The Language of Deception: Applying Findings on Opinion Spam to Legal and Forensic Discourses

Alibek Jakupov, Julien Longhi, Besma Zeddini

languages

MDPI

Article

The Language of Deception: Applying Findings on Opinion Spam to Legal and Forensic Discourses

Alibek Jakupov [1], Julien Longhi [2,*] and Besma Zeddini [1]

[1] SATIE Laboratory CNRS–UMR 8029, CY Tech, CY Cergy Paris University, 95000 Cergy, France; jakupovalibekdev@gmail.com (A.J.); besma.zeddini@cyu.fr (B.Z.)
[2] AGORA Laboratory EA 7392, CY Cergy Paris University, 95000 Cergy, France
* Correspondence: julien.longhi@cyu.fr

Abstract: Digital forensic investigations are becoming increasingly crucial in criminal investigations and civil litigations, especially in cases of corporate espionage and intellectual property theft as more communication occurs online via e-mail and social media. Deceptive opinion spam analysis is an emerging field of research that aims to detect and identify fraudulent reviews, comments, and other forms of deceptive online content. In this paper, we explore how the findings from this field may be relevant to forensic investigation, particularly the features that capture stylistic patterns and sentiments, which are psychologically relevant aspects of truthful and deceptive language. To assess these features' utility, we demonstrate the potential of our proposed approach using the real-world dataset from the Enron Email Corpus. Our findings suggest that deceptive opinion spam analysis may be a valuable tool for forensic investigators and legal professionals looking to identify and analyze deceptive behavior in online communication. By incorporating these techniques into their investigative and legal strategies, professionals can improve the accuracy and reliability of their findings, leading to more effective and just outcomes.

Keywords: digital investigation; NLP-based forensics; deceptive opinion spam; feature engineering; stylometry; sentiment analysis

Citation: Jakupov, Alibek, Julien Longhi and Besma Zeddini. 2024. The Language of Deception: Applying Findings on Opinion Spam to Legal and Forensic Discourses. *Languages* 9: 10. https://doi.org/10.3390/languages9010010

Academic Editor: Alan Garnham

Received: 16 October 2023
Revised: 10 December 2023
Accepted: 15 December 2023
Published: 22 December 2023

1. Introduction

Digital communication mediums like emails and social networks are crucial tools for sharing information and communication, but they can also be misused for criminal and political purposes. A notable instance of this misuse was the spread of false information during the U.S. election. Lazer et al. highlighted that "misinformation has become viral on social media" (Lazer et al. 2018). They underscored the importance for researchers and other relevant parties to encourage cross-disciplinary studies aimed at curbing the propagation of misinformation and addressing the root issues it exposes. Reports and worries have also arisen about terrorists and other criminal groups taking advantage of social media to promote their unlawful endeavors, such as setting up discrete communication pathways to share information (Goodman 2018). Therefore, it is not unexpected that government bodies are closely scrutinizing these platforms or communication paths. Most existing studies focus on creating a map of individual relationships within a communication network. The primary goal in these methods is to pinpoint the closest associates of a known target. These methods aim to enhance precision, recall, and/or the F1 score, often overlooking the significance of the content within conversations or messages. As a result, these methods can be highly specific (tailored for particular outcomes), may lack accuracy, and may not be ideal for digital investigations (Keatinge and Keen 2020). For example, in the tragic incident at the Gilroy Garlic Festival, the shooter had reportedly expressed his anger on his Facebook page before the incident. This post, however, did not attract the attention of pertinent parties until after the tragedy. This lack of attention is not surprising, given that

the shooter was not a recognized threat on the social network, and his post might not have been given high priority using traditional methods (Sun et al. 2021).

The example mentioned above demonstrates how written information can be employed to influence public opinion and impact the outcome of important events. There is a field within Natural Language Processing (NLP) that concentrates on scrutinizing on a similar phenomenon, called Deceptive Opinion Spam. Therefore, certain findings within this field could significantly enhance our comprehension of forensic linguistic analysis. Opinion Spam refers to reviews that are inappropriate or fraudulent, which can take on various forms such as self-promotion of an unrelated website or blog, or deliberate review fraud that could lead to monetary gain (Ott et al. 2011). Organizations have a strong incentive to detect and eliminate Opinion Spam via automation. This is because the primary concern with Opinion Spam is its influence on customer perception, particularly with regards to reviews that inaccurately praise substandard products or criticize superior ones (Vogler and Pearl 2020). Compared to other NLP tasks like sentiment analysis or intent detection, there has been relatively little research on using text classification approaches to detect Opinion Spam (Barsever et al. 2020). One can easily identify certain types of opinion spam, such as promotional content, inquiries, or other forms of non-opinionated text (Jindal and Liu 2008). The described situations can be classified as Disruptive Opinion Spam, characterized by irrelevant comments that are easily recognizable by the audience and pose a minimal threat, as individuals are empowered to disregard them if they so choose (Ott et al. 2011). When it comes to Deceptive Opinion Spam, which involves more nuanced forms of fake content, the task of identifying it is not as simple; the reason being that these statements are intentionally constructed to seem authentic and mislead the assessor (Ott et al. 2011). Deceptive Opinion Spam is a type of fraudulent behavior where a malicious user creates fictitious reviews, either positive or negative, with the intention of either boosting or damaging the reputation of a business or enterprise (Barsever et al. 2020). Thus, the deliberate intention to deceive readers in certain statements makes it challenging for human reviewers to accurately identify such deceptive texts, resulting in a success rate that is not significantly better than chance (Vogler and Pearl 2020). Consequently, discoveries in Deceptive Opinion Spam could prove valuable for designing digital investigation techniques for studying different communication channels, such as social networks. In contrast to traditional methods, the strategy that incorporates NLP techniques, particularly those used for Deceptive Opinion Spam analysis, places emphasis on both the interaction among individuals and the substance of the communication which may significantly improve the investigation process (Sun et al. 2021).

The problem is commonly addressed as a task of classifying text. Text classification systems typically consist of two key elements: a module for vectorization and a classifier. The vectorization module is tasked with creating features from a provided text sequence, while the classifier assigns category labels to the sequence using a set of matching features. These features are usually categorized into lexical and syntactic groups. Lexical features may include metrics such as total words or characters per word, as well as the frequency of long and unique words. On the other hand, syntactic features primarily consist of the frequency of function words or word groups, such as bag-of-words (BOW), n-grams, or Parts-Of-Speech (POS) tagging (Brown et al. 1992). In addition to vocabulary and sentence structure aspects, there are also methods known as lexicon containment techniques. These techniques symbolize the presence of a term from the lexicon in a text as a binary value, with positive indicating its existence and negative denoting its absence (Marin et al. 2014). The lexicons for such kind of features are constructed by a human expert (Pennebaker et al. 2001; Wilson et al. 2005) or generated automatically (Marin et al. 2010). Several approaches suggest integrating the text's morphological relationships and reliant linguistic components as input vectors for the classification algorithm (Brun and Hagege 2013). In addition to this, there are semantic vector space models which serve to characterize each word via a real-valued vector, determined using the distance or angle between pairs of word vectors (Sebastiani 2002). In the field of automatic fraudulent text

detection, various approaches have been applied, mostly relying on linguistic features, such as n-grams (Fornaciari and Poesio 2013; Mihalcea and Strapparava 2009; Ott et al. 2011), discourse structure (Rubin and Vashchilko 2012; Santos and Li 2009), semantically related keyword lists (Burgoon et al. 2003; Pérez-Rosas et al. 2015), measures of syntactic complexity (Pérez-Rosas et al. 2015), stylometric features (Burgoon et al. 2003), psychologically motivated keyword lists (Almela et al. 2015), and parts of speech (Fornaciari and Poesio 2014; Li et al. 2014).

These vectorization strategies are typically utilized to examine the significance of the features, which helps to highlight recurring patterns in the framework of fraudulent statements that are less prevalent in truthful texts. Although this technique shows some effectiveness, it has significant drawbacks due to the difficulty in controlling the quality of the training set. For example, while many of the classification algorithms, trained using this method, show acceptable performance within their specific fields, they struggle to generalize effectively across different domains, thereby lacking resilience in adapting to domain changes. (Krüger et al. 2017). As an illustration, a mere alteration in the polarity of fraudulent 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 (Ott et al. 2013). This observation holds when the training and the testing dataset originate from different domains (Mihalcea and Strapparava 2009). Additionally, specific categorization models that rely on semantic vector space models could be significantly influenced by social or personal biases embedded in the training data. This can lead the algorithm to make incorrect deductions. (Papakyriakopoulos et al. 2020). Furthermore, certain studies suggest that deceptive statements differ from truthful ones more in terms of their sentiment then other linguistic features (Newman et al. 2003). According to certain cases, the deceivers display a more positive affect in order to mislead the audience (Zhou et al. 2004), whereas certain instances demonstrate that deception is characterized by more words reflecting negative emotion (Newman et al. 2003).

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 forensic investigations. 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, and syntactic rules) (Vogler and Pearl 2020). 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.

2. Our Approach

Researchers in the forensic domain typically address investigative questions via linguistic analysis, such as identifying authors of illegal activities, understanding the content of documents, and extracting information about the timing, location, and intent of the text (Longhi 2021). Alternatively, studies into Deceptive Opinion Spam, which focus on fraudulent analysis, have proposed techniques for examining linguistic semantics by identifying patterns in the expression and content from a statistical standpoint. In fact, this method aligns with a forensic science approach, combining quantitative identification and qualitative analysis based on the analysis corpus consisting of different texts related to criminal acts, particularly involving terrorist groups, mostly in the same manner as scholars studying misleading discourse, but with the Ott Deceptive Opinion Spam corpus and the Multi-Domain Deceptive corpus instead (Jakupov et al. 2022). The goal is to assist investigators in finding stylistic similarities or exclusions between texts and potentially their authors.

In this paper, we explore the effectiveness of a novel linguistically defined implementation of stylometric and sentiment-based features for digital investigation. We begin by examining prior approaches to automatic fraudulent text detection, emphasizing tech-

niques that employ linguistic features such as n-grams, which provide the best performance within the domain. Following that, we outline the diverse corpora used to evaluate our approach and its cross-domain performance. Next, we explore the suggested sentiment-based features, confirming their possible significance in forensic examination within these collections. We also investigate the stylometric features and diagnostic potential of non-functional words, but without incorporating them into the classifier. Finally, we describe our classification scheme, which leverages these features.

2.1. Our Contributions

Our contributions can be summarized as follows.

- Novel approach to automatic digital forensic investigation that applies sentiment-based features;
- Comprehensive analysis of previous approaches to digital investigation, 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 forensic analysis;
- Investigation of the diagnostic potential of non-functional words as stylometric features

The significance of our contributions towards the advancement of automated digital forensic investigation lies in the incorporation of sentiment-based features, thereby transforming the paradigm of digital investigation methodologies. It particularly emphasizes the importance and diagnostic potential of non-functional words as stylometric features, which are typically overlooked by researchers.

Outline

The rest of the paper is organized as follows: in Section 3, we provide an overview of related work; in Section 4, we summarize our methodology for topic modeling and we present and discuss the experimental results as well as the datasets used to benchmark our approaches; finally, conclusions and discussions are provided in Sections 5 and 6.

3. Related Work

The idea of employing machine learning and deep learning methods to identify dubious activities in social networks has garnered general attention. For instance, Bindu et al. introduced an unsupervised learning method that can automatically spot unusual users in a static social network, albeit assuming that the network's structure does not change dynamically Bindu et al. (2017). Hassanpour et al. applied deep convolutional neural networks for images and long short-term memory (LSTM) to pull out predictive characteristics from Instagram's textual data, showing the capability to pinpoint potential substance use risk behaviors, aiding in risk evaluation and strategy formulation (Hassanpour et al. 2019). Tsikerdekis used machine learning to spot fraudulent accounts trying to enter an online sub-community for prevention purposes (Tsikerdekis 2016). Ruan et al. also used machine learning to detect hijacked accounts based on their online social behaviors (Ruan et al. 2015). Fazil and Abulaish suggested a mixed method to detect automated spammers on Twitter, using machine learning to examine related aspects like community-based features (e.g., metadata, content, and interaction-based features) (Fazil and Abulaish 2018). Cresci et al. employed machine learning to spot spammers using digital DNA technology, with the social fingerprinting technique designed to distinguish between spam bots and genuine accounts in both supervised and unsupervised manners (Cresci et al. 2017). Other applications focused on urban crime perception utilizing the convolutional neural network as their learning preference (Fu et al. 2018; Shams et al. 2018).

Certain studies showed the potential of focusing purely on textual data, especially in the context of social network analysis (Ala'M et al. 2017). One example of this application was in 2013, when Keretna et al. used a text mining tool, Stanford POS tagger, to pull out

features from Twitter posts that could indicate a user's specific writing style (Keretna et al. 2013). These features were then used in the creation of a learning module. Similarly, Lau et al. used both NLP and machine learning techniques to analyze Twitter data. They found that the Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) methods yielded the best results in terms of the Area Under the ROC Curve (AUC) (Lau et al. 2014). In addition, Egele et al. developed a system to identify compromised social network accounts by analyzing message content and other associated features (Egele et al. 2015). Anwar and Abulaish introduced a unified social graph text mining framework for identifying digital evidence from chat logs based on user interaction and conversation data (Anwar and Abulaish 2014). Wang et al. treated each HTTP flow produced by mobile applications as text and used NLP to extract text-level features. These features were then used to create an effective malware detection model for Android viruses (Wang et al. 2017). Al-Zaidya et al. designed a method to efficiently find relevant information within large amounts of unstructured text data, visualizing criminal networks from documents found on a suspect's computer (Al-Zaidy et al. 2012). Lastly, Louis and Engelbrecht applied unsupervised information extraction techniques to analyze text data and uncover evidence, a method that could potentially find evidence overlooked by a simple keyword search (Louis and Engelbrecht 2011).

Li et al. applied their findings to detect fraudulent hotel reviews, using the Ott Deceptive Opinion spam corpus, and obtained a score of 81.8% by capturing the overall dissimilarities between truthful and deceptive texts (Li et al. 2014). The researchers expanded upon the Sparse Additive Generative Model (SAGE), which is a Bayesian generative model that combines both topic models and generalized additive models, and this resulted in the creation of multifaceted latent variable models via the summation of component vectors. Since most studies in this area focus on recognizing deceitful patterns instead of teaching a solitary dependable classifier, the primary difficulty of the research was to establish which characteristics have the most significant impact on each classification of a misleading review. Additionally, it was crucial to assess how these characteristics affect the ultimate judgment when they are paired with other attributes. SAGE is a suitable solution for meeting these requirements because it has an additive nature, which allows it to handle domain-specific attributes in cross-domain scenarios more effectively than other classifiers that may struggle with this task. The authors discovered that the BOW method was not as strong as LIWC and POS, which were modeled using SAGE. As a result, they formulated a general principle for identifying deceptive opinion spam using these domain-independent features. Moreover, unlike the creator of the corpus (Ott et al. 2011), they identified the lack of spatial information in hotel reviews as a potential indicator for identifying fraudulent patterns, of which the author's findings suggest that this methodology may not be universally appropriate since certain deceptive reviews could be authored by experts in the field. Although the research found that the domain-independent features were effective in identifying fake reviews with above-chance accuracy, it has also been shown that the sparsity of these features makes it difficult to utilize non-local discourse structures (Ren and Ji 2017); thus, the trained model may not be able to grasp the complete semantic meaning of a document. Furthermore, based on their findings, we can identify another significant indication of deceptive claims: the existence of sentiments. This is because reviewers often amplify their emotions by utilizing more vocabulary related to sentiments in their statements.

(Ren and Ji 2017) built upon earlier work by introducing a three-stage system. In the first stage, they utilized a convolutional neural network to generate sentence representations from word representations. This was performed by employing convolutional action, which is commonly used to synthesize lexical n-gram information. To accomplish this step, they employed three convolutional filters. These filters are effective at capturing the contextual meaning of n-grams, including unigrams, bigrams, and trigrams. This approach has previously proven successful for tasks such as sentiment classification. (Wilson et al. 2005). Subsequently, they created a model of the semantic and discourse relations of these sentence

vectors to build a document representation using a two-way gated recurrent neural network. These document vectors are ultimately utilized as characteristics to train a classification system. The authors achieved an 85.7% accuracy on the dataset created by Li et al. and showed that neural networks can be utilized to obtain ongoing document representations for the improved understanding of semantic features. The primary objective of this research was to practically show the superior efficacy of neural features compared to conventional discrete feature (like n-grams, POS, LIWC, etc.) due to their stronger generalization. Nevertheless, the authors' further tests showed that by combining discrete and neural characteristics, the total precision can be enhanced. Therefore, discrete features, such as the combination of sentiments or the use of non-functional words, continue to be a valuable reservoir of statistical and semantic data.

(Vogler and Pearl 2020) conducted a study investigating the use of particular details in identifying disinformation, both within a single area and across various areas. Their research focused on several linguistic aspects, including n-grams, POS, syntactic complexity metrics, syntactic configurations, lists of semantically connected keywords, stylometric properties, keyword lists inspired by psychology, discourse configurations, and named entities. However, they found these features to be insufficiently robust and adaptable, especially in cases where the area may substantially differ. This is mainly because most of these aspects heavily rely on specific lexical elements like n-grams or distinct keyword lists. Despite the presence of complex linguistic aspects such as stylometric features, POS, or syntactic rules, the researchers consider these to be of lesser importance because they do not stem from the psychological basis of verbal deceit. In their research, they saw deceit as a product of the imagination. Consequently, in addition to examining linguistic methods, they also explored approaches influenced by psychological elements, like information management theory (Burgoon et al. 1996), information manipulation theory (McCornack 1992), and reality monitoring and criteria-based statement analysis (Vogler and Pearl 2020). Since more abstract linguistic cues motivated by psychology may have wider applicability across various domains (Kleinberg et al. 2018), the authors find it beneficial to use these indicators grounded in psychological theories of human deception. They also lean on the research conducted by Krüger et al. which focuses on identifying subjectivity in news articles and proposes that linguistically abstract characteristics could potentially be more robust when used on texts from different fields (Krüger et al. 2017). For their experiment, Vogler and Pearl employed three different datasets for the purpose of training and evaluation, accommodating shifts in the domain, ranging from relatively subtle to considerably extensive: the Ott Deceptive Opinion Spam Corpus (Ott et al. 2011), essays on emotionally charged topics (Mihalcea and Strapparava 2009), and personal interview questions (Burgoon et al. 1996). The linguistically defined specific detail features the authors constructed for this research proved to be successful, particularly when there were notable differences in the domains used for training and testing. These elements were rooted in proper nouns, adjective phrases, modifiers in prepositional phrases, exact numeral terms, and noun modifiers appearing as successive sequences. The characteristics were derived from appropriate names, descriptive phrase clusters, prepositional phrase changes, precise numerical terms, and noun modifiers that showed up as successive sequences. Each attribute is depicted as the total normalized number and the average normalized weight. The highest F score they managed to obtain was 0.91 for instances where content remained consistent, and an F score of 0.64 for instances where there was a significant domain transition. This suggests that the linguistically determined specific detail attributes display a broader range of application. Even though the classifier trained with these features showed fewer false negatives, it struggled to accurately categorize truthful texts. The experimental results clearly indicate that a combination of n-gram and language-specific detail features tends to be more dependable only when a false positive carries a higher cost than a false negative. It is worth noting that features based on n-grams might have a superior ability for semantic expansion when they are built on distributed meaning representations like GloVe and ELMo. In their technique, however, n-gram

features rely only on single words without considering the semantic connection among them. This stands in stark contrast to our method, which revolves around analyzing the semantic essence of statements by evaluating the overall sentiment.

4. Materials and Methods

4.1. Model

Stylometry is a quantitative study of literary style that employs computational distant reading methods to analyze authorship. This approach is rooted in the fact that each writer possesses a distinctive, identifiable, and fairly stable writing style. This unique writing style is apparent in different writing components, including choice of words, sentence construction, punctuation, and the use of minor function words like conjunctions, prepositions, and articles. The fact that these function words are used unconsciously and independent of the topic makes them especially valuable for stylometric study.

In our research, we investigate the use of stylometric analysis in identifying misinformation, concentrating on the distinctive language patterns that can distinguish between honest and dishonest writings. Through the scrutiny of multiple stylometric aspects, our goal was to reveal the hidden features of dishonest language and establish a trustworthy approach for forensic investigation.

To obtain a better understanding of how lies are expressed in text, we utilized the Burrows' Delta method, a technique that gauges the "distance" between a text whose authorship is uncertain and another body of work. This approach is different from others like Kilgariff's chi-squared, as it is specifically structured to compare an unidentified text (or group of texts) with the signatures of numerous authors concurrently. More specifically, the Delta technique assesses how the unidentified text and groups of texts authored by an arbitrary number of known authors deviate from their collective average. Notably, the Delta method assigns equal importance to every characteristic it measures, thereby circumventing the issue of prevalent words dominating the outcomes, an issue often found in chi-squared tests. For these reasons, the Delta Method developed by John Burrows is typically a more efficient solution for authorship identification. We modified this method to discern the usage of non-functional words by deceivers and ordinary internet users. As this method extracts features that are not topic-dependent, we are able to establish a model that is resilient to changes in the domain.

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

- Compile a comprehensive collection of written materials from a variable number of categories, which we will refer to as x (such as deceptive and truthful).
- Identify the top n words that appear most often in the dataset to utilize as attributes.
- 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 the deceptive's subcorpus.
- Next, compute the average and standard deviation of these x values and adopt them as the definitive average and standard deviation for this characteristic across the entire body of work. Essentially, we will employ an average of the averages, rather than determining a sole value that symbolizes the proportion of the whole body of work represented by each term. We do this because we want to prevent a larger subsection of the body of work from disproportionately affecting the results and establish the standard for the body of work in a way that everything is presumed to resemble 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.

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

- Next, calculate identical z scores for each characteristic in the text, where the authorship needs to be ascertained.
- Finally, compute a delta score to compare the unidentified text with each candidate's subset of text. This can be performed by calculating the mean of the absolute differences between the z scores for each characteristic in both the unidentified text and the candidate's text subset. 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. The formula below 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 class, or "winning" candidate, is most likely determined by finding the one with the least amount of difference in the score between their respective subcorpus and the test case. This indicates the least variation in writing style, which makes it the most probable class (either deceptive or truthful) for the text being examined.

In our methodology, we also incorporated a measure of exaggeration, consistently applied across various domains. The fundamental idea suggests that the intensity of the sentiment remains unchanged, irrespective of the text expressing a positive or negative sentiment (for instance, "I love the product" and "I detest the product" indicate the same level of sentiment, although in contrary directions). In order to examine false opinion spam, we made use of Azure Text Analytics API[1], which facilitates the analysis of the overall sentiment and the extraction of three aspects: positive, negative, and neutral. This was innately similar to the RGB color model, leading us to assign the values in the same way: Negative was paired with Red, Positive with Green, and Neutral with Blue. Following this, we displayed the pattern that began to form.

To illustrate the emotional trends in both honest and dishonest reviews, we initially utilized color-coding derived from sentiment analysis findings. To begin, we converted the sentiment ratings (positive, negative, and neutral) into a blue–green–red (BGR) format, which allowed us to represent each review as a pixel. Considering that Azure Text Analytics offers percentages for every sentiment component (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.

After recognizing visual patterns, we used these figures as attributes for our categorizer. To prevent the categorizer from making incorrect inferences by evaluating sentiments instead of hyperbole, we initially determined the total sentiment. If the sentiment was adverse, we exchanged the green and red channels, as hyperbole is steady for both negative and positive sentiments. We then standardized this set of attributes, as the percentage of neutral aspect is generally much higher than the other sentiments in most situations. Finally, we input these features into our classifier and examined the subsequent results as shown in Algorithm 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

4.2. Data

Our initial approach involved examining labeled fraudulent reviews in order to train the model. One of the first large-scale, publicly available datasets for the research in this domain is Ott Deceptive Opinion Spam corpus (Ott et al. 2011), composed of 400 truthful and 400 gold-standard deceptive reviews. In order to obtain deceptive reviews of high quality via Amazon Mechanical Turk, a set of 400 Human-Intelligence Tasks (HITs) were created and distributed among 20 selected hotels. To ensure uniqueness, only one submission per Turker was allowed. To obtain truthful reviews, the authors gathered 6977 reviews from the 20 most popular Chicago hotels on Trip Advisor. Despite the dataset, the authors have discovered that detecting deception is a challenge for human judges, as most of them performed poorly.

To prevent our model from identifying inaccurate features that are related to the domain rather than deceptive cues, we augmented our training dataset with cross-domain data. For cross-domain investigation, we applied a dataset consisting of hotel, restaurant, and doctor reviews (Li et al. 2014) 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.

Following the model's training, we opted to assess its usefulness in forensic investigations by evaluating it on real-world email data. Email serves as a crucial means of communication within most businesses, facilitating internal dialogue between staff members and external communication with the broader world. Consequently, it offers a wealth of data that could potentially highlight issues. However, this brings up the issue of privacy, as the majority of employees would not be comfortable knowing their employer has access to their emails. Therefore, it is critical to adopt methods to manage this issue that are as non-invasive as possible. This is also beneficial to the organization, as implementing a system that literally "reads" employees' emails could prove to be excessively costly.

Theories of deceptive behavior, fraud, or conspiracy suggest that changes in language use can signal elements such as feelings of guilt or self-awareness regarding the deceit, as well as a reduction in complexity to ease the consistency of repetition and lessen the mental load of fabricating a false narrative (Keila and Skillicorn 2005). The potential presence

of some form of monitoring may also lead to an excessive simplicity in messages, as the senders strive to avoid detection. This simplicity could, in itself, become a telltale sign. It is also probable that messages exchanged between collaborators will contain abnormal content, given that they are discussing actions that are unusual within their context.

The Enron email dataset was made publicly available in 2002 by the Federal Energy Regulatory Commission (FERC). This dataset consists of real-world emails that were sent and received by ex-Enron employees. The dataset contains 517,431 emails from the mail folders of 150 ex-Enron employees, including top executives such as Kenneth Lay and Jeffrey Skilling. While most of the communication in the dataset is mundane, some emails from executives who are currently being prosecuted suggest the presence of deceptive practices. The emails contain information such as sender and receiver email addresses, date, time, subject, body, and text, but do not include attachments. This dataset is widely used for research purposes and was compiled by Cohen at Carnegie Mellon University. We initiated a preprocessing phase to polish the dataset, which involved eliminating redundant entries, junk emails, unsuccessful and blank emails, along with punctuation symbols (essential for applying sentiment analysis). This purification process resulted in a remaining total of 47,468 emails, all of which were either dispatched or obtained by 166 previous Enron employees. Among these employees, 25 were marked as "criminals", a term denoting those who were supposedly involved in fraudulent acts.

5. Results

At first, we analyzed a group of deceptive reviews which consisted of the Ott Deceptive Opinion Spam Corpus and the cross-domain corpus of reviews for hotels, restaurants, and doctors curated by Li et al. Our aim was to confirm that the use of non-essential words remained 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 of the negative deceptive test indicated 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. As for the positive tests, the deceptive test had a delta score of 1.4003 for deceptive and 1.8459 for truthful, whereas the truthful test had a delta score of 2.9074 for deceptive and 2.2098 for truthful. Overall, the model accurately detected 65% of deceptive texts and 68% of truthful texts, taking into account both positive and negative cases.

The study primarily investigated the stylometric characteristics and potential usefulness of non-functional words, but decided not to include them in the classifier due to the inherent methodological limitation that necessitates analyzing the entire corpus for vectorizing individual statements. However, the results uncovered interesting patterns that require further exploration and may be potentially applied to forensic investigation.

After exploring the fraudulent reviews, we focused on extracting sentiment-based features. To observe emotional trends in truthful and deceptive reviews, we colored the reviews using a blue–green–red (BGR) format based on their sentiment scores (positive, negative, and neutral). This allowed us to depict each review as a pixel, with blue indicating neutral sentiment, green representing positive sentiment, and red signifying negative sentiment. To convert the sentiment scores into pixel format and create an image from the BGR values, we developed support functions. Each image showcased 400 pixels (20×20), symbolizing 400 reviews.

We created images for different categories of reviews, such as deceptive positive, deceptive negative, truthful positive, and truthful negative, and compared their visual patterns. The analysis showed that fake negative reviews had a brighter appearance with less green spots, whereas fake positive reviews had more vibrant colors with fewer red spots. This suggests that there is an element of exaggeration and insincere praise in deceitful reviews. Conversely, truthful reviews appeared to be more authentic and impartial in their emotional tone.

In order to achieve a consistent color that conveys deception, we took all the pixels in the images and computed their average values across three color channels: blue, green, and red. Afterward, we combined the channels to create a single color that symbolizes the mean sentiment of the dishonest reviews, as shown in Figure 1.

According to the study, negative reviews that were truthful appeared to be less red in color than negative reviews that were deceptive. On the other hand, positive reviews that were fake appeared to be greener than positive reviews that were truthful. This indicates that deceptive reviews tend to contain more exaggerated expressions of sentiment, which can be represented through the use of color.

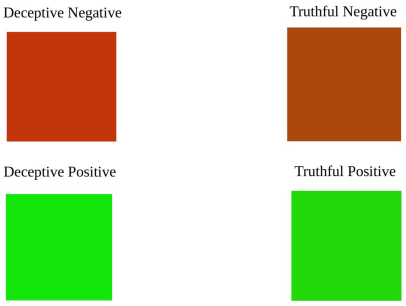

Figure 1. Deceptive datasets: colorized sentiments.

With this in mind, we trained multiple classifiers with features extracted using Algorithm 1. The training was conducted with the Ott Deceptive Opinion Spam dataset, while the Li et al. cross-domain dataset was used for testing. Once we identified the optimal model, we applied it to the Enron email corpus.

In order to ensure that the input features used in a machine learning model have a consistent scale or distribution, we applied different normalization techniques such as MaxAbsScaler, StandardScaler Wrapper, and Sparse Normalizer in our experiment. We chose AUC Weighted as the primary metric to assess the performance of our models. AUC Weighted was selected because it is capable of measuring the classifier's performance across varying thresholds, while also considering the potential class imbalance present in the cross-domain dataset. This guarantees a more reliable and strong evaluation of the model's ability to differentiate truthful and deceptive opinions.

Table 1 clearly indicates that the classifier's performance is consistent, signifying that the features are robust even in cross-domain situations. It should be emphasized that the merged dataset encompasses various fields and includes both favorable and unfavorable evaluations. This implies that the suggested characteristics can proficiently endure changes in the sentiment as well.

Table 1. Classifiers utilizing sentiment-based features

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

While there is a reduction in accuracy compared to related work, we can still achieve relatively high and stable results, which is more important since it reduces the risk of

overfitting. Our progress in this area is leading us towards developing a universal method for detecting deception, rather than creating a classifier that is only suitable for a particular dataset. This approach proves to be more effective in identifying instances of deception on the internet.

The model trained on the deceptive training set was finally applied to the Enron email dataset, including mails from high-ranking executives like Kenneth Lay (ex-Chairman and CEO) and Jeffrey Skilling (ex-CEO). Although the majority of the communication is innocuous and uneventful, the emails of several executives who are currently facing prosecution are included in the dataset, suggesting that evidence of deception could potentially be found within the data. We cross-referenced the name list on the website to confirm the authenticity of the email and determine whether it is misleading. Our model was able to obtain the F1 score of 0.43, but due to the dataset being imbalanced, with only 25 out of 166 employees being identified as criminals, our evaluation of the model takes into account some level of uncertainty.

In order to comprehend how our model can be applied in practical scenarios, we assessed its performance against other top-performing models such as SIIMCO (Taha and Yoo 2016) and LogAnalysis (Ferrara et al. 2014), despite them not being rooted in NLP. These methods were devised by building an extensive graph detailing the suspected individuals' connections, with those particularly active in the communication network frequently being strongly implicated as criminals. For example, "employee 57", who exchanged 3247 and 847 emails, respectively, was identified as a criminal as per both existing techniques, or in other words, a true negative.

Upon examining Table 2, it is clear that our approach yields a lower F1 score and precision rate. This disparity can be attributed to several factors.

Firstly, our classifier was trained exclusively on online reviews, excluding emails or any other communication types involving two or more parties. This specificity could affect the textual patterns we can detect. As a result, it would be beneficial to enrich our training set with anonymized conversation data.

Secondly, our preprocessing stage overlooked the removal of email signatures and conversation history. This oversight could distort the analysis results, as the response may not be deceptive itself, but it could contain traces of a previous deceptive email. Consequently, we must refine our text preprocessing pipeline and integrate a layout analysis to distinguish the message body from the metadata, such as signatures or conversation history.

Lastly, the level of exaggeration, which is commonplace in online reviews, may not translate accurately to the corporate communication realm. Therefore, we should consider introducing a variable exaggeration level that adapts to the specific domain.

Table 2. Performance of SIIMCO and LogAnalysis: A comparative summary.

Approach	F1 Score	Precision	Recall
LogAnalysis	0.51	0.49	0.53
SIIMCO	0.59	0.58	0.60
Our proposed approach	0.43	0.26	1

6. Discussion

Current state-of-the-art models, based on common features like n-grams or embeddings, have demonstrated their effectiveness within specific domains, with improvements achieved when combined with other features. However, cross-domain performance tends to decrease as content differences between training and testing datasets increase. The utilization of more abstract linguistic features, such as syntax-based features and psychologically motivated categories, has shown to enhance cross-domain deception detection performance.

Our method has been shown to be effective in detecting deception in various deceptive reviews. Stylometric analysis, which focuses on unique linguistic patterns in writing, has

demonstrated promise in uncovering the underlying characteristics of deceptive language. Sentiment analysis and visualization techniques have also been explored to identify patterns in deceptive and truthful reviews. Converting sentiment scores into color formats and generating images to represent reviews allows for visual comparison and insights into exaggeration levels present in online communication.

However, for better performance on email data, like the Enron dataset, one alternative approach we could have used is a transductive method, specifically by employing topic modeling, such as the LDA model, on the entire dataset. Moreover, we would recommend evaluating the model using a 5×2 Nested Cross Validation method. This involves splitting the preprocessed dataset into five folds, with each fold potentially being chosen as the test set, while the remaining four are used for a 2-fold validation. The training set should then be used to train the classifier, with each generator building a group of classifiers for each possible number of topics from zero up to the number given by the LDA, with the smallest perplexity. The validation set should be used to test these classifiers in terms of precision, recall, and F1 score. Only the best classifiers for each metric should be recommended to the investigator and evaluated in the test set.

To sum up, the insights gained from studying the linguistic and psychological aspects of deception can be leveraged to improve existing tools used by investigators and legal professionals tasked with identifying deceptive behavior in online communication. By providing these individuals with a deeper understanding of the subtle markers that indicate deception, they may be better equipped to assess the credibility of information and make informed decisions in high-stakes situations.

7. Conclusions

The results of our study have significant implications for cross-domain approaches in the future and we have specific suggestions. Firstly, it should be expected that there will be a decline in classification performance when transitioning from within-domain to cross-domain detection, regardless of the approach used. Our study has investigated specific details in this regard, but they are unable to completely negate this drop in performance. Therefore, if possible, it is recommended to use training data that is closely related to the testing data in terms of domain, with a closer match being preferable.

However, when this is not feasible, and the training content differs significantly from the test content, it is important to weigh the tradeoff between false negatives and false positives. If false negatives are a greater concern, relying solely on linguistically defined specific details can be advantageous. On the other hand, if false positives are the greater concern, it is preferable to use a combination of n-gram and linguistically defined specific detail features.

Our study draws on insights from prior deception detection methods, including both within-domain and cross-domain approaches, to identify linguistically defined sentiment and stylometric features that can effectively be applied for forensic investigation across domains under specific circumstances. These features are particularly useful when there are significant content differences between training and test sets, as well as when the cost of false negatives is greater than that of false positives. We anticipate that future research will use these findings to improve general-purpose forensic investigation strategies.

In essence, the advancements made in the field of Deceptive Opinion Spam detection not only hold the potential to improve trust and transparency in online communications, but also contribute to the broader domains of online threat investigation. As research in this area continues to evolve, it is crucial that the knowledge and methodologies developed are shared and adapted across disciplines, thereby maximizing their impact and benefit to society as a whole.

Author Contributions: Conceptualization, A.J. and J.L.; methodology, A.J. and J.L.; software, A.J.; validation, J.L. and B.Z.; formal analysis, B.Z.; investigation, A.J.; resources, J.L.; data curation, A.J.; writing—original draft preparation, A.J.; writing—review and editing, J.L.; visualization, A.J.;

supervision, J.L.; project administration, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available in the author's GitHub repository: https://github.com/ajakupov/ColorizeComments (accessed on 7 July 2023), as well as on the following websites: https://myleott.com/op-spam.html, https://nlp.stanford.edu/~bdlijiwei/Code.html (accessed on 1 June 2023) and https://www.cs.cmu.edu/~enron/ (accessed on 20 August 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
BOW	Bag of Words
POS	Part of Speech
LSTM	Long Short-Term Memory Networks
DNA	Deoxyribonucleic Acid
LDA	Latent Dirichlet Allocation
SVM	Support Vector Machine
FERC	Federal Energy Regulatory Commission
ROC	Rate of Change
AUC	Area under the ROC Curve
HTTP	Hypertext Transfer Protocol
SAGE	Sparse Additive Generative Model
LIWC	Linguistic Inquiry and Word Count
GloVe	Global Vectors for Word Representation
ELMo	Embeddings from Language Model

Note

[1] https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/sentiment-opinion-mining/overview, accessed on 21 September 2023.

References

Al-Zaidy, Rabeah, Benjamin C. M. Fung, Amr M. Youssef, and Francis Fortin. 2012. Mining criminal networks from unstructured text documents. *Digital Investigation* 8: 147–60. [CrossRef]

Ala'M, Al-Zoubi, Ja'far Alqatawna, and Hossam Paris. 2017. Spam profile detection in social networks based on public features. Paper presented at 2017 8th International Conference on information and Communication Systems (ICICS), Irbid, Jordan, April 4–6. pp. 130–35.

Almela, Ángela, Gema Alcaraz-Mármol, and Pascual Cantos. 2015. Analysing deception in a psychopath's speech: A quantitative approach. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 31: 559–72. [CrossRef]

Anwar, Tarique, and Muhammad Abulaish. 2014. A social graph based text mining framework for chat log investigation. *Digital Investigation* 11: 349–62. [CrossRef]

Barsever, Dan, Sameer Singh, and Emre Neftci. 2020. Building a better lie detector with bert: The difference between truth and lies. Paper presented at 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, July 19–24. pp. 1–7.

Bindu, P. V., P. Santhi Thilagam, and Deepesh Ahuja. 2017. Discovering suspicious behavior in multilayer social networks. *Computers in Human Behavior* 73: 568–82. [CrossRef]

Brown, Peter F., Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18: 467–80.

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

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

Burgoon, Judee K., J. Pete Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. Detecting deception through linguistic analysis. Paper presented at Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2–3; Proceedings 1, Berlin/Heidelberg: Springer, pp. 91–101.

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. Social fingerprinting: Detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing* 15: 561–76. [CrossRef]

Egele, Manuel, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2015. Towards detecting compromised accounts on social networks. *IEEE Transactions on Dependable and Secure Computing* 14: 447–60. [CrossRef]

Fazil, Mohd, and Muhammad Abulaish. 2018. A hybrid approach for detecting automated spammers in twitter. *IEEE Transactions on Information Forensics and Security* 13: 2707–19. [CrossRef]

Ferrara, Emilio, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. 2014. Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications* 41: 5733–50. [CrossRef]

Fornaciari, Tommaso, and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial Intelligence and Law* 21: 303–40. [CrossRef]

Fornaciari, Tommaso, and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. Paper presented at 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 26–30. Toronto: Association for Computational Linguistics: pp. 279–87.

Fu, Kaiqun, Zhiqian Chen, and Chang-Tien Lu. 2018. Streetnet: Preference learning with convolutional neural network on urban crime perception. Paper presented at 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, November 6–9. pp. 269–78.

Goodman, Anka Elisabeth Jayne. 2018. When you give a terrorist a twitter: Holding social media companies liable for their support of terrorism. *Pepperdine Law Review* 46: 147.

Hassanpour, Saeed, Naofumi Tomita, Timothy DeLise, Benjamin Crosier, and Lisa A. Marsch. 2019. Identifying substance use risk based on deep neural networks and instagram social media data. *Neuropsychopharmacology* 44: 487–94. [CrossRef]

Jakupov, Alibek, Julien Mercadal, Besma Zeddini, and Julien Longhi. 2022. Analyzing deceptive opinion spam patterns: The topic modeling approach. Paper presented at 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), Macao, China, October 31–November 2, pp. 1251–61.

Jindal, Nitin, and Bing Liu. 2008. Opinion spam and analysis. Paper presented at 2008 International Conference on Web Search and Data Mining, Palo Alto, CA, USA, February 11–12. pp. 219–30.

Keatinge, Tom, and Florence Keen. 2020. Social media and (counter) terrorist finance: A fund-raising and disruption tool. In *Islamic State's Online Activity and Responses*. London: Routledge, pp. 178–205.

Keila, Parambir S., and David B. Skillicorn. 2005. Detecting unusual and deceptive communication in email. Paper presented at Centers for Advanced Studies Conference, Toronto, ON, Canada, October 17–20. Pittsburgh: Citeseer, pp. 17–20.

Keretna, Sara, Ahmad Hossny, and Doug Creighton. 2013. Recognising user identity in twitter social networks via text mining. Paper presented at 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, October 13–16. pp. 3079–82.

Kleinberg, Bennett, Maximilian Mozes, Arnoud Arntz, and Bruno Verschuere. 2018. Using named entities for computer-automated verbal deception detection. *Journal of Forensic Sciences* 63: 714–23. [CrossRef]

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

Lau, Raymond Y. K., Yunqing Xia, and Yunming Ye. 2014. A probabilistic generative model for mining cybercriminal networks from online social media. *IEEE Computational Intelligence Magazine* 9: 31–43. [CrossRef]

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and et al. 2018. The science of fake news. *Science* 359: 1094–96. [CrossRef]

Li, Jiwei, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. Paper presented at 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, June 22–27. pp. 1566–76.

Longhi, Julien. 2021. Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International* 318: 110564. [CrossRef]

Louis, A. L., and Andries P. Engelbrecht. 2011. Unsupervised discovery of relations for analysis of textual data. *Digital Investigation* 7: 154–71. [CrossRef]

Marin, Alex, Mari Ostendorf, Bin Zhang, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, and Emily M. Bender. 2010. Detecting authority bids in online discussions. Paper presented at 2010 IEEE Spoken Language Technology Workshop, Berkeley, CA, USA, December 12–15. pp. 49–54.

Marin, Alex, Roman Holenstein, Ruhi Sarikaya, and Mari Ostendorf. 2014. Learning phrase patterns for text classification using a knowledge graph and unlabeled data. Paper presented at Fifteenth Annual Conference of the International Speech Communication Association, Singapore, September 14–18.

McCornack, Steven A. 1992. Information manipulation theory. *Communications Monographs* 59: 1–16. [CrossRef]
Mihalcea, Rada, and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. Paper presented at ACL-IJCNLP 2009 Conference Short Papers, Singapore, August 4, pp. 309–12.
Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29: 665–75. [CrossRef]
Ott, Myle, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. Paper presented at 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, June 9–14. pp. 497–501.
Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv* arXiv:1107.4557.
Papakyriakopoulos, Orestis, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. Paper presented at 2020 Conference on Fairness, Accountability, and Transparency,Barcelona, Spain, January 27–30. pp. 446–57.
Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: Liwc 2001*. Mahway: Lawrence Erlbaum Associates, vol. 71.
Pérez-Rosas, Verónica, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. Paper presented at 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 9–13. pp. 59–66.
Ren, Yafeng, and Donghong Ji. 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences* 385: 213–24. [CrossRef]
Ruan, Xin, Zhenyu Wu, Haining Wang, and Sushil Jajodia. 2015. Profiling online social behaviors for compromised account detection. *IEEE Transactions on Information Forensics and Security* 11: 176–87. [CrossRef]
Rubin, Victoria L., and Tatiana Vashchilko. 2012. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. Paper presented at Workshop on Computational Approaches to Deception Detection, Avignon, France, April 23; pp. 97–106.
Santos, Eugene, and Deqing Li. 2009. On deception detection in multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40: 224–35. [CrossRef]
Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 34: 1–47. [CrossRef]
Shams, Shayan, Sayan Goswami, Kisung Lee, Seungwon Yang, and Seung-Jong Park. 2018. Towards distributed cyberinfrastructure for smart cities using big data and deep learning technologies. Paper presented at 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), Vienna, Austria, July 2–6. pp. 1276–83.
Sun, Dongming, Xiaolu Zhang, Kim-Kwang Raymond Choo, Liang Hu, and Feng Wang. 2021. Nlp-based digital forensic investigation platform for online communications. *Computers & Security* 104: 102210.
Taha, Kamal, and Paul D. Yoo. 2016. Using the spanning tree of a criminal network for identifying its leaders. *IEEE Transactions on Information Forensics and Security* 12: 445–53. [CrossRef]
Tsikerdekis, Michail. 2016. Identity deception prevention using common contribution network data. *IEEE Transactions on Information Forensics and Security* 12: 188–99. [CrossRef]
Vogler, Nikolai, and Lisa Pearl. 2020. Using linguistically defined specific details to detect deception across domains. *Natural Language Engineering* 26: 349–73. [CrossRef]
Wang, Shanshan, Qiben Yan, Zhenxiang Chen, Bo Yang, Chuan Zhao, and Mauro Conti. 2017. Detecting android malware leveraging text semantics of network flows. *IEEE Transactions on Information Forensics and Security* 13: 1096–1109. [CrossRef]
Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. Paper presented at Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, October 6–8. pp. 347–54.
Zhou, Lina, Judee K. Burgoon, Douglas P. Twitchell, Tiantian Qin, and Jay F. Nunamaker, Jr. 2004. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems* 20: 139–66. [CrossRef]