a modifier of 10% and a polar word such as good, with the value +3, the polarity of the word *"good"* in the phrase *"really good"* will equal $3 + (3 \times 10\%) = 3.30$. For sentiment analysis, this type of method is key when further breaking down how someone may emphasise particular words. However, deciding whether or not to use alternate meaning to intensifiers such as *"pretty"* where the term has differing effects to its partnered word, for instance, *"pretty dress"* is much more positive to *"pretty good"*, can pose an issue for learning-based models as context-based classification methods will need to be in place to differentiate which word it is intensifying.

IX. Evaluation of Ambiguous Sentiment Analysis Methods

Ambiguous, hybrid methods allow us to not restrict our scoping or methods to just one feature of the text. Written text, especially those taken from movie-review corpora are biased and subjective as well as each being mostly written by different people. Methods involving word-based and structure-based features allow for a much clearer understanding as both are influenced by one another. Sentences can be written in different styles and use different words, however, they may have the same opinions, or in our case, sentiment to one another - therefore a combination of both features (word and structure) feels like the natural progression for sentiment analysis on text.

A. Advantages

- A broader understanding of the text and how it was composed can be deciphered allowing for a broader range of sentimental information to be elicited about either structure-based or word-based tropes.
- An understanding of a structural and word-based synergy may allow us to understand, from a machine-learning perspective, how we put together sentences, and sentiment-information elicitation can be proportionate to our emotional status when comprising the text.

B. Disadvantages

- Combining particular methods of word and structure-based understanding can widen the scope of analysis on a given input-text too much. As such, a broader search prevent smaller nuances from being considered. If they are, however, it could lead to much, much larger computational times.
- ambiguously defined intensifiers are potentially harmful to learning-based models – more specifically, exaggerations or sarcastic remarks made my the author of an input text.

X. Discussion and Opinions

Conventional methods of sentiment analysis can provide a substantial amount of accuracy from their simplistic approaches. For small corpora (Twitter data), conventional methods provide us with a considerable amount of detail as there is not much sentiment-heavy structure data within them (only spanning ≤ 2 sentences most of the time.) Using more complex, structure-based, analysis methods may only slow the processing of the input-corpora and may not produce as accurate of a result due to the potential lack of structural data present in the first place. A lot of noise in the input-text can be present - potentially hoodwinking the implementation. This is where RST shines as it allows us to abstract the document into a tree-structure with potentially more-relevant information closer to the root of the tree. This allows us to "prune" such tree, removing, potentially, irrelevant information. However, when it comes to larger input-text, structural information, such as positions of words; negation-terms; or, contextual data; can provide us with much more information regarding sentiment. The drawbacks to this method, however, is that small word-based nuance may be lost in the analysis which may have had large inflexions on the overall polarity.

A. Hybrid Methods

Its is clear to see that a culmination of both methods provides much more information as we can take into account both sets of features - word-based and structural. Furthermore, a combination of structure-based methods with conventional methods could be devised on the same corpus to create a novel method or workflow with the intent of providing a more accurate method of deciding the polarity of the text. More specifically, combining text discourse-structure sentiment analysis with multiple conventional sentiment analysis methods

and compare the results to determine the viability of the combination as well as then determine the best method, should the combination produce comparable results to the baselines of Naive Bayes. The combination will involve several models being implemented based on the above-discussed methods, and then subsequently compared to our novel set of potential models (the combination of RST and conventional classifiers). Exploring the addition of OCR techniques, expressed in [33] and [34], may also supplement the research and may provide some extra information regarding the more explicit, physical, structure of the document. We believe that structural analysis provides a promising approach to sentiment analysis – based on the underlying hypothesis and one potential research question to answer, after our study, *"Whether or not our emotional state can affect how we structure text when we create a tweet or a message from an NLP perspective?"*. This statement is partially proven already in [27] where neuroticism, narcissism and *most importantly* depression can affect the way in which we structure the text.

REFERENCES

[1] Z. Hailong, G. Wenyan, and J. Bo, "Machine learning and lexicon based methods for sentiment classification: A survey," in *2014 11th Web Information System and Application Conference*, pp. 262–265, Sep. 2014.

[2] O. Tas and F. Kiyani, "A survey automatic text summarization," *PressAcademia Procedia*, vol. 5, pp. 205 – 213, 2017.

[3] H. Su, "A study on long-short term memory networks with," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[4] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[5] B. Yu, "An evaluation of text classification methods for literary study," *Literary and Linguistic Computing*, vol. 23, pp. 327–343, 09 2008.

[6] W. C. Mann and S. A. Thompson, *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute, 1987.

[7] B. WEBBER, M. EGG, and V. KORDONI, "Discourse structure and language technology," *Natural Language Engineering*, vol. 18, no. 4, p. 437–490, 2012.

[8] B. Chardon, F. Benamara, Y. Mathieu, V. Popescu, and N. Asher, "Measuring the effect of discourse structure on sentiment analysis," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 25–37, Springer, 2013.

[9] M. Taboada, K. Voll, and J. Brooke, "Extracting sentiment as a function of discourse structure and topicality," *Simon Fraser Univeristy School of Computing Science Technical Report*, 2008.

[10] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, pp. 415–463, Springer, 2012.

[11] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, vol. 2016, p. 595, NIH Public Access, 2016.

[12] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of facebook statuses using naive bayes classifier for language learning," in *IISA 2013*, pp. 1–6, IEEE, 2013.

[13] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREc*, vol. 10, pp. 1320–1326, 2010.

[15] N. Zainuddin and A. Selamat, "Sentiment analysis using support vector machine," in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pp. 333–337, IEEE, 2014.

[14] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *2014 Seventh International Conference on Contemporary Computing (IC3)*, pp. 437–442, IEEE, 2014.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "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, 2002.

[17] M. Taboada, C. Anthony, and K. D. Voll, "Methods for creating semantic orientation dictionaries.," in *LREC*, pp. 427–432, 2006.

[18] N. Mehra, S. Khandelwal, and P. Patel, "Sentiment identification using maximum entropy analysis of movie reviews," *St anford Univer sity, USA in*, 2002.

[19] A. Ratnaparkhi, "A simple introduction to maximum entropy models for natural language processing," *IRCS Technical Reports Series*, p. 81, 1997.

[20] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, "Automatic sentiment analysis in on-line text.," in *ELPUB*, pp. 349–360, 2007.

[21] A. Athar, "Sentiment analysis of citations using sentence structure-based features," in *Proceedings of the ACL 2011 student session*, pp. 81–87, Association for Computational Linguistics, 2011.

[22] M.-C. De Marneffe and C. D. Manning, "The stanford typed dependencies representation," in *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pp. 1–8, Association for Computational Linguistics, 2008.

[23] B. Webber and R. Prasad, "Discourse structure: Swings and roundabouts," *Oslo Studies in Language*, vol. 1, no. 1, 2009.

[24] B. Heerschop, F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong, "Polarity analysis of texts using discourse structure," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1061–1070, ACM, 2011.

[25] M. Kraus and S. Feuerriegel, "Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees," *Expert Systems with Applications*, vol. 118, pp. 65–79, 2019.

[26] M. E. Ireland and J. W. Pennebaker, "Language style matching in writing: Synchrony in essays, correspondence, and poetry.," *Journal of personality and social psychology*, vol. 99, no. 3, p. 549, 2010.

[27] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[28] J. M. Chenlo, A. Hogenboom, and D. E. Losada, "Sentiment-based ranking of blog posts using rhetorical structure theory," in *Natural Language Processing and Information Systems* (E. Métais, F. Meziane, M. Saraee, V. Sugumaran, and S. Vadera, eds.), (Berlin, Heidelberg), pp. 13–24, Springer Berlin Heidelberg, 2013.

[29] U. Farooq, H. Mansoor, A. Nongaillard, Y. Ouzrout, and M. A. Qadir, "Negation handling in sentiment analysis at sentence level.," *JCP*, vol. 12, no. 5, pp. 470–478, 2017.

[30] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, "A survey on the role of negation in sentiment analysis," in *Proceedings of the workshop on negation and speculation in natural language processing*, pp. 60–68, 2010.

[31] S. Li and C.-R. Huang, "Sentiment classification considering negation and contrast transition," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pp. 307–316, 2009.

[32] R. Quirk, *A comprehensive grammar of the English language*. Longman, 1985.

[33] G. Maderlechner, P. Suda, and T. Brückner, "Classification of documents by form and content," *Pattern Recognition Letters*, vol. 18, no. 11-13, pp. 1225–1231, 1997.

[34] N. Chen and D. Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10, no. 1, pp. 1–16, 2007.