Combining Stylometric and Sentiment Mining Approaches for Deceptive Opinion Spam Detection

Alibek Jakupov, Besma Zeddini, Julien Longhi, Julien Mercadal

Combining Stylometric and Sentiment Mining Approaches for Deceptive Opinion Spam Detection

Alibek Jakupov
SATIE Laboratory CNRS – UMR 8029
CY Tech, CY Cergy Paris University
Cergy, France
ajakupov@expertime.com

Besma Zeddini
SATIE Laboratory CNRS – UMR 8029
CY Tech, CY Cergy Paris University
Cergy, France
besma.zeddini@cyu.fr

Julien Longhi
AGORA Laboratory EA 7392
CY Tech, CY Cergy Paris University
Cergy, France
julien.longhi@cyu.fr

Julien Mercadal
CY Tech, CY Cergy Paris University
Cergy, France
jml@cy-tech.fr

***Abstract*—Current methods for detecting deception automatically tend to focus on specific words or abstract linguistic features that may not be rooted in the psychology of deception. While these methods can perform well when training and testing data have similar content, they struggle when the content changes. In this study, we explore new features that capture stylistic patterns and sentiments, which are psychologically relevant aspects of truthful and deceptive language that may be useful across different content domains. To assess these features' potential, we test them on various datasets containing deceptive language, such as hotel, restaurant, and doctor reviews. We also evaluate these features within a deception detection classifier. Our findings show that sentiment-based features are most effective for cross-domain deception detection when the content of training and testing data significantly differs, and they especially improve classification accuracy on deceptive documents. The results have implications for developing general-purpose deception detection approaches.**

***Index Terms*—feature engineering, deceptive opinion spam, natural language processing, stylometry, sentiment analysis**

I. Introduction

Due to the rising popularity of online reviews, there has been an explosion in web authorship from individuals, some of which may include fraudulent reviews or Opinion Spam. Opinion Spam is inappropriate or fraudulent reviews which can range from self-promotion of an unrelated website or blog to deliberate review fraud with a potential for monetary gain [1]. Companies are highly motivated to automatically detect and remove Opinion Spam because one of the main risks of Opinion Spam in terms of its impact on customer opinion mainly concerns the reviews that falsely praise inferior products or criticize superior products [2]. When compared to other Natural Language Processing (NLP) tasks, such as sentiment analysis or intent detection, detecting Opinion Spam using text classification approaches has received very little research [3]. A human reader can quickly recognize some sorts of opinion spam, such as advertisements, questions or other non-opinion texts [4]. These instances fall under the category of Disruptive Opinion Spam, which consists of unrelated remarks that are easily noticeable by readers and present a low level of risk, as users have the option to disregard them [1]. However, in the case of more subtle forms of fake content, such as Deceptive Opinion Spam, the challenge is not straightforward, as these statements are deliberately crafted to appear genuine and deceive the evaluator [1]. Deceptive Opinion Spam commonly takes the form of fictitious reviews (positive or negative) posted by a malicious internet user to inflate or hurt an enterprise's image [3]. As these statements have been intentionally written to mislead the reader, human reviewers are faring little better than a chance in detecting these deceptive texts [2]. Consequently, there is a dire necessity to tackle this matter as identifying textual patterns in deceptive texts and obtaining significant substructures continues to be a challenging task. [3].

The problem is commonly addressed as a task of classifying text. In the majority of cases, text classification systems are composed of two main components: a vectorization module and a classifier. The former is responsible for generating features from a given text sequence, while the latter assigns class labels to this sequence based on a list of corresponding features. The features are commonly grouped into lexical and syntactic. For instance, such measures as total words or characters per word, frequency of large and unique words belong to the lexical group, whereas syntactic features are mainly composed of frequency of function words or word combinations, like bag-of-words (BOW), n-grams, or Parts-Of-Speech (POS) tagging [5]. Besides lexical and syntactic features, there also exist lexicon containment techniques which represent the occurrence of a term from the lexicon within a text as a binary value (positive signifying presence, and negative indicating absence) [6]. The lexicons for such kind of features are constructed by human expert [7, 8] or generated automatically [9]. Certain methodologies propose the incorporation of morphological connections and dependent linguistic elements present within the text as input vectors for the classification algorithm [10]. In addition to this, there are semantic vector space models, which serve to characterize each word through a real-valued vector, determined by the distance or angle between pairs of word vectors [11]. In the

field of automatic deception detection various approaches have been applied, mostly relying on linguistic features, such as n-grams [1, 12, 13], discourse structure [14, 15], semantically-related keyword lists [16, 17], measures of syntactic complexity [17], stylometric features [16], psychologically-motivated keyword lists [18] and parts of speech [19, 20].

These vectorization techniques are generally served to analyse the weights of the features, which allows to shed light on the common patterns in the structure of deceptive statements that is less present in truthful texts. While this methodology demonstrates certain efficiency, it has notable disadvantages due to the challenge in regulating the quality of the training set. For instance, while most of the classification algorithms, trained with this approach, exhibit satisfactory performance within their specific domains, they encounter difficulties in effectively generalizing across varying domains, thus lacking robustness in adapting to domain alterations [21]. As an illustration, a mere alteration in the polarity of hotel evaluations (that is, training the model on positive reviews while testing it on negative ones) has the potential to significantly reduce the F-score [22]. This observation holds when the training and the testing dataset originate from different domains [12]. Moreover, certain classification models based on the semantic vector space models may be strongly impacted by social or personal attitudes present in the training data, which makes the algorithm draw wrong conclusions [23]. Furthermore, certain researches suggest that deceptive statements differ from truthful ones more in terms of their sentiment then other linguistic features [24]. According to them in certain cases the deceivers display more positive affect in order to mislead the audience [25] whereas certain instances demonstrate that deception is characterized by more words reflecting negative emotion [24].

Based on the evidence mentioned above, it can be inferred that feature extraction methodologies utilized in classical NLP tasks exhibit limited reliability when applied to Deceptive Opinion Spam. This is primarily due to their strong association with particular lexical elements (like n-grams and specific keywords) or linguistically abstract components that may not be directly influenced by the style of verbal deception (such as specific parts of speech, stylometric features, syntactic rules) [2]. From this point of view it is more favorable to develop a novel set of features based on domain-independent approaches like sentiment analysis or stylometric features, as it offers superior generalization capabilities and independence from the training dataset domain.

Our approach

In this paper, we explore the effectiveness of a novel linguistically-defined implementation of stylometric and sentiment-based features for automatic deception detection. We begin by examining prior approaches to automatic deception detection, emphasizing techniques that employ linguistic features such as n-grams, which provide the best performance within the domain. Subsequently, we provide an overview of the varied databases utilized to assess our methodology and its effectiveness across different domains. We then delve into the proposed sentiment-based features, validating their potential value in detecting deception within these corpora. We also investigate the stylometric features and diagnostic potential of non-functional words but without incorporating them into classifier. Finally, we describe our classification scheme, which leverages these features.

Our contributions

Our contributions can be summarized as follows.

- Novel approach to automatic deception detection that applies sentiment-based features
- Comprehensive analysis of previous approaches to deception detection, highlighting the strengths and weaknesses of different techniques and emphasizing the importance of linguistic features
- Demonstration of the effectiveness of our approach using diverse corpora, showcasing its potential for cross-domain performance.
- Investigation of the diagnostic potential of non-functional words as stylometric features

Outline: The rest of the paper is organized as follows: in Section II we provide an overview of relevant research and studies; in Section III we summarize our methodology for forensic investigation; in Section IV we present and discuss experimental results as well as the datasets used to benchmark our approaches; finally, conclusions and discussions are provided in Section V.

II. Related Work

Ott *et al.* were the first to study this problem with the help of the Machine Learning approach [1]. A significant contribution of their research is the proof of the necessity of considering both the context and motivations underlying a deception, rather than solely focusing on a pre-defined set of deceptive indicators, such as Linguistic Inquiry and Word Count (LIWC), which is extensively applied to investigate personality traits [26] and study tutoring dynamics [27]. Accordingly, they combined the common text vectorization techniques with psycho-linguistic features, and succeeded to achieve the 89.8% performance with the model based on LIWC and bigrams. Nonetheless, evidence has demonstrated that these attributes lack robustness in response to topic change, as they can do well only if training and testing datasets are of the same domain [21, 2]. Moreover as all these methods were assessed within a single domain, both the training and testing sets were from the same subject area. As a result, it remains uncertain how well the performance can be generalized across different domains, especially when the classifier has access to features defined by specific linguistic details. Approach based on sentiment analysis, in this context, is more flexible as it allows extracting the deceptive patterns independently from the domain. This is because the sentiment-based features are thought to correlate with psychological mechanisms underlying the generation of fake reviews, such as the specificity

of the memory trace the deceiver is relying on and strategic avoidance of potentially verifiable information [2].

Li *et al.* achieved a score of 81.8% on Ott dataset by capturing the overall dissimilarities between truthful and deceptive texts [20]. In their research they extended Sparse Additive Generative Model (SAGE), a generative bayesian approach, that integrates topic models and generalized additive models. This results in the creation of multifaceted latent variable models through the summation of component vectors. Due to the fact that the majority of research in this field concentrates on identifying deceptive patterns rather than training a single reliable classifier, the primary challenge of the research was to determine the features that have the greatest influence on each category of deceptive review. Furthermore, it was necessary to evaluate the impact of these features on the final decision when combined with other features. SAGE is well-suited to address these requirements due to its additive nature, whereas other classifiers may encounter difficulties when dealing with domain-specific attributes in cross-domain scenarios. According to the authors' findings, the BOW approach was less robust than LIWC and POS modeled using SAGE, therefore they constructed the general rule of deceptive opinion spam with these domain-independent features. Moreover, unlike Ott *et al.* [1], who considered the absence of spatial data in hotel reviews as a clue to detect deceptive patterns, Li *et al.* demonstrated that this approach may not be universally applicable, as some fraudulent reviews may be composed by domain experts. Despite the fact that the domain-independent features constructed during the research demonstrated to be effective and enabled the identification of fake reviews with above-chance accuracy, it has been proven that the sparsity of these features makes it complex to leverage non-local discourse structures [28], thus the trained model will be incapable of capturing the overall semantic information of a document. Additionally, based on the results of their investigation, another noteworthy indicator of fraudulent statements is the presence of sentiments, as reviewers tend to exaggerate sentiment by using more sentiment-related vocabulary in their statements.

Ren and Ji [28] extended the previous research by proposing a three-stage system. At first, they utilized a convolutional neural network to construct sentence representations from word representations, as the convolution action has been widely employed to synthesize lexical n-gram information. For this step they utilized three convolutional filters as they are capable of capturing local semantics of n-grams, such unigrams, bigrams and trigrams, a method that has already been demonstrated to be successful for such tasks as sentiment classification [8]. Afterwards, they modeled the semantic and discourse relations of these sentence vectors to construct a document representation using a bi-directional gated recurrent neural network. These document vectors are finally used as features to train a classifier. The authors achieved 85.7% accuracy on the dataset created by Li *et al.* and demonstrated that neural networks can be employed to acquire continuous document representations to better capture semantic characteristics. The main goal of this study was to empirically demonstrate the better performance of neural features over traditional discrete features (like n-grams, POS, LIWC, etc.) due to their stronger generalization. However, additional experiments conducted by the authors revealed that by integrating discrete and neural features the overall accuracy may be improved, thus discrete features, like sentiment combination or non-functional word usage, still remain a rich source of statistical and semantic information.

Vogler and Pearl [2] explored the application of specific details for detecting deception both within a single domain and across multiple domains. The linguistic features they examined in the study were comprised of n-grams, POS, measures of syntactic complexity, syntactic structure, semantically-related keyword lists, stylometric features, psychologically-motivated keyword lists, discourse structure and named entities. The authors concluded that these features were not sufficiently robust, not sufficiently resilient, particularly in situations where the domain may differ significantly, as most of them tend to depend on clues that are very reliant on specific lexical items, such as n-grams or specific keyword lists. Though there are some linguistically abstract features like stylometric features, POS or syntactic rules, the authors regard them as less pertinent since they are not motivated by the psychology of verbal deception. In their study, deception was viewed as an act of imagination, therefore besides analyzing the linguistic approaches the authors also investigated methods that were motivated by psychological factors, such as information management theory [29], information manipulation theory [30], reality monitoring and criteria-based statement analysis [2]. Since more abstract linguistic cues motivated by psychology may have wider applicability across various domains [31] the authors consider it advantageous to apply these cues with a basis in the psychological theories of how humans produce deceptive statements. They have also relied on the findings presented by Krüger *et al.* whose research centers around the detection of subjectivity in newspaper articles and suggests that linguistically abstract features may exhibit greater resilience when utilized on texts from various domains [21]. For the experimentation Vogler and Pearl utilized three datasets for training and testing with domain shifts ranging from relatively subtle to considerably extensive, the Ott Deceptive Opinion Spam Corpus [1], essays on emotionally charged topics [12] and personal interview questions [29]. The linguistically-defined specific detail features the authors constructed for this research were demonstrated to be effective when there were significant variations in the domains used for training and testing. The features were based on proper nouns, adjective phrases, prepositional phrase modifiers, exact number words and noun modifiers that appeared as consecutive sequences. Each feature is represented as the total normalized number and the average normalized weight. They succeeded to achieve the best F-score of 0.91 for cases in which there was no change in content and the best F-score of 0.64 for cases in which there was a substantial shift in domain, which demonstrates that the linguistically-defined specific detail features are

more generalizable across domains. Despite the fact that the classifier trained on these features had fewer false negative, it poorly classified the truthful texts. As is evident from the results of the experiments, a mix of n-gram and linguistically-defined specific details features tends to be more reliable only in case the false positive is more costly than false negative. It should also be mentioned that the n-gram-based features may possess a greater capacity for semantic generalization when based on distributed meaning representations, such as GloVe and ELMo, whereas n-gram features in their approach are based solely on individual words and do not capture the semantic relatedness between them. This is in contrast to our approach, as our proposed method involves examining the semantic content of statements through an assessment of the overarching sentiment.

Barsever *et al.* constructed a state-of-the art classifier with the help of BERT and then examined this classifier to detect the patterns BERT learned to distinguish the deceptive reviews [3]. BERT is a neural network architecture that has been pre-trained on millions of words and utilizing the Masked Language Modeling (MLM) by jointly conditioning on left and right context in all layers to train deep bidirectional language encoding [32]. The primary advantage of BERT lies in its ability to learn rules and features in an unsupervised manner, which allows BERT looking for the best solution unrestricted by preconceived rules. With their model, Barsever *et al.* achieved an accuracy of 93.6%, which proves the existence of features allowing to distinguish truth and deception. Given that the primary aim of the study was to uncover rules and patterns of deceptive language, the authors conducted an ablation study, wherein they removed each part-of-speech (POS) one at a time and monitored the network's performance. Additionally, the researchers identified what are known as 'swing' sentences, which hold greater importance for the classifier than other sentences within the text, to run POS analysis on them and shed light on the inner rules BERT constructed. Finally, based on their BERT model, the authors developed a Generative Adversarial Network (GAN), whose objective is to deceive the classifier and uncover the trends replicated in the synthetic data. Their findings suggest that particular POS, such as singular nouns, hold greater significance for the classifier than others. Moreover, the research shows that truthful texts exhibit greater variance in terms of POS, whereas deceptive reviews tend to follow a more formulaic pattern. Nevertheless, the approach applied by Barsever *et al.* may present significant challenges. In fact, one of the primary drawbacks of BERT is the lack of separate sentence embeddings, which can play an important role as a higher means of abstraction. Not surprisingly, the authors had to manually eliminate sentences from the original dataset by replacing them with [MASK] tokens, and exclude the entries comprising only one sentence. In addition, the rules generated by BERT are still not entirely clear to the authors, and the results of the ablation study may uncover other similarities rather than accurately identifying the patterns of deception. For instance, the removal of singular nouns resulted in a significant decline in the performance of the mode, which is interpreted as a strong weight of this POS in the classifier. Nevertheless, based on these results, it can be also inferred that due to the prevalence of nouns in natural language, their replacement may result in text that is difficult to comprehend and interpret by the classifier. In this context, the sentiment vector is much easier to reason about, due to its sparsity.

III. Model

This section describes the methodology for our approach.

Stylometric Approach

Stylometry is a quantitative study of literary style that employs computational distant reading methods to analyze authorship. It is based on the observation that authors have relatively consistent, recognizable, and unique writing styles. These unique styles are evident in various aspects of writing, such as vocabulary, sentence structure, punctuation usage, and the use of small function words like articles, prepositions, and conjunctions. The unconscious and topic-independent nature of function words makes them particularly useful for stylometric analysis.

In our study, we explore the application of stylometric analysis in detecting deceptive opinion spam, focusing on the unique linguistic patterns that can differentiate between truthful and deceptive texts. By examining various stylometric features, we aimed to uncover the underlying characteristics of deceptive language and develop a reliable method for identifying deceptive opinion spam.

For our experiments we applied the Burrows' Delta method, a measure of the "distance" between a text whose authorship we want to ascertain and some other corpus. Unlike other methods like Kilgariff's chi-squared, the Delta Method is designed to compare an anonymous text (or set of texts) to many different authors' signatures at the same time. More precisely, Delta measures how the anonymous text and sets of texts written by an arbitrary number of known authors all diverge from the average of all of them put together. Furthermore, the Delta Method gives equal weight to every feature that it measures, thus avoiding the problem of common words overwhelming the results, which was an issue with chi-squared tests. For all of these reasons, John Burrows' Delta Method is usually a more effective solution to the problem of authorship. We adjust this approach to identify how non-functional words are used by deceivers and regular internet users. As the features extracted by this approach are topic-independent, this allows us to construct a model which is robust to the domain change.

Our adaptation of Burrows' original algorithm can be summarized as follows:

- Assemble a large corpus made up of texts written by an arbitrary number of classes; let's say that number of classes is x (deceptive and truthful).
- Find the n most frequent words in the corpus to use as features.

- For each of these n features, calculate the share of each of the x classes' subcorpora represented by this feature, as a percentage of the total number of words. As an example, the word "the" may represent 4.72% of the words in deceptive's subcorpus.
- Then, calculate the mean and the standard deviation of these x values and use them as the offical mean and standard deviation for this feature over the whole corpus. In other words, we will used a mean of means instead of calculating a single value representing the share of the entire corpus represented by each word. This is because we want to avoid a larger subcorpus over-influencing the results in its favor and defining the corpus norm in such a way that everything would be expected to look like it.
- For each of the n features and x subcorpora, calculate a z-score describing how far away from the corpus norm the usage of this particular feature in this particular subcorpus happens to be. To do this, subtract the "mean of means" for the feature from the feature's frequency in the subcorpus and divide the result by the feature's standard deviation. Below is the z-score equation for feature 'i', where C(i) represents the observed frequency, the μ represents the mean of means, and the σ, the standard deviation.

$$Z_i = \frac{C_i - \mu_i}{\sigma_i} \tag{1}$$

- Subsequently, compute the same z-scores for each feature in the text for which authorship is to be determined.
- Lastly, calculate a delta score comparing the anonymous text with each candidate's subcorpus. This can be achieved by averaging the absolute values of the differences between the z-scores for each feature in both the anonymous text and the candidate's subcorpus. This process ensures that equal weight is given to each feature, regardless of the frequency of words in the texts, preventing the top 3 or 4 features from overwhelming the others. Below formula presents the equation for Delta, where Z(c,i) represents the z-score for feature 'i' in candidate 'c', and Z(t,i) denotes the z-score for feature 'i' in the test case.

$$\Delta_c = \sum_i \frac{Z_c(i) - Z_t(i)}{n} \tag{2}$$

The "winning" candidate, or the most likely class, is determined by identifying the one with the lowest delta score between their respective subcorpus and the test case. This signifies the smallest divergence in writing style, making them the most probable class (deceptive or truthful) of the text in question.

Sentiment Approach

In our methodology, we incorporated a measure of exaggeration, consistently applied across various domains. The underlying rationale posits that the sentiment intensity remains constant regardless of whether the text conveys a positive or negative sentiment (e.g., "I adore the product" and "I hate the product" represent equivalent degrees of sentiment, albeit in opposing directions). To scrutinize deceptive opinion spam, we employed Azure Text Analytics API [1], which facilitates the analysis of overall sentiment and extraction of three aspects: positive, negative, and neutral. This naturally resembled the RGB color model, prompting us to map the values accordingly: Negative to Red, Positive to Green, and Neutral to Blue. Subsequently, we visualized the emerging pattern.

To depict sentiment patterns in both deceptive and truthful reviews, we first employed colorization based on sentiment analysis results. Firstly, we transformed the sentiment scores (positive, negative, and neutral) into a blue-green-red (BGR) format, enabling each review to be represented as a pixel. Given that Azure Text Analytics provides percentages for each sentiment aspect (e.g., 80% positive, 15% neutral, and 5% negative), we multiplied these values by 255 to facilitate visualization. Next, we devised auxiliary functions to convert sentiment scores into pixel format and generate an image utilizing the BGR values.

Upon identifying visual patterns, we utilized these values as features for our classifier. To preclude the classifier from drawing erroneous conclusions by analyzing sentiments rather than exaggeration, we first ascertained the overall sentiment. If the sentiment was negative, we switched the green and red channels, since exaggeration remains consistent for both negative and positive sentiments. We then normalize this feature set, as in most of the cases the neutral aspect percentage is significantly higher than the other sentiments. Finally, we input these features into our classifier and examined the subsequent results as showin in 1.

Algorithm 1 Extract Sentiment Features

$features \leftarrow []$
for all $items \in Corpus$ **do**
  $sentiment \leftarrow mean(item.sentiments)$
  $aspect_{pos}, aspect_{neg}, aspect_{neut} \leftarrow item.sentiments$
  if $sentiment == Positive$ **then**
    $feature_r \leftarrow aspect_{neg} * 255$
    $feature_g \leftarrow aspect_{pos} * 255$
    $feature_b \leftarrow aspect_{neut} * 255$
  else
    $feature_r \leftarrow aspect_{pos} * 255$
    $feature_g \leftarrow aspect_{neg} * 255$
    $feature_b \leftarrow aspect_{neut} * 255$
  end if
  $feature \leftarrow (feature_r, feature_g, feature_b)$
  $feature \leftarrow normalize(feature)$
  $features \leftarrow feature$
end for

[1]https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/sentiment-opinion-mining/overview

IV. Experiments

Data

One of the first large-scale, publicly available datasets for the research in this domain is Ott Deceptive Opinion Spam corpus [1] composed of 400 truthful and 400 gold-standard deceptive reviews. To solicit these high-quality deceptive reviews using Amazon Mechanical Turk, a pool of 400 Human-Intelligence Tasks (HITs) has been created. These HITs have been then allocated across 20 chosen hotels. They have also ensured that opinions are written by unique reviewers, by allowing only a single submission per Turker. For truthful opinions they mined 6977 reviews from the 20 most popular Chicago hotels on Trip Advisor. With their dataset the authors have shown that the detection of deception is challenging for human judges, as most of them performed roughly at-chance.

For cross-domain investigation we applied a dataset consisting of hotel, restaurant, and doctor reviews [20], obtained from various sources, including TripAdvisor and Amazon. The deceptive reviews were primarily procured from two sources: professional content writers and participants from Amazon Mechanical Turk. This approach allowed the researchers to capture the nuances of deceptive opinions generated by both skilled and amateur writers. To ensure the quality and authenticity of truthful reviews, the authors relied on reviews with a high number of helpful votes from other users. This criterion established a baseline of credibility for the truthful reviews in the dataset. Furthermore, the dataset included reviews with varying sentiment polarities (positive and negative) to account for the sentiment intensity and exaggeration aspects in deceptive opinion spam.

Stylometric Approach

In this study, we integrated both datasets to investigate our hypothesis that the usage of non-functional words remains consistent across various domains. The combined dataset was divided into a 25% test set and a 75% training set, and the training set was used to evaluate the accuracy of correct identification. The results for the negative deceptive test showed a delta score of 1.3815 for deceptive and 1.8281 for truthful, while the negative truthful test had a delta score of 1.4276 for deceptive and 1.0704 for truthful. In the positive tests, the deceptive test had a delta score of 1.4003 for deceptive and 1.8459 for truthful, and the truthful test had a delta score of 2.9074 for deceptive and 2.2098 for truthful. In summary, the model accurately identified 65% of deceptive texts and 68% of truthful texts, considering both positive and negative cases.

In this study, we focused on examining the stylometric attributes and diagnostic potential of non-functional words, but opted not to incorporate them into the classifier due to the inherent methodological limitation that necessitates analyzing the entire corpus for vectorizing individual statements. Nevertheless, the findings unveil intriguing patterns that warrant further investigation.

Sentiment Approach

Before training the classifier, to visualize sentiment patterns in deceptive and truthful reviews, we first colorized the reviews by converting the sentiment scores (positive, negative, and neutral) to a blue-green-red (BGR) format. This allowed us to represent each review as a pixel, with blue representing neutral sentiment, green representing positive sentiment, and red representing negative sentiment. We then created helper functions to convert the sentiment scores into pixel format and generate an image from the BGR values. Each image consisted of 400 pixels (20x20), representing 400 reviews.

After generating images for different subsets of reviews (deceptive positive, deceptive negative, truthful positive, and truthful negative), we compared their patterns visually. The comparison revealed that negative deceptive reviews were brighter with fewer green spots, while positive deceptive reviews exhibited more vivid colors with fewer red spots. This indicates exaggeration in fake comments and false flattery in deceptive reviews. In contrast, truthful reviews appeared more realistic and balanced in their sentiment expression.

To obtain a uniform color representing deception, we averaged all the pixels in the images by splitting them into three channels (blue, green, and red) and calculating the average for each channel. We then merged the channels to create a single color representing the average sentiment of the deceptive reviews.

The analysis showed that truthful negative reviews were less red than deceptive negative reviews, while fake positive reviews were greener than truthful positive reviews. This suggests that deceptive reviews tend to exhibit more extreme sentiment expressions, which can be visualized through colorization.

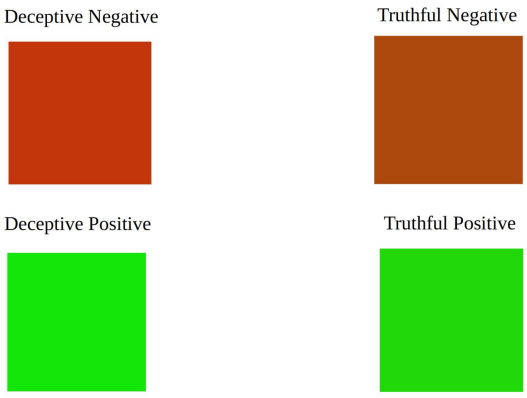

Fig. 1. Ott Deceprive dataset: colorized sentiments

With this in mind we trained multiple classifiers with features extracted using the algorithm 1. We used the Ott Deceptive Opinion Spam dataset for training and the cross-domain dataset constructed by Li *et al.* for testing.

Considering these factors, we employed various classifiers with features derived using the algorithm 1. For training

purposes, we utilized the the cross-domain dataset constructed by Li *et al.* , while the Ott Deceptive Opinion Spam dataset was employed for testing.

For this experiment, we implemented various normalization techniques, including MaxAbsScaler, StandardScaler Wrapper, and Sparse Normalizer, to ensure that the input features used in a machine learning model have a consistent scale or distribution. To evaluate the performance of our models, we employed the AUC Weighted as the primary metric. The choice of using AUC Weighted for model evaluation, as opposed to other metrics, stems from its ability to effectively measure the classifier's performance across different thresholds, while accounting for the potential class imbalance present in the cross-domain dataset. This ensures a more robust and reliable evaluation of the model's ability to discern between truthful and deceptive opinions.

Algorithm	Normalizer	AUC Weighted
Light GBM	Sparse Normalizer	0.67
Random Forest	Sparse Normalizer	0.68
Light GBM	Standard Scaler Wrapper	0.68
Light GBM	Max Abs Scaler	0.69
Random Forest	Max Abs Scaler	0.69
Random Forest	Standard Scaler Wrapper	0.70
Logistic Regression	Standard Scaler Wrapper	0.71
Extreme RandomTrees	Max Abs Scaler	0.73
Light GBM	Standard Scaler Wrapper	0.74
Extreme Random Trees	Max AbsScaler	0.74

TABLE I
CLASSIFIERS UTILIZING SENTIMENT-BASED FEATURES

Table I clearly indicates that the classifier's performance is consistent, signifying that the features are robust even in cross-domain situations. It is worth noting that the combined dataset comprises different domains as well as both positive and negative reviews. This suggests that the proposed features can effectively withstand shifts in sentiment as well.

While there is a reduction in accuracy compared to related work, we can still achieve relatively high and consistent outcomes, which is more crucial as it lowers the possibility of overfitting. This brings us closer to establishing a general rule for deception detection rather than merely tailoring a classifier to a specific dataset, which would be less effective in identifying deception on the internet.

V. CONCLUSION AND FUTURE WORK

Our findings have broader implications for future cross-domain approaches, leading to specific recommendations. Firstly, when transitioning from within-domain to cross-domain detection, a noticeable decline in classification performance should be anticipated, regardless of the approach used. The specific details investigated in this study are unable to completely counteract this performance drop. Therefore, if feasible, using training data closely related to the testing data in terms of domain is recommended, with the closer the better.

However, in situations where this is not possible, particularly when the training content significantly differs from the test content, it is crucial to consider the trade-off between false negatives and false positives. If false negatives pose a greater concern, relying solely on linguistically-defined specific details can be beneficial. Conversely, if false positives are of greater concern, it is preferable to use a combination of n-gram and linguistically-defined specific detail features.

Drawing on insights from prior deception detection methods, encompassing both within-domain and cross-domain approaches, we have identified linguistically defined sentiment and stylometric features that effectively detect deception across domains under specific circumstances. Notably, our features prove most valuable when considerable content differences exist between training and test sets, and when the cost of false negatives outweighs that of false positives. We anticipate that future research will leverage these findings to enhance general-purpose deception detection strategies.

REFERENCES

[1] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

[2] Nikolai Vogler and Lisa Pearl. Using linguistically defined specific details to detect deception across domains. *Nat. Lang. Eng.*, 26(3):349–373, 2020.

[3] Dan Barsever, Sameer Singh, and Emre Neftci. Building a better lie detector with bert: The difference between truth and lies. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[4] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230, 2008.

[5] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992.

[6] Alex Marin, Roman Holenstein, Ruhi Sarikaya, and Mari Ostendorf. Learning phrase patterns for text classification using a knowledge graph and unlabeled data. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[7] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[8] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354, 2005.

[9] Alex Marin, Mari Ostendorf, Bin Zhang, Jonathan T Morgan, Meghan Oxley, Mark Zachry, and Emily M Bender. Detecting authority bids in online discussions.

In *2010 IEEE Spoken Language Technology Workshop*, pages 49–54. IEEE, 2010.

[10] Caroline Brun and Caroline Hagege. Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science*, 70(79.7179):5379–62, 2013.

[11] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[12] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 309–312, 2009.

[13] Tommaso Fornaciari and Massimo Poesio. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21:303–340, 2013.

[14] Eugene Santos and Deqing Li. On deception detection in multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(2):224–235, 2009.

[15] Victoria L Rubin and Tatiana Vashchilko. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 97–106, 2012.

[16] Judee K Burgoon, J Pete Blair, Tiantian Qin, and Jay F Nunamaker. Detecting deception through linguistic analysis. In *Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2–3, 2003 Proceedings 1*, pages 91–101. Springer, 2003.

[17] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 59–66, 2015.

[18] Ángela Almela, Gema Alcaraz-Mármol, and Pascual Cantos. Analysing deception in a psychopath's speech: a quantitative approach. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 31:559–572, 2015.

[19] Tommaso Fornaciari and Massimo Poesio. Identifying fake amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287. Association for Computational Linguistics, 2014.

[20] Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576, 2014.

[21] Katarina R Krüger, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707, 2017.

[22] Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501, 2013.

[23] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 446–457, 2020.

[24] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[25] Lina Zhou, Judee K Burgoon, Douglas P Twitchell, Tiantian Qin, and Jay F Nunamaker Jr. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–166, 2004.

[26] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500, 2007.

[27] Whitney L Cade, Blair A Lehman, and Andrew Olney. An exploration of off topic conversation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 669–672, 2010.

[28] Yafeng Ren and Donghong Ji. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385:213–224, 2017.

[29] Judee K Burgoon, David B Buller, Laura K Guerrero, Walid A Afifi, and Clyde M Feldman. Interpersonal deception: Xii. information management dimensions underlying deceptive and truthful messages. *Communications Monographs*, 63(1):50–69, 1996.

[30] Steven A McCornack. Information manipulation theory. *Communications Monographs*, 59(1):1–16, 1992.

[31] Bennett Kleinberg, Maximilian Mozes, Arnoud Arntz, and Bruno Verschuere. Using named entities for computer-automated verbal deception detection. *Journal of forensic sciences*, 63(3):714–723, 2018.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.