INTERNATIONAL
HELLENIC
UNIVERSITY

Disinformation analysis based on linguistic features

Sofia Savvaidou

SID: 3305190047

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

JANUARY 2022

THESSALONIKI – GREECE

INTERNATIONAL HELLENIC UNIVERSITY

THESIS TITLE:

Disinformation analysis based on linguistic features

Sofia Savvaidou

SID: 3305190047

Supervisor:	Prof. Vassilios Peristeras
Supervising Committee Members:	Assist. Prof. Eleni Kapantai
	Assist. Prof. Ioannis Konstantinidis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

JANUARY 2022

THESSALONIKI – GREECE

Abstract

Disinformation meaning comes from its latin prefix dis- to the word information which ads a negative sign. People tend to research, find a solution and try to fix anything wrong. There are many tools created that detect deceptive language based on algorithms, artificial intelligence, etc. and some of them even reach 90% accuracy, but the result does not exceed the precision of human detection. For this aim, linguistic elements based on human language, speech and expression are important tools in the task of detecting misleading information and powerful weapons in the fight against misinformation. Our contribution in this battle is defining the specific linguistic features that are used in the different types of disinformation which are conspiracy theories, hoaxes, rumors, clickbait, misleading connection, fake reviews, trolling, fabricated, biased or one-sided and pseudoscience. This classification will help us deal with deceptive content and especially businesses. For example, it is important for marketing professionals to know if the reviews for the company they represent are fake in order to treat them properly. However, distinguishing true from fake is not an easy job and to achieve it right and effectively we must be very careful and constantly evolving.

Sofia Savvaidou

6/1/2022

Contents

1 Introduction

It is not like misinformation is a new phenomenon. Deceptive advertising (in business and politics), government propaganda, doctored images, falsified papers, and fake maps are all examples. Operation Bodyguard, a World War II deception effort designed to hide the planned site of the D-Day invasion, is a good example. In a successful attempt to persuade the Germans that a massive force in East Anglia was poised to assault Calais rather than Normandy, the Allies sent out fake radio signals and generated fraudulent military bulletins, among other deceptions (Farquhar, 2005, p. 72). Disinformation, on the other hand, appears to have been considerably more pervasive in recent years.

What makes today's information disorder different is the speed and worldwide reach it can achieve (Niklewicz, 2017), as well as the magnitude, complexity, and abundance of communication (Blumler, 2015). Through decentralized and dispersed networks, anyone may utilize digital media, particularly social media, to easily produce and circulate false information (Benkler et al., 2018). The purpose is often malevolent, with the objective of disseminating pre-determined views with possibly negative societal effects.

According to Bennett and Pfetsch (2018), this new, hyper-dynamic environment appears to usher in a new age in information flows and political communication, which necessitates a reformulation of research frameworks to account for conceptual effects from social media and digital networks.

Social media has evolved into a strong tool for freely interacting with others (Appel et al., 2020). Subscribers may simply utilize social media as a "megaphone" attracting attention thanks to social media, affecting people along with the damaging of the reputation of businesses (Herhausen et al., 2019). Social media, on the other hand, have created the conditions for the spread of problematic content as well as relevant channels for people to communicate freely (Choi & Sung, 2018). In addition, social networks is now a fertile ground for various sorts of misinformation (Allcott & Gentzkow, 2017; Appel et al., 2020).

We may say that fake news is one of the best known (and hazardous) social media phenomena. Since false news is expanding to traditional media as well, having a reliable method of detecting it is more important than ever. The internet is an open source where

you can find anything you need, yet it may also be deceiving at times. Articles, websites, films, and social media posts can all be used to try to influence people; they can also be used as a type of cyberwarfare between states, to raise someone's fame and influence, or to discredit their opponents. Due to the widespread usage of social media, fake news may quickly travel around the world, and it's sometimes difficult to track down the source of a story once it's been extensively disseminated. Fact-checking companies like Newtral, Snopes, and BuzzFeed can only handle a small part of the most popular rumors.

Due to the large amount of material published on social media, Manual Fake News detection is usually impossible or at the very least time consuming. As a result, automated methods are more adapted to detecting Fake News on a regular basis. Such methods can be classified based on the sources of their major characteristics, for example, some depend on linguistic cues while others use network analysis to uncover behavioral patterns (Aldwairi et al., 2018). Following feature extraction, both strategies use machine learning algorithms to solve the problem. In fact, depending on the attributes, they reveal if something is fake or real news.

Google, Facebook, and Twitter are just a few of the internet companies that have tried to solve the problem. Nevertheless, these attempts have seldom been successful in fixing the issue, as companies have responded to refusing people associated with these kinds of websites of the profit which would have derived from more visitors. People, however, are still dealing with websites that give inaccurate information and have a negative influence on the viewer's ability to interact with real news (Aldwairi& Al-Salman, 2011). The fact why firms like Facebook are involved in the issue of fake news is due to the fact that the establishment and subsequent proliferation of social media platforms has only added to the problem's severity (Westerman et al., 2014). In particular, most sites that feature this type of material, include a sharing option that encourages visitors to further share the web page's contents with others. Because social networking platforms allow for effective and speedy information transmission, users can swiftly propagate erroneous information. Regarding the data hack of millions of accounts by Cambridge Analytica, Facebook and other internet companies committed to do more to stop the spread of fake news (Smith et al., 2018).

Nowadays, people are finding it simpler to develop and spread false and misleading information thanks to new information technology (Hancock, 2007). Hackers have spread fake information on news websites like Yahoo! News and the New York Times (Fiore

& Francois, 2002). Also, websites that "impersonate" legitimate sources of information, such as Bloomberg News, have deceived investors (Fowler, Franklin, & Hyde, 2001). People can now modify visual images convincingly thanks to software (Farid, 2009). Furthermore, anybody with an internet connection can quickly (and anonymously) edit Wikipedia, "the free online encyclopedia that anyone can edit," to add false and misleading material. When the page on journalist John Seigenthaler was altered to imply that he was "actively implicated in the Kennedy assassinations," it was a prominent case (Fallis, 2008, p. 1665).

In each approach, disinformation can gain the ability to deceive people. The originator of most types of misinformation, such as lies and propaganda, intends for the information to be misleading. Other types of misinformation, such as conspiracy theories and false alarm calls, are deceptive merely because the source profits from their deception (Fallis and Don, 2015). Even though the means by which that function was achieved differs, all examples of misinformation have one thing in common: they all serve a purpose. And it's no coincidence that the information is deceptive, regardless of how that function was gained.

Given that individuals utilize cues to comprehend and form perceptions of others, language plays a critical part in social relationships (Xu & Zhang, 2018). In fact, the way people use words reflects their personal characteristics (Hirsh & Peterson, 2009), identities (McAdams, 2001), and emotional states (Tausczik & Pennebaker, 2010). Lately, marketing research has begun to emphasize on the impact of text linguistic style on social media virality, with particular attention paid to the use of function words (Aleti et al., 2019), pronoun choices (Labrecque et al., 2020), and text narrative style (Van Laer et al., 2019).

It is important to look at how these conceptions disseminate online on platforms like Twitter, which is one of the most popular places for people to discuss current issues (Kietzmann et al., 2011, Rust et al., 2021). Understanding the origins, characteristics, and evolution of viral misinformation is an important role of social media surveillance (Di Domenico et al., 2021), and timely study is required to counteract the spread of disinformation (Ahmed et al., 2020; Chang et al., 2020; Chou et al., 2018). The language and writing styles we employ reveal our psychological systems and our personalities (Berger et al., 2020; Humphreys & Wang, 2018; Netzer et al., 2019; Pennebaker et al. 2010). In reality, not only the content (what you say) but also the tone (how you say it),

reflects how individuals interpret the world and influences the listener (e.g., Aleti et al., 2019; Bertele et al., 2020).

Markowitz and Hancock (2014) achieved a 71.4 % accuracy rate by employing linguistic cues to categorize scientific papers as fake or genuine. Adjective, amplifier, and diminish her frequencies, as well as certainty term frequencies, were shown to be the most descriptive characteristics in the studied dataset. Relevant characteristics identified in this study have also been proven to be beneficial in the identification of fake reviews as well as studies on reality monitoring. To distinguish legitimate from fake news, Hardalov et al. (2016) used a mix of linguistic, credibility, and semantic factors. (Weighted) n-grams and a normalized number of unique words per article are linguistic characteristics included in their work. Capitalization, punctuation, pronoun use, and sentiment polarity characteristics created from lexicons were all borrowed from the literature as credibility features. On the basis of self-created datasets, all feature categories were examined separately and in combination. In two of the three examples, the greatest results were obtained by utilizing all available characteristics.

In addition to language or environmental factors, argumentation and textual structure can be employed to identify false information. By assessing posts solely on the basis of word similarity, Lendavi and Reichel (2016) studied how discrepancies in rumored micro-post sequences might be uncovered. The authors propose that even for small and loud texts, word and token sequence overlap scores may be used to produce authenticity evaluation cues. Furthermore, Ma et al. (2015) built on previous research by tracking changes in the linguistic features of messages over the course of a rumor's existence. They were able to show solid results in the early identification of a developing rumor using SVM (support vector machine) based on time series characteristics.

The COVID-19 pandemic in 2020 provided a further more chance to manipulate public perception (World Health Organization, 2020). Since many people were compelled to remain into their houses this time, they were online and actively followed news and engaged in public conversation on different social media platforms, when they faced a variety of misinformation and rumors that quickly spread throughout the world (Frenkel et al., 2020). Users are free to submit a wide variety of content, engage with anybody, and create numerous sorts of groups, agendas, and debate subjects under no expense, making public conversation favorable to deception and rumors (Kirman, 2012).

Provocative headlines, unproven facts, imprecise wording, and disinformation are just a few of the issues that should be the focus of a pandemic's research of receivers' media

literacy. The influence of the media on society's mental health (Riles, 2019) should be closely monitored during the pandemic phase. Another issue is finding a way out of the so-called post-pandemic information syndrome, in which faith in the media is not eroded in the coverage of genuine events, when social networks take center stage in expert opinion and authority, and news is not seen as "fake" ahead of time. "Viral news" is spreading alongside the virus, causing harm to the current recipient's view of so-called COVID news.

According to a survey conducted by Shevchenko et al. (2021), headlines containing lexical manipulative resources are the most popular (44 percent), whereas headlines including phonetic manipulative resources are less impacted by the receiver. The majority of those questioned are still willing to double-check the accuracy of information in a media text (the selection of fake news was deliberate). Some recipients (22%) pay attention to fake news because of a catchy title (or because they believe a reliable media source) and are willing to share it without verifying it. Because of the widespread interest in the COVID-19 story, journalists have utilized both objective and false news to sway the public. As evidenced in the survey, even if the popularity of news or media sources covering the COVID-19 pandemic declines, as indicated by this poll, journalists "artificially" resort to the inclusion of the phrase "COVID-19" in the text, to which receivers actively respond.

It has also been stated in the past that ignorant communities (for example, anti-vaxxers) employ far more pronouns, implying a highly narrative discourse structure. According to Memon et al. (2020) analysis, in contrast to misinformed users, informed users in the COVID-19 discourse use significantly more pronouns, more functional words, mention more family-related keywords, are less analytical, and are more authentic and honest. All of this suggests that knowledgeable users are likely to utilize many more narratives than those who are ignorant.

The ongoing struggle against fake news on COVID-19, as well as the uncertainties surrounding it, demonstrates the necessity for a hybrid strategy to detecting fake news. In this process, both human expertise and technology must be employed (De Beer and Matthee, 2020). Several of these safeguards, hopefully, will stay that way, and digital media platform owners and the public will be held responsible for recognizing and combating fake news.

2 Structure

The following is how the rest of the article is organized as follows. In the "Methodology" section we analyze the approach that we chose to follow in order to realize the systematic literature review. In "Categories of linguistic features" we mention our findings of the categorization of linguistic features according to a variety of scientific papers and books. As about "Detection systems based on linguistic features" we examine and determine the different detection tools that exist and are based on linguistic features. The chapter "Linguistic features in different types of disinformation" is and the classification of the linguistic features in disinformation typology that we accomplished after extensive research. In the "Results" section we present the results of our SLR and the useful tables that we made out of this. Finally, in "Conclusion" and "Limitations and Future work", we present our conclusions, limitations we had during the research and some ideas for future work.

3 Methodology

In order to address our research questions, we conducted a systematic literature review about disinformation analysis based on linguistic features. The information collection was done via popular scientific databases like Science Direct, Research Gate but mainly through Scopus.

The query that was used to collect our data was:

TITLE-ABS-KEY (("linguistics" OR "linguistic features" OR "linguistic analysis" OR "computational linguistics" OR “stylometric features1” OR “linguistic style”) AND ("disinformation" OR "fake news" OR "fabrication" OR "propaganda" OR "pseudoscience" OR "conspiracy theory" OR "hoaxes" OR "fake reviews" OR trolling OR clickbait)) AND (LIMIT-TO (LANGUAGE , "English"))

The above query gave us the result of 522 scientific books and papers, but by keeping only the documents that are chronologically from 2010 to 2021, we scanned the titles and the abstracts of 484 scientific articles. After a rapid look over all those documents, we excluded most of these because of the lack of relevance with our topic and we found useful the information of 65 different papers. Moreover, we include only the papers written in English.

4 Categories of linguistic features

4.1 Linguistic features categories according to different researchers

Different researchers follow different approaches that can lead to additional categories of linguistic features (the studies are presented on a chronological order):

Linguistic features categories based on the fundamentals of speech

- *Burgoon et al. (2003),* divides linguistic features in terms of quantity (syllables, words, sentences), vocabulary complexity (big words, syllables per word), grammatical complexity (short sentences, long sentences, Flesh Kincaid grade level avg of words per sentence, sentence complexity, number of conjunctions), specificity and expressiveness(emotiveness index, rate of adjectives and adverbs, affective terms).
- *Newman et al. (2003),* proposes a different approach of linguistic dimensions such as dictionary words, words with more than six letters, total pronouns, first person singular, total first person, total third person, negations, articles, and prepositions.
- *Zhou et al. (2004),* distinguishes linguistic characteristics in the same way Burgoon et al. does, in terms of quantity (words, verbs, noun phrases, sentences), complexity (avg clauses, avg sentence length, avg word length, avg noun phrase length, pausality), uncertainty (modifiers like modal verbs, uncertainty, other reference), non immediacy (passive voice, objectification, generalizing terms, self reference, group reference), expressivity (emotiveness, lexical diversity, content word diversity, redundancy, typographical error ratio), specificity (spatio-temporal information, perceptual information) and affect (positive effect, negative affect).

Linguistic features categories when writing an essay

- *Mary J. Schleppegrell (2006)* describes linguistic resources for writing an *essay* as nominal expressions for naming the points to be made, verbs that construct relational processes for defining key terms and projection, modality for constructing possibility and necessity in making judgments, markers of consequential relationships (purpose, condition, cause, concession) for drawing conclusions or supporting assertions, thematic choices that enable smooth progression in presenting information and internal connectors for signposting the organization of the text.

Linguistic features categories based on the nine parts of speech

The *nine parts of speech* that are typically taught in *schools* are noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. There are, however, a plethora of other categories and subcategories. In addition, the components of speech in most foreign languages differ from one another:

- According to *Mahyoob et al. (2020)*, the relative frequencies of all the tested linguistics attributes in Politi-fact site articles are personal pronoun, proper noun, adverb, to-infinitive, stative verb, passive voice, reported speech, comparative adjective, modal verb, superlative adjective, interrogative, conjunction, long sentence, negation, and quotes.
- *Syntax-based* features, according to *Choudhary et al. (2021)*, are made up of numerous linguistic dimensions. The extraction of essential evidence, such as count statistical evidence, sentence sentiment data, and grammatical property evidence, takes place in this step:

SYNTAX-BASED FEATURES	**DESCRIPTION**
CHAR COUNT	TOTAL NUMBER OF CHARACTERS WITH AND WITHOUT SPACES
WORD COUNT	TOTAL NUMBER OF WORDS IN A GIVEN SENTENCE
TITLE WORD COUNT	COUNT THE NUMBER OF WORDS IN A GIVEN TITLE
STOP WORD COUNT	COUNT THE TOTAL NUMBER OF STOP WORDS IN A GIVEN SENTENCE
UPPER CASE WORD COUNT	COUNT THE NUMBER OF UPPERCASE WORDS IN A GIVEN SENTENCE
WORD DENSITY	NUMBER OF OCCURRENCES OF THE CHOSEN KEYWORD OVER THE TOTAL NO. OF WORDS IN A GIVEN TEXT.

Making a hypothesis, the sentiment is expressed in a news article that depends on decision-making elements in the classification of news as fake or not fake. The following are the main sentiment-based qualities that are associated to the evaluation or strength of specific emotions:

SENTIMENT-BASED FEATURES	DESCRIPTION
POLARITY	IT REFERS TO POSITIVE AND NEGATIVE STATEMENTS. IT LIES IN THE RANGE OF [-1,1]
SUBJECTIVITY	EXPRESSING AN OPINION, VIEWS, OR A PERSON'S FEELINGS. IT LIES IN THE RANGE OF [0,1]

Subjectivity refers to the difference between objective and subjective manifestations of sentiment. Objective expressions are facts, whereas subjective expressions are a person's thoughts, beliefs, or feelings about a certain issue. Subjectivity and polarity, for example, are 0.9 and 0.81 for the statement "Donald Trump is a great politician."

Grammatical characteristics, which are retrieved using parts of speech (POS) tag evidence features, are a crucial aspect in inspecting true and false news. For the targeted problem, noun, verb, adjective, and pronoun are plausible attributes to determine its authenticity out of all parts of speech (POS) features. These traits are intended to detect deception indicators in writing style in order to distinguish fake news. Details are shown below:

GRAMMATICAL-BASED FEATURES	EXAMPLE	REPRESENTATION
NOUN	TRUMP SAYS NOBODY REALLY KNOWS IF CLIMATE CHANGE IS REAL.	TRUMP
VERB		IS
ADJECTIVE		REAL
PRONOUN		IF, NOBODY
ADVERB		REALLY

The model is built on in-depth linguistic analysis, which aids in the researcher's understanding of language structure, writing style (word density and word count), grammatical use (POS tag), and reading ability (ARI, Gunning

fog).Sentiment analysis gathers subjective information from news content and is heavily weighted in the identification and classification of fake news. In this study (Choudhary et al., 2021), the linguistic characteristics are retrieved based on literature. Various computational techniques were utilized to extract language aspects for the news dataset, including syntactic, grammatical, emotive, and readability factors.

4.2 Linguistic features extracted from Twitter

Deng et al. (2021) identifies six linguistic elements that impact brand engagement, as measured by *Twitter likes and retweets*: post length, language complexity, visual complexity, emotional cues, interpersonal signals, and multimodal cues in rich media. They can call linguistic traits such as hashtag (a metadata tag preceded by the hash symbol, e.g., #TimeToAct), cashtag (a corporate ticker symbol preceded by the U.S. dollar sign, e.g., $TSCO), and URL (a reference to a web site, e.g., ibm.co/3jYd9IH) social media-specific qualities. These features were developed by social media platforms to assist users in dealing with the massive volume of data available. For example, *hashtag* and *cashtag* can help users find postings about comparable subjects or companies, follow conversations, and promote their ideas more effectively. *URLs* can be used to link to external resources that can give further information to users. These characteristics have become essential to social media communications because they make it easier for users to find, share and engage with content (Davis et al., 2019; McShane et al., 2019).

Another study by *Zhou et al. (2021)*, aims to extract linguistic features from misinformation material and examine how those qualities impact misinformation propagation behavior in users on social media. The phrases that may induce misinformation propagation behavior in various ways are the features explored. They look at *four language features*: persuasive words, comparison words, emotional words (both positive and negative emotion words), and uncertainty words. Furthermore, Yin and Zhang (2020) show that a piece of information's representational richness might alter its credibility and argument quality. Users may now readily generate and use multimedia material (such as text-only, text and photos, or text and videos) thanks to social media. The richness of information, according to Chen et al. (2020a), is described as "the display format level of information on social media." Visual material is more likely to be shared and liked on social media than plain text stuff (Chen et al., 2020a). As a result, misinformation rich-

ness is proposed in this study, and its moderating effect in the link between the four linguistic qualities and misinformation transmission is investigated.

Zhou's et al study (2021) divides emotional words into two categories: positive and negative emotion words. Emotion is not considered as a binary variable in this study for the following reasons: First, this treatment is consistent with Long et al. (2017) in terms of the other three linguistic characteristics; second, the misinformation creator frequently uses both positive and negative emotion words at the same time (Long et al., 2017), and previous research has shown that both positive (Wang et al., 2017a) and negative emotion words increase the spread of information (Zhang & Qu, 2018). As a result, if the emotion is considered as binary, it is impossible to investigate if positive and negative emotion words could enhance the spread of disinformation.

A comparison analysis *(Verma et al. 2021)* of five linguistic feature categories is shown below:

1. *Readability index* quantifies the text's complexity (i.e., reading difficulty) based on word length, number of syllables, and sentence length.
2. *Psycho-linguistic* features describe emotions, behaviors, persona, and mindset.
3. *Stylistic* features explain the style of a sentence.
4. *User credibility* features describe user information.
5. *Quantity* features explain sentence information such as the number of words and number of sentences.

4.3 Linguistics in Fake News

In terms of substance, fake articles are far shorter, with fewer technical words, smaller words, less punctuation, fewer quotations, and more lexical repetition. They also utilize simpler language, which results in fewer analytic terms, more personal pronouns, fewer nouns, and more adverbs on a linguistic level (Horneetal., 2017). Finally, the emotional response that the essay aims to elicit is a good indicator. Emotional words and phrases attract greater attention and spread more quickly (Ruchanskyetal., 2017).

The type and number of words used in Fake News

Rashkin et al. (2017) demonstrated that fake news used more subjectives, superlatives, and modal adverbs, all of which may be utilized employed to exaggerate. Comparatives, money, and numbers — words used to convey specific data – occur more frequently in

accurate news. This adds to Ott et al.'s (2011) results on the distinction between superlative and comparative use.

Horne and Adali (2017) discovered how there is a difference between real and fake news. Fake news articles have a shorter content and utilize less punctuation than actual news articles, while fake political articles had greater lexical variety than real political publications. Fake news stories also include less analytic terms. In one dataset, they discovered that reading false news stories requires a lower educational level, but the converse is true for the GossipCop dataset. Fake titles with additional appropriate nouns have been discovered. In BuzzFeedNews and GossipCop, fake titles contain more proper nouns than actual titles and fewer nouns. Fake political titles are also lengthier, include more capitalized words, and contain fewer words, according to research.

Furthermore, Shrestha and Spezzano (2021) findings reveal new patterns that were not apparent in Horne and Adali's investigation. They discovered that real news articles utilize more nouns, determinants, wh-determinants, verbs, past tense verbs, Wh-pronouns, and adjectives, as well as a more positive tone. They additionally notice that fake titles reflect more negative emotions (anger, sadness, fear, and contempt) as well as more negative sentiment than real titles. This trend holds true for the corpus of fake news as well. Real titles, on the other hand, tend to communicate more positive feelings and sentiment (trust, posemo, joy). People are sensitive to negative information while choosing information. People are more likely to pay attention to bad news because of this negativity bias, therefore fake news tiles, bodies, and even linked pictures that portray negative emotions are more appealing and circulate more widely.

Syntax-based features are formed of numerous linguistic dimensions that fulfill a specified pattern to identify false news, according to Choudhary et al. (2021). The extraction of essential evidence, such as count statistical evidence, sentence sentiment data, and grammatical property evidence, takes place in this step. During the creation of false news, the developer purposefully uses title words, uppercase terms, and even considers content length and word density. Because of these essential traits, digital natives are drawn to and affected by news, making statistical proof of false news content a significant and notable aspect.

Grammatical characteristics, which are retrieved using parts of speech (POS) tag evidence features, are a crucial aspect in inspecting true and false news. For the targeted problem, noun, verb, adjective, and pronoun are plausible features to identify its validity

out of all POS features (Choudhary et al., 2021). These traits are intended to detect deception indicators in writing style in order to distinguish fake news.

There are certain linguistic indicators for lying. Several researches comparing honest and dishonest speech discovered that when it comes to lying, the regular lexical composition of utterances alters (Arciuli et al. 2010, DePaulo et al. 2003, Hancock et al. 2008, Newman et al. 2003, VanSwol et al. 2012). For example, liars employ more negations and generalizing phrases such as always, never, nobody, or everybody than serious speakers, according to Vrij (2008, pp. 102, 112–14); liars also use less self-references such as I, me, or mine, and have lower lexical variety (i.e., the number of different words in a statement divided by the total number of words used in that statement).Furthermore, liars are said to use fewer emotive terms (e.g., sad) and more motion verbs (e.g., go) than honest speakers (Van Swol 2014, p. 608). The reason for their selectiveness is that the liars want to maintain control over how they present the facts and hide their true intentions (Meibauer & Jörg, 2018).

Intentionally false information is less verifiable, which is rather unsurprising. Lasorsa and Dai (2007) compare deceptive news stories against genuine news articles and conclude that false news stories frequently deal with themes that encourage secrecy, which might mask the lack of reliable facts. The latter also has to do with the usage of sources, which refers to people who offer information or quotations for news stories. Intentionally false tales include more allusions to sources, but they are rarely presented in a fashion that allows them to be traced. Because sources are often anonymous and unclear, no identifiable identities are supplied (Bonet-Jover et al. 2021). Others point to the use of questionable and conspiratorial sources without fact-checking (Bradshaw et al. 2020; Marchal et al. 2019; Neudert, Howard, and Kollanyi 2019). The use of unverifiable sources may also be seen in language traits, such as the usage of pronouns rather than particular source names. Furthermore, sources are used less frequently for direct quotes, since quoted material in false news is found to be lower than in real news (Reddy et al. 2020). A lack of reliable data corresponds to research that suggests source cues are becoming less and less essential for how people access and digest information about current affairs, which might lead to increased consumption of false messages (e.g., Kalogeropoulos, Fletcher, and Nielsen 2019).

When reporting fraudulent information, the language is more sociable and pleasant, displays greater conviction, and concentrates on current and future activities. Similarly,

Benjamin D et al. (2017) attained accuracies of 71%–78% in their study to distinguish false from true news from text. The authors' method leads to general results such as: headlines are a key differentiator between false and real news, and legitimate news articles convince people via good reasoning, but fake news pieces persuade users using heuristic. This is a significant result that opens up possibilities for future study into quantifying the gap between the title and substance of an information piece and uses it as a misleading information indication. This research reveals a concerning tendency. As a result, for the automatic identification of false news, a method uses the title as the primary element. It's known as "stance-based detection" (Lahlou et al., 2019).

4.4 Linguistic traits in Fake News that sway the public

Capitalization, pronouns, modal verbs and casual language

Additionally, a number of elements related to the precise language employed in intentionally fake news-like articles or social media postings may be detected based on the examined literature. To summarize, purposefully misleading messages are more likely to employ capitalization, pronouns, and casual language or swearing. Furthermore, there are three linguistic traits that are often investigated but for which there is no unequivocal proof. The literature provides differing cues for lexical diversity, text length, and punctuation use. There's evidence that purposely fake news-like texts use a lot of capitalization, both in the title and in the body of the piece, to draw people's attention (e.g., Bradshaw et al. 2020; Marchal et al. 2019; Neudert, Kollanyi, and Howard 2017; Reddy et al. 2020). This might also apply to tweets, where an overabundance of capital letters (more than 70%) has been identified as a sign of fake material (Srivastava, Rehm, and Schneider 2017).

Modal verbs (would, could, might, etc.) express uncertainty and are frequently employed by deceivers. This occurs because deceivers are unsure of the information they spread, therefore they prefer to conjecture and suggest linkages between occurrences that aren't clearly linked (Zhou et al., 2003). Their findings reveal that modal verbs are used more frequently in fake news.

First-person and second-person usage

In authentic articles, the usage of first-person singular is more common. This is due to the fact that deceitful authors strive to distance themselves from the material they spread (Zhou et al., 2003). On Kasseropoulos and Tjortjis. (2021) research it is found, as expected that authentic news use the first person more frequently, whereas fake news use the third person, plural and singular, more frequently.

Rashkin et al. (2017) discovered that first-person and second-person pronouns are more frequently utilized in misleading news texts when applied to news material. The writers use journalistic methods to explain the distinction. Editors of reliable sources are likely to be fairly strict about deleting wording that appears to be too personal, but similar methods are not used in the creation of falsified news articles. In a similar line, Asubiaro and Rubin (2018) discover that in false news, the usage of (all sorts of) pronouns is often higher. Informal vocabulary and phrasing are often kept to a minimum in regular news reports. Slang and curse ("inflammatory language") was found to be rather common in purposely fake news pieces (e.g., Asubiaro and Rubin 2018; Gupta et al. 2014; Neudert, Howard, and Kollanyi 2019; Rashkin et al. 2017; Zhou and Zafarani 2020).Hameleers, van der Meer, and Vliegenthart (2021) undertake a content analysis of misleading assertions fact-checked by Politifact.org and Snopes.com in a recent work. They discovered that material that is wholly untrue is more likely to contain hate speech and incivility. As a consequence, Damstra et al. (2021) come to the conclusion that hate speech and incivility may be used to detect deception.

5 Detection systems based on linguistic features

The WELFake system

According to Verma et al. (2021), the distribution of fake and real news in the *WELFake dataset* is balanced across all of the following four feature categories, as it is shown below:

1. The number of short sentences (less than 10 words) describing real news outnumbers those describing fake news.
2. Fake news text readability is lower than real news text readability.
3. Fake news stories have a higher level of subjectivity than real news.
4. The number of articles containing real news is higher than the number of articles containing fake news.

Further research

There are variations in the words capture based on real and fake news from different views, which not only learn the conflict aspects typically recorded by existing models, but also get the differences from strong emotion and writing styles perspectives. True news use more emotional smooth words, such as “delightful”, “joyful” and “sad”, but fake news use more negative or extreme terms, such as “extremely”, “rage” and “manic”, demonstrating that pattern-shared unit successfully boosts category-differentiated characteristics at emotion level. Both patterns are able to catch some words in various ways. The unit for fake news, in particular, might capture terms like “shocking”, “unexpected” and “totally”, whereas the unit for legitimate news focuses more on objective style related words, like “should”, “consider” and “reported” (Wu et al., 2021).

Burgoon et al. (2003)’s 16 language variables

The use of language signals for detecting fraud in written narratives was a popular strategy that gained traction in the mid-2000s. Experiments conducted by psychologists in collaboration with linguists and computer scientists revealed that potential deceivers use specific language patterns, such as short sentences, phrasal verbs in abundance, certain tenses, and so on (Burgoon, Blair, Qin, & Nunamaker, 2003; Hancock, Curry, Goorha,

& Woodworth, 2007; Newman, Pennebaker, Berry, & Richards, 2003; Tausczik & Pennebaker, 2010; Zhou, Burgoon, Nunamaker, & Twitchell, 2004).

In particular, Burgoon et al. (2003) investigated 16 language variables that might assist distinguish between deceitful and genuine messages. To construct a database, they conducted two tests in which, on either face to face or on computer-based discussions, one volunteer acted as the deceiver and the other acted honestly. The writers then transcribed the dialogues for additional study, and they concluded to certain language cue classes that may disclose the deceiver. They employed the C4.5 Decision Tree technique with 15-fold cross-validation to cluster and produce a hierarchical tree structure of the suggested characteristics. In a short sample of 72 cases, their method's overall accuracy was 60.72 %. Grammatical Complexity, Vocabulary Complexity, Quantity, and Specificity/Expressiveness are the four categories of features proposed (Gravanis et al. 2019).

5.1 The rising problem of misleading texts and the demand for linguistic research linked to them

According to Mahyoob et al. (2020), reported speech, passive voice, negation, and proper nouns are the four most commonly employed linguistic elements in misleading articles. In these pieces of writing, the to-infinitive, modals, and long sentences are the least employed.

Corpus linguists have shown that the structure of language varies consistently depending on its communicative goal, based on the examination of huge samples of spoken language. People utilize more past tense verbs and third person pronouns while telling stories. People employ more nouns and prepositions while giving explanations. People utilize more questions and interjections when they interact. The aim of a text is reflected in its grammar. This is why, rather than its content, the language of fake news – its structure – might be the key to detecting it. An author attempting to educate the public has a completely different purpose than one attempting to mislead the public. This distinction in communicating intent will have linguistic impact. The challenge is to figure out what these implications are, in order to describe the fake news style.

Even though fake news is omnipresent, it is difficult to tell what's genuine and what's not, especially when people debate so fiercely over what's real and what's not. Many

organizations are compiling databases of fake news, mostly through fact-checking, but their primary goal isn't to create a balanced corpus for linguistic analysis, and these collections inevitably reflect the curators' political prejudices. It is also uncertain how to gather a similar sample of real news.

The major concern here is how to assess the purpose of millions of anonymous internet authors, and problems like these are crucial to knowing what distinguishes false news. Consider a research in which the false news consist of sensationalist stories shared on social media, while the true news consist of more serious articles published by established news organizations. Although a linguistic analysis of these two datasets would surely reveal differences in formality, it would be incorrect to suppose that this is the only difference between honest and dishonest reporting.

According to Jack Grieve's (2018) preliminary study, the solution to these issues is to create corpora that reflect the whole internet news media scene, not just false news. We can only begin to grasp how the language of fake news and genuine news differs by collecting material from a variety of formats, outlets, markets, writers, themes, and opinions, and then submitting this data to meticulous linguistic analysis.

The fact that people handle the discomfort produced by lying by separating themselves from the misleading message they constructed is one example of the psychological side effects of deceit (DePaulo et al., 2003). Psychological distancing is manifested by a decrease in self-reference (e.g., "I," "me," "myself") and an increase in group reference (e.g., "they," "he"), both of which are methods that imply a lack of commitment to the misleading assertion (DePaulo et al., 2003; Hancock et al., 2007). These pronouns serve as powerful linguistic cues for deception (Addawood et al., 2019).

According to a research, Russian troll accounts that discussed the 2016 US election on social media used misleading language to sway public opinion and promote false political information. A classifier was created to detect these trolls using several false language signals, and it had a high accuracy rate.

5.2 Software systems that detect fake news and their use in various studies

Because most communication is text-based and done asynchronously, social media that focus heavily on content are particularly vulnerable to deceit. An increasing body of ev-

idence shows that counting and classifying the words individuals use to communicate, whether spoken or written, can reveal a lot about their underlying ideas, feelings, and intentions. Several research on deception detection have shown that linguistic cue identification is successful, as truth-tellers' vocabulary differs from that of deceivers (Larcker and Zakolyukina, 2012).Previous research has examined deceptive language in a variety of contexts, including fake reviews (Ott et al., 2011; Feng, Banerjee, and Choi, 2012), online games (Zhou et al., 2004), online dating profiles (Toma and Hancock, 2012), interview dialogues (Levitan, Maredia, and Hirschberg, 2018), and controversial topic opinions (Mihalcea and Strapparava,2009).

Changes in word quantity, pronouns, emotive phrases, and differentiation markers may imply dishonesty, according to literature on linguistic analysis of deception (Burgoon et al., 2003; DePaulo et al., 2003). Furthermore, expressing emotions, particularly negative emotions, has been connected to deceit (Zhou et al., 2004; Burgoon et al., 2003). Parts of speech tags and other syntactic elements have been shown to be beneficial for structured data (Ott et al., 2011; Feng, Banerjee, and Choi, 2012). New approaches to evaluate such data have evolved as a result of past research on deception detection using language, such as building software that can automate the identification of linguistic indicators.Linguistic Inquiry and Word Count (LIWC) (Pennebaker and King, 1999) is one of the most well-known software systems for text-based deception detection. LIWC organizes words into psychologically driven categories. LIWC coding's fundamental concept is text categorization based on truth criteria. The LIWC has been widely used to investigate deception detection (Vrij, 2000; Hancock et al., 2007; Mihalcea and Strapparava, 2009). When traditional classification techniques like decision trees and logistic regression are used to identify dishonesty, it obtains a 74 % accuracy rate (Fuller,Biros, and Wilson, 2009).The classifier gets an average accuracy rate of 70% when utilizing known psycholinguistic lexicons such as LIWC for detecting misleading beliefs (Mihalcea and Strapparava, 2009). Human judges, on the other hand, only succeed 50-63 % of the time in detecting dishonesty (Rubin and Conroy, 2011).

Rashkin et al. evaluated linguistic elements such personal pronouns and swear words that were retrieved using the LIWC lexicon. Other researchers investigated how fake news affected people's emotions. Vosoughi et al., for example, found that false rumors elicited fear, disgust, and surprise in responses, but real rumors elicited joy, sadness, trust, and anticipation. Giachanou et al. (2021) introduced emoCred, an LSTM-based

neural network that used emotions from text to distinguish between credible and non-credible articles, demonstrating the importance of emotions in credibility detection. For false news identification, Wang suggested a hybrid CNN that combines user metadata with text.

Some scholars have looked into the language used in various forms of deception. Addawood et al. (2019) examined the language used by Russian trolls during the 2016 US presidential election to try to sway public opinion. They discovered and analyzed 49 linguistic signals as possible markers of misleading language in their study, and found that the frequency of hashtags and retweets are the most crucial indications of trolls. Giachanou et al. (2021) employed common language dictionaries to detect linguistic signals, such as LIWC, which has been used in a variety of prediction tasks, including gender and age prediction and prediction of changes in mental and physical health. They also use LIWC to extract linguistic data in a similar way as those works.

LIWC is a text analysis tool that generates features for 93 semantic groups based on normalized word counts. Many researchers have employed LIWC factors to predict outcomes such as personality (Pennebaker and King, 1999), deception (Newman et al., 2003), and health (Pennebaker and King, 1999). (Pennebaker, Mayne, and Francis, 1997). With the exception of word count, words per sentence, and question marks, which are reported frequencies, LIWC calculates the percentage of each variable type by dividing the observed variable's frequency by the total amount of words in the sample. Except for LIWC characteristics computed as percentages, the number of tweets each of them submitted normalizes other features computed for the users.

According to Addawood et al. (2019), political trolls use deception in order to fool people about their genuine intentions. The deceiver pays an emotional and cognitive price for deception, which may frequent lye seen in the language used to deceive. Studies investigated at the deceiver's physiological reactions using highly-trained experts and behavioral coding, as well as applying content-based criteria to written transcripts for deception detection (Zhou et al., 2004). Then, to evaluate the linguistic profiles of misleading language, automated linguistic tools were created to analyze the linguistic features of texts (Newman et al., 2003; Zhou et al., 2004).

Due to the lack of facial expressions, gestures, and body posture and distance standards in Twitter communications, someone can rely only on the text to deduce human ideas and attitudes and verify message trustworthiness. Furthermore, earlier research has

identified dishonesty as a trait that may be assessed using linguistic signals (Tsikerdekis and Zeadally, 2014). Automated linguistic approaches, which utilize computer algorithms to assess text's linguistic features, have recently been applied to investigate the linguistic profiles of misleading language—see (Newman et al., 2003; Zhou et al., 2004; Bond and Lee, 2005).

Deceivers, according to the IDT (innovation diffusion) hypothesis, employ less organized and evasive language. Truth-tellers, on the other hand, are more confident in their remarks. Certain linguistic markers, such as "always" or "never," are powerful indications of sincerity (Levitan, Maredia, and Hirschberg, 2018; Rubin, Liddy, and Kando, 2006). Subjective language has been proven in the past to aid in the recognition of certainty in textual information (Rubin, Liddy, and Kando, 2006). Deceivers use more modifiers and modal verbs in their writing than truth tellers to indicate their doubt (Zhou et al., 2004; Buller and Burgoon, 1996). More uncertainty has been connected to the greater usage of hedges (Rubin, Liddy, and Kando, 2006; Levitan, Maredia, and Hirschberg, 2018).

Distinct language patterns (e.g., use of personal pronouns, swear words) between conspiracy and anti-conspiracy propagators' tweets must be analyzed and compared in order to find the qualities of conspiracy and anti-conspiracy propagators.

Additional studies

Other studies concentrated on identifying users who spread fake news. Shu and Wang (2018) researched user profiles to see if they shared fraudulent or authentic news. According to the findings, there are differences in characteristics (such as registration time) between those who share false news and those who share legitimate news. They also investigated how successful those qualities are in detecting false news, and discovered that combining user profile variables with document psycho-linguistic characteristics can be quite effective at detecting fake news. In contrast to their study, they concentrate on conspiracy theorists and conduct a more in-depth investigation of the users' characteristics. They also filter out accounts that are most likely trolls or bots, with the goal of analyzing the characteristics of actual people who share posts that support well-known theories (e.g., vaccines lead to autism).

Another study by Vo and Lee (2019), examined the linguistic features of fact-checking tweets (tweets that affirm that an article is false) and developed a deep learning frame-

work for generating fact-checking answers. According to their findings, fact-checkers are more likely to rebut fake news and use formal language. Furthermore, fact-checker tweets focused on what happened in the past, whereas random tweets focused on the present and future. Giachanou et al. (2021) developed a CNN-based system for distinguishing between prospective false news spreaders and fact-checkers based on a variety of psycho-linguistic variables and inferred personality qualities of the users. El Azab et al. (2015) investigated the efficiency of several characteristics and identified the most essential ones for detecting fake accounts. Klein et al. (2019) also investigated individuals who posted conspiracy theories on Reddit to evaluate whether there were any changes in the language and social surroundings they employed compared to other users. Finally, Rangel et al. (2020) organized a shared evaluation task in which participants had to create a system that could detect whether or not a person is a potential false news spreader.

Furthermore, to create their data, Giachanou (2021) employed the Twitter Application Programming Interface (API) to gather tweets on some of the most well-known conspiracy theories. They attempted to come up with a hashtag for every conspiracy theory that supports it (for example, #vaccinesCauseAutism versus #vaccinesWork). This was not achievable for all hashtags, though.

Psycho-linguistic characteristics include the following:

- Personality traits: the inferred personality traits of the user
- Sentiment: the polarity of sentiment conveyed in a user's tweets
- Emotions: the number of times a person has conveyed emotion in a tweet.
- Linguistic patterns: the number of various linguistic patterns that may be found in tweets.

Every writing may use a simple or complicated linguistic style (Tausczik & Pennebaker, 2010). Even while tweets are often brief, this is also the case (Kietzman et al., 2011). Users of deceptive speech tend to talk and write in a less complicated manner, using fewer words, fewer propositions, fewer big words, and less terms associated with cognitive functions (words like “think”, “know” and “question”) (Tausczik & Pennebaker, 2010).

The use of PrivacyDIC and ConspiracyDIC

According to Visentin et al. (2021), text complexity in misleading speech may be lowered due to the cognitive load required to sustain a tale that contradicts experience and the effort required to persuade others that something untrue is true. Despite the fact that prior research has failed to provide empirical proof or direct causal links between text complexity and favorable behaviors toward it, we may claim that the more complicated a text is, the less persuasive it will be and, as a result, the less likely it will be shared.

The amount of words, the density of prepositions, the density of words with more than six letters, and the density of terms related with cognitive mechanisms all have a negative impact on the chance of a tweet being retweeted. Because PrivacyDIC and ConspiracyDIC are in Italian, Visentin et al. (2021) collected tweets in that language. They chose two separate contexts for privacy concerns and conspiracy theories, which is noteworthy because no scenario that had both privacy issues and conspiracy theories was available when the authors undertook these additional studies.

Previous research has proven that language traits may be used to identify fact from fiction (Clarke et al., 2020; Luca & Zervas, 2016; Purda & Skillicorn, 2015; Zhang & Ghorbani, 2020; Zhao et al., 2020). The work by Zhou et al. (2021) adds to the body of knowledge on detecting disinformation in social media by providing new sorts of linguistic features. The validity of all four types of linguistic characteristics described in identifying disinformation is additionally confirmed in this investigation. This study finds that persuasive, comparative, and uncertainty words are practical and effective in detecting disinformation on social media, as compared to previous research. These findings imply that, rather than depending on subject, sentiment, and high-dimensional textual features (e.g., "n-grams"), other forms of linguistic characteristics should be considered for the identification of disinformation.

6 Linguist features in different types of disinformation

One of our society's most serious concerns is online information disorder. Although fake news and conspiracy theories are not new, the exponential rise of social media has provided an accessible venue for their dissemination that is often quicker than true news. In online social media, a large deal of misinformation, such as rumors, propaganda, and conspiracy theories, is spread with the goal of deceiving people and forming certain attitudes. Depending on the degree of falsehood and the desire to hurt, information disorder can be characterized as misinformation, disinformation, or malinformation (Giachanou et al, 2021).

Two-thirds of Americans acquire their news from social media, according to Pew Research Center (Gottfried and Shearer, 2016). However, although social media has become an important source of information for many people, it has also become a source of disinformation, hoaxes, and false news for others. This is due to the fact that, unlike traditional news channels, social media platforms offer nothing in the way of personal responsibility or fact-checking (Addawood et al, 2019). Misinformation, such as conspiracy theories, hoaxes, and rumors, spreads just as easily on social media as accurate information. For example, when the Ebola crisis erupted in 2014, research found that on the Twitter social media network, falsehoods, half-truths, and rumors spread as rapidly as real facts (Jin et al., 2014).

6.1 Fabricated

Fabricated stories are those that are utterly devoid of any factual foundation, and are completely fake. The goal is to deceive and damage people (Wardle and Derekshan, 2017). One of the most serious forms (Zannettou et al., 2018) is when the fabrication mimics the style of a news story in a way that the receivers feel it is genuine (Tandoc et al., 2017). It might be written, but it could also be in a visual format (Ireton and Posetti, 2018).

People readily get reliant on social media as a route for exchanging information since it is like a blank piece of paper on which anything may be written (Yaraghi, 2019). This is precisely why material provided on social media sites (such as Twitter and Facebook) are closely reviewed (Haralabopoulos et al., 2015). Although these platforms have attempted to curb the dissemination of false news, they have mostly failed to do so.

Disinformation, on the other hand, is not only a technological issue. Uncertain socio-psychological elements can contribute to the spread of incorrect information. According to Chadwick et al. (2018), people who shared tabloid news articles were more likely to spread false or inflated information. The proportional number of generalized terms was largest in the fabricated story, according to Dilmon and Rakefet (2009).

William Yang Wang (2017) used a contemporary publicly available data collection named LIAR to address the low availability of labelled data sets for countering false news using statistical methodologies. Using language patterns, this data collection was used to analyze falsified news. The findings were based on a comparison of numerous methods, including Logistic Regression (LR), the Convolution Neural Network (CNN) model, Long Short-Term Memory (LSTM) networks, and Support Vector Machines (SVM). They came to the conclusion that combining meta-data with text increases the identification of false news greatly. The authors claim that this data set may be used to detect rumors, categorize positions, and conduct topic modeling, argument mining, and political Natural Language Processing (NLP) research.

Several exploratory studies were used to uncover the language distinctions between real and fake news in a separate technique for automatic fake news identification (P'erez-Rosas et al., 2017). It entailed the introduction of two new data sets, the first of which was compiled using both manual and crowdsource annotation, and the second of which was created entirely from the internet. Based on this, a series of exploratory studies were conducted to determine the most prevalent language features associated with fake news (Khan et al. 2021). Second, a model for detecting fake news was created using the retrieved language traits. They came to the conclusion that the suggested system exceeded humans in some circumstances involving more serious and diversified news sources. In the celebrity sector, however, humans outperformed the suggested approach.

On “Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017)” (2018), 276 digital items were gathered between November 2016 and June 2017 in the sphere of US politics in order to uncover discrepancies be-

tween manufactured (i.e., 'fake') and authentic news coming from the US. The writers used pattern.en (De Smedt, and Daelemans, 2012) and the NLTK packages (Loper& Bird, 2002) of the Python language libraries to content-analyze the matched datasets of 276 fake and authentic news headlines and texts using natural language processing (NLP) approaches. On pairings of authentic and faked news, a paired sample t-test was performed with a significance threshold of 0.05.

According to Asubiaro et al. (2018), real news stories include more words than those of manufactured ones. In the same way, real news has more paragraphs than fake news. Legitimate news, on the other hand, has shorter paragraphs with larger letters and fewer words per paragraph than falsified news, which has longer paragraphs with smaller fonts. Fabricated news headlines, on the other hand, include more words and punctuation marks than authentic news headlines samples. People may struggle to keep track of punctuation marks in the body of the article, but finding unneeded punctuation marks in the title is easy.

Differences in psycho-linguistic aspects also demonstrate that fabricated news articles have more positive and negative effect and their titles contain greater emotiveness, indicating that the bodies of such pieces and their headings seek to make substantial emotional pleas to the readers. Asubiaro et al. (2018) discovered that fabricated news articles used more casual language. Furthermore, fabricated news headlines feature more demonstratives (pronouns and unspecific) and, on the other hand, less verified facts (specific names), as seen frequently in clickbait. Fabricated news headlines, on the other hand, contain more of he, she, they, etc.' (i.e., pronouns), whereas real news headlines have more particular names. Fabricated political news stories by comparison to their likely legitimate counterparts, tend to have fewer words, fewer but lengthier paragraphs; they also contain more slang, swear, and affective words. There are more words, demonstratives, pronouns, and punctuation marks in fabricated news headlines, but less verified facts (or named entities).

6.2 Imposter

Imposters are genuine sources that are impersonated by fraudulent, made-up sources in order to promote a fictitious story. It's actually quite deceptive, because the source or

author is regarded as a major criterion for determining legitimacy (House of Commons, 2018; Zannettou et al., 2018; Wardle and Derekshan, 2017).

The Times Mexico (times.com.mx) and Before It's News (beforeitsnews.com) are two of the most notable non-credible instances. PolitiFact labeled the first (now offline) as an 'Imposter site,' since it poses as a legitimate media - a division of The Times. Fox News and TheBlaze, which have been considered as the most distrusted and least known of the trustworthy sources featured, are the most problematic instances of credible sources whose approach mirrors that of non-credible sources (Mitchell et al. 2014).

6.3 Conspiracy theories

They are stories without factual foundations are impossible as there is no set standard for reality. They frequently describe major events as government or powerful individual conspiracies (Zannettou et al., 2018). Conspiracies are difficult to prove as genuine or untrue by definition, and they are usually started by those who believe they are true (Allcott and Gentzkow, 2017). Evidence that contradicts the conspiracy is viewed as more confirmation of the conspiracy (EAVI, 2018). Some conspiracy theories may have negative ramifications.

This type of fake news offers answers for news items about entities that are in the spotlight, but most of the time, these explanations are based on pseudo-scientific findings (Shahsavari et al., 2020). Conspiracy theories foster a style of thinking that is diametrically opposed to the scientific method of explanation, encouraging those who are predisposed to them to share knowledge and speak out against disinformation (Potthast et al., 2017).

Researchers have recently been more interested in automated detection of information disorder. Emotions, user metadata, language traits, visual information, and temporal aspects have all been investigated for the purpose of detecting deception. Users, in addition to the content, play an important role in detecting erroneous content. Shu et al. (2018) studied users who spread fake news and discovered that several characteristics, such as registration time, differ between people who spread fake news and those who spread real news. They also demonstrated that integrating user profile variables with psycho-linguistic characteristics of tweets is a highly effective method for detecting false news. Vo and Lee (2019) studied the linguistic characteristics of fact-checking

tweets and discovered that fact-checkers prefer to employ formal language in their messages.

6.3.1 The ConspiDetector model

ConspiDetector is a model based on a convolutional neural network (CNN) that detects conspiracy propagators by combining word embeddings with psycho-linguistic characteristics derived from user tweets. In terms of F1-metric, the results reveal that ConspiDetector can enhance performance in detecting conspiracy propagators by 8.82 % when compared to the CNN baseline (74 %).

Can psycholinguistic characteristics, on the other hand, be utilized to distinguish conspiracy and anti-conspiracy propagators? Giachanou et al. (2021) present ConspiDetector, a CNN model for evaluating the usefulness of linguistic and psychological variables derived from users' tweets in distinguishing between conspiracy and anti-conspiracy propagators. In comparison to anti-conspiracy propagators, conspiracy propagators exhibit different profile characteristics, according to this study. Furthermore, it is demonstrated that the language patterns found in the tweets of conspiracy and anti-conspiracy propagators differ. More importantly, variances in their personality characteristics have been discovered. Finally, the findings of the experiments demonstrate that psycholinguistic factors are valuable for detecting conspiracy and anti-conspiracy propagators, while profile characteristics are not. The rest of the article focuses on similar work on misinformation detection, with a concentration on conspiracy theories. After that, they describe the steps they used to construct the dataset of posts pulled from conspiracy and anti-conspiracy propagators' timelines. The profile and psycho-linguistic characteristics of conspiracy and anti-conspiracy propagators are next examined. The ConspiDetector model, as well as the assessment procedure and performance, are then shown. Finally, the findings and limitations of this research are discussed, followed by conclusions and future research.

In recent years, the detection of internet deception has gotten a lot of attention. Researchers have worked on detecting fake news, rumors, clickbaits, bots, and fact checking, among other things.

ConspiDetector is a tool that uses a CNN and psycholinguistic variables that are derived from individuals' tweets to categorize them as conspiracy or anti-conspiracy propaga-

tors. The model has two branches: a content-based branch and a lexicon-based branch. An embedding layer is followed by convolutional, max pooling, and dense layers in the content-based model.

- Since the classification objective is binary (conspiracy propagators versus anti-conspiracy propagators), they employ a sigmoid layer as an output. Dropout is used after the embedding layer and before the dense layer in this approach. After the max pooling layer, convolutional filters of various sizes concatenate their outputs into a single vector. They concatenate all of the user's tweets into a single document to feed it to this branch. It's worth noting that they used CNN over LSTM since the process of concatenating all the tweets ignores the input document's sequential structure. This was further supported by the fact that when they used LSTM on our data, they had a poorer performance than CNN.
- The second branch is based on the four types of psycho-linguistic variables collected from the user's tweets (i.e., personality traits, emotions, sentiment, and linguistic patterns). They start by counting how many terms from the categories/lexicons occur in a user's tweet (count frequency feature vector). They perform this for all of the user's tweets, and then total the tweet vectors and divide by the number of tweets to get an average vector for that person. ConspiDetector's second branch receives the final averaged vector.
  For the llinguistic patterns, they used LIWC, a typical technique for mapping text to 73 psychologically significant categories, to compare the psychological characteristics of conspiracy and anti-conspiracy propagators. In particular, they extract pronouns (I, we, you, she or he, they), personal concerns (work, leisure, home, money, religion, death), time focus (past, present, future), informal language (swear, assent, non-fluencies, fillers), cognitive processes (causation, discrepancy, tentative, certainty) and affective processes (anxiety).

6.3.2 The use of words in conspiracy theories

In comparison to users who tend to accept conspiracy theories, users who oppose conspiracy theories use more of the third singular (i.e., she or he) and the first plural person (i.e., us). In comparison to individuals who support conspiracies, those who share content debunking conspiracy theories exhibit a greater use of personal issues. In comparison to conspiracy propagators, anti-conspiracy propagators use a lot of words about work (e.g., work, class, boss). "In the office today after being gone for a week, greeting me was a pile of cards, letters and flowers" for example, is an example of a work-related tweet. Furthermore, anti-conspiracy propagators employ terms like leisure (e.g., house, TV, music), money (audit, cash, owe), home (e.g. house, kitchen, lawn) and death.

Conspiracy theorists, on the other hand, appear to be more concerned about religion. A conspiracy theorist, for example, tweeted the tweet "Pope Francis opens the door for future female deacons."

Both conspiracy and anti-conspiracy propagators emphasize more on the present than the past or future when it comes to time focus. This may be explained by the sort of medium utilized to broadcast what is going on at any given time (i.e., tweeting). In addition, users who embrace conspiracies are less concerned with the present and future than those who oppose theories. Furthermore, as compared to anti-conspiracy propagators, conspiracy theorists use more curse words. In addition, when compared to anti-conspiracy propagators, conspiracy propagators employ more acquiescence terms (e.g., agree, yup, okey).

Finally, when it comes to cognitive processes, anti-conspiracy propagators utilize causality (because, effect, hense) more frequently than conspiracy propagators. The fact that people who deny conspiracy theories offer more explanations and arguments in their postings explains this. Anti-conspiracy propagators also use statistically significant more discrepancy (should, would, could) and tentative terms (e.g., maybe, perhaps).

In terms of psycho-linguistic characteristics, conspiracy theorists are found to employ a greater quantity of swear words. Anti-conspiracy propagators, on the other hand, appear to have more personal interests (such as employment), more cognitive processes, and a higher anxiety level. This study is extremely useful in identifying the language variations between conspiracy and anti-conspiracy propagators, as well as providing more insights into conspiracy theory research (Giachanou et al. 2021).

6.3.3 Conspiracy theories and privacy concerns

According to the findings of Visentin et al. (2021), terms connected with privacy concerns and conspiracy theories belong to two distinct domains. Overall, conspiracy theories increase the likelihood of someone retweeting a text, while privacy concerns, contrary to popular belief, reduce the virality of a tweet. Retweets are negatively affected by the complexity of the language used, implying that tweets are more successful when simple texts are utilized, and avoiding large cognitive demands for the reader. The data also paints a fascinating picture of how emotions and certainty affect language. In fact, Aleti et al. (2019) discovered that emotions in text might decrease retweets, whereas Pezzuti et al. (2021) discovered that certainty in language can also lower users' in-

volvement (i.e., retweets). This pattern of findings contributes to the scholarly argument, and it reflects the high level of ambiguity that defines the Twitter debate about the use of a contact tracing software to combat the COVID 19 pandemic.

These findings add to the literature in a variety of ways. For starters, they add to the expanding corpus of work that focuses on the interaction of psychology, marketing, and languages (e.g., Aleti et al., 2019; Berger et al., 2020; Netzer et al., 2019; Packard & Berger, 2019; Labrecque et al., 2020). This body of literature contends that a work reflects and reveals something about its creator, as well as having an affect on the audience (e.g., Berman et al., 2019; Cruz et al., 2017; Labrecque et al., 2020; Massara et al., 2020; Van Laer et al., 2019).

Rather than implying any type of social media censorship, it is discovered that privacy problems may be easily identified by monitoring a smaller selection of terms. As a result, we should assume that Twitter users who employ complicated perigrams to express privacy concerns and/or conspiracy theories in a tweet will restrict the virality of their tweets.

Visentin et al. (2021) with the automated text analysis that proceeded by including two dictionaries PrivacyDIC and ConspiracyDIC were included in LIWC (Linguistic Inquiry and Word Count; Tausczik & Pennebaker, 2010) and ran automatic text analysis on the entire collection of 166 Clubhouse tweets and 1311 vaccinations tweets. The LIWC categories for text complexity, word count, presence of prepositions, words with more than six letters and cognitive mechanisms were also included in the study. They incorporated terms conveying tentativeness and certainty to account for certain language. They incorporated negative and positive emotions to account for emotions (negative emotions and positive emotions, respectively).

6.3.4 COVID-19 and Conspiracy

For COVID-19 misinformation, the range of sentiment was substantially higher, with tweets more frequently displaying rising unfavorable sentiment, notably in April and May 2020. The themes of 5G, Bill Gates or the Bill & Melinda Gates Foundation, SARS-CoV-2 being laboratory-released or human-made, and vaccinations are discussed in these conspiracy theories tweets. To define the language properties of COVID-19 conspiracy theories as they change over time, downstream sentiment analysis (AFINN, NRC) and dynamic topic modeling were utilized. Each conspiracy theory is divided into

eight emotions, as well as a general negative or positive sentiment. Although the results for misinformation and non-misinformation in tweets linked to 5G conspiracies are similar, there are significant differences in the other four conspiracy theories. When compared to tweets that aren't labeled as disinformation, misinformation tweets have greater levels of negative sentiment, fear, rage, and contempt.

To classify the tokenized tweets, two well-documented sentiment dictionaries were employed. The first, AFINN, assigned an integer score to each word in the dictionary ranging from –5 (negative emotion) to +5 (positive emotion). The National Research Council (NRC) Word-Emotion Association Lexicon was used to associate words with emotion categories, offering labels for eight emotions: anger, anticipation, contempt, fear, joy, sorrow, surprise, and trust, as well as a general "positive" or "negative" feeling. The emotion for each identified data set was then contrasted over time by Gerts et al. (2021). The total of integer scores and the numbers for each emotion label were computed for each tweet as aggregate sentiment metrics.

These findings are in line with previous research, which found that those who believe in one conspiracy theory are more likely to believe in others or are more open to conspiratorial thinking in general.

The first step in establishing measures to address misinformation that is harmful to public health is to identify it. Instead of reacting to existing incorrect ideas, the capacity to analyze conspiracy theories before they spread would allow public health practitioners to develop effective communications to prevent misperceptions. Too frequently, health professionals fail to create successful communications campaigns because they focus on what they want to promote rather than correcting the recipients' preconceptions. Misinformation spreads quickly and without a clear path. This work shows that using Twitter data, it is feasible to discover and characterize prevalent and long-lived COVID-related conspiracy theories, even when the messages' content and tone change over time.

6.4 Hoaxes

Hoaxes are fabrications that are quite intricate and large-scale, and may involve deceptions that go beyond the boundaries of a joke or a ruse, resulting in material loss or injury to the victim (Rubin et al., 2015). They comprise either fake or inaccurate information that are presented as actual facts. This category is also known as half-truth or

factoid stories (Zannettou et al., 2018) and is capable of persuading readers of the veracity of a paranoia-fueled story (Rashkin et al., 2017).

A hoax is a sort of deception that is intended to fool the reader (Volkova et al., 2018). The following is an example of a fake tweet: “BREAKING! Massive Volcano Eruption Only 32 Miles Away From MAJOR Nuclear Plant!” enlightened on purpose.

A hoax is a lie that is purposefully created to appear to be true. It's distinct from mistakes in observation or judgment. It is frequently used as a practical joke, to embarrass someone, or to drive social or political change by raising people's awareness of a topic. It might also be the result of a marketing or advertising strategy.

A hoax is an act or statement that deceives, conceals the truth, or promotes a false belief, notion, or idea. It is frequently done for the sake of personal benefit or advantage.

It encompasses a variety of statements or omissions aimed at distorting or omitting the entire truth. Wrong statements and misleading assertions, in which crucial information is removed and the recipient is led to draw false conclusions, are examples of hoaxes. A person can sometimes disguise themselves by their appearance to give the idea of being someone or something different; for a well-known individual, this is referred to as incognito. Passing is more than just dressing up; it also entails concealing one's true speaking pattern (Huseynova, 2021).

6.4.1 Linguistic characteristics of hoaxes

There is a claim that when people disguise their writing style, some language traits change, and that by recognizing such features, misleading texts may be identified. According to the Undeutsch Hypothesis,"Statements that are the product of experience will contain characteristics that are generally absent from statements that are the product of imagination”. Deception necessitates extra cognitive effort in order to conceal information, which frequently results in minor alterations in human behavior. These changes in behavior have an impact on both verbal and written communication. Several language indicators were discovered to distinguish between deceitful and genuine speech. Deceivers, for example, use shorter phrases, have fewer average syllables per word, and employ simpler sentences than truth tellers. As a result, misleading language appears to be simpler and easier to understand. Although stylistic deception is not the same as lying, similar language elements alter in this type of deception.

According to Afroz et al. (2012), stylistic deception can be distinguished from conventional writing with 96.6 % accuracy (F-measure) and distinct forms of deception (imitation vs. obfuscation) may be identified with 87 % accuracy (F-measure), using linguistic and contextual cues (big feature set).

Long-term deception detection is akin to identifying fiction as deceit. Fiction and sophisticated deception have distinct linguistic characteristics than short-term on-the-spot deception, since the author has enough time and topic to write descriptively and edit sufficiently to make it look as a true document in the long-term deception. This is why detecting long-term hoaxes and deceit requires a different method. Regular authorship recognition can aid in the detection of discrepancies in writing and the identification of the true authors of deceitful materials (Afroz et al., 2012).

Vukovic et al. concentrate on hoaxes and propose the usage of an email detection system. The suggested system is made up of a feed-forward neural network and a self-organizing map (SOM), and it has been trained on a corpus of 298 hoaxes and 1370 legitimate emails. The method has a 73 % accuracy rate with a 4.9 % false positive rate. Afroz et al. concentrate on spotting hoaxes through changes in writing style. The assumption is that when people try to obscure or modify information from users, they employ various language traits. Their tests on diverse datasets show that the suggested method can detect hoaxes with a 96 % accuracy (Zannettou et al. 2019).

6.5 Biased or one-sided

These are stories that are biased heavily in favor of a person, party, circumstance, or event, causing conflict and polarization. The context of this style of news is significantly biased (i.e., left or right wing), provocative, emotive, and frequently filled with lies. They include either a combination of accurate and incorrect information or largely false information, resulting in misleading information intended to validate a specific ideological viewpoint (Zannettou et al., 2018; Potthast et al., 2018).

6.5.1 Examples of Biased News

On Damstra et al.'s (2021) paper it is implied that information that is purposely false is ideologically biased in favor of the right. The 2016 presidential election in the United States has sparked academic study on misinformation, with a lot of prominent work focusing on the US context in the years leading up to and after the election. Faris et al.

(2017), for example, investigated at election coverage in the mainstream and on social media. Between May 1, 2015, and Election Day, almost two million pieces were published by about 70,000 online media sites (November 8, 2016). While partisan prejudice exists on both sides of the political divide (Bradshaw et al. 2020), information on the right receives more amplification and legitimacy, especially on social media (Faris et al. 2017). In fact, purposely false material in news-like texts was discovered to have a pro-Trump signature more often than articles favoring Clinton (Silverman 2016b), and stories promoting Trump were disseminated more extensively than stories favoring Clinton (Allcott and Gentzkow 2017; Lazer et al. 2017). Benkler, Faris, and Roberts (2018) gathered and analyzed two million stories published during the 2016 presidential election campaign and 1.9 million pieces regarding Trump's first year in office. The dissemination and reach of misinformation were investigated in both studies, by examining cross-linking patterns between media sources, including Twitter and Facebook sharing activity. The sheltered right-wing media environment has been proven to be significantly more vulnerable to deception than publications on the opposite side of the spectrum, ranging from center-right to far-left. Marwick and Lewis (2017) look at how far-right organizations used the media in the run-up to the 2016 election in the United States. They come to the conclusion that most Clinton supporters acquired their news from normal sources, but many Trump followers were surrounded by a far-right network that "peddled heavily in misinformation, rumors, conspiracy theories, and attacks on the mainstream media" (Marwick and Lewis 2017, 21).

Outside of the United States, some research has been done on the partisan component of misinformation. For example, Pierri et al. (2020) investigate Italian misinformation on Twitter and discover that the majority of themes are contentious discussions about immigration, crime, and national security—issues that resemble the conservative and far-right political agenda. Similarly, right-wing connotations are found in German misinformation, including "skepticism toward the European Union (...) and, most of all, the exclusion of migrants and refugees" (Zimmermann and Kohring 2020, 221).

Intentionally false news articles employ emotions in a very different way: emotions are more evident, prominent, and unpleasant. There is a number of research that back up this assertion. Emotionally driven language is a significant aspect of this sort of false information; propaganda methods are employed to influence readers on an emotional rather than logical level. Neudert, Kollanyi, and Howard (2017) take a similar method

in their study of the transmission and reach of purposefully false information during campaign seasons in Germany and afterwards in Germany, France, and the United Kingdom (Neudert, Howard, and Kollanyi 2019). They also discovered that emotionally charged phrases are crucial in these communications. Scholars who examined purposely misleading material on Breitbart's Facebook timeline came to the same conclusion: the content is affective, with the goal to arouse voter wrath, fear, and contempt (Benkler, Faris, and Roberts 2018; Faris et al. 2017). It's vital to remember that emotional language mostly refers to negative feelings, as fake news or tweets are designed to make people feel bad (see also Hameleers, van der Meer, and Vliegenthart 2021). Horne and Adali (2017) explore the linguistic aspects of real news, fake news, and satire using three different datasets. They come to the conclusion that deliberately false news contains more negative emotional phrases than real news (see also Zhou and Zafarani 2020).

Fake headlines also have a high level of emotionality (Asubiaro and Rubin 2018, see for similar findings Volkova et al. 2017). Van Der Zee et al. (2018) perform a linguistic study on 447 tweets posted by former US President Trump that were fact-checked by The Washington Post in connection to social media. According to the study, honest tweets include more happy feelings, and dishonest tweets have more negative emotions. According to research on Russian misinformation tactics on social media, these statements are intended to elicit indignation and resentment among opposed outgroups. Defaming political and social opponents is a popular tactic for achieving this result (Freelon and Lokot 2020; Howard et al. 2019).

6.6 Rumors

Rumors are stories whose veracity is disputed or never proven (gossip, innuendo, unverified claims). This type of fake material is regularly disseminated on social media sites (Peterson and Gist, 1951).

A rumor is defined as "a piece of circulating information whose veracity status is yet to be verified at the time of spreading" (Zubiaga et al., 2018). Rumors thrive in today's world of social media, making detection more difficult. The transmission style differs significantly from real news, according to studies, and is used to categorize rumors on the internet (Liu & Xu, 2016).

A rumor is an unconfirmed allegation that originates from one or more sources and travels from node to node in a network over time. A rumor on Twitter is a group of tweets all expressing the same unconfirmed allegation (although, the tweets might be, and probably likely are, written differently from one another), spreading across Twitter in a cascade. A rumor can take one of three paths: it can be resolved as true (factual), false (non-factual), or unresolved. Usually, there are multiple rumors regarding the same subject, any of which might be real or untrue. When one or more rumors are resolved, all other rumors concerning the same issue are resolved as well (Vosoughi et al., 2017)

6.6.1 Rumors and Social Media

Previous study has shown that hashtags might be effective rumor signals (Castillo, Mendoza, and Poblete, 2011). The current emphasis of research is on developing an automated rumor detection program. A rumor detection approach (Zhao et al., 2015) was developed for this purpose. Two types of clusters were created using this technique: one for posts including terms like "Really," "What," and "Is it true?" and another for posts containing words like "Really," "What," and "Is it true?" The results of these questions were then utilized to find rumor clusters. Similarly, postings that did not contain words of inquiry were placed in a separate cluster. Both clusters yielded statements that were similar. The clusters were then rated according to how likely they were to include these terms. Later, the whole cluster was searched for any assertions that were challenged. These tests, which used Twitter data, resulted in earlier and more successful rumor identification (almost 50 rumor clusters were identified). However, there is certainly a lot of room for improvement here (Zhao et al., 2015). For example, developing a classifier might enhance the manual collection of inquiry terms, and investigating new characteristics for the rumor cluster method could improve the ranking process (Khan et al. 2021).

Kwon et al. (2013) investigate how rumors spread on Twitter, taking into account results from social and psychological studies. They discovered that users who spread rumors and non-rumors have similar registration age and number of followers, that rumors have a distinct writing style, that sentiment in news depends on the topic rather than the credibility of the post, and that words related to social relationships are more frequently used in rumors, by analyzing tweets obtained from.

Kwon et al.'s (2013) research uses Twitter data as a starting point. The dataset includes 54 million user profiles, 1.9 billion follow links between them, and 1.7 billion public tweets from March 2006, when Twitter was debuted, until August 2009. They were able to investigate user behaviors around genuine information dispersion thanks to the whole set of users, links, and tweets. They had to first identify true rumor incidents from Twitter data in order to study the dissemination characteristics of rumors. They used three websites: snopes.com, urbanlegends.about.com, and networkworld.com to find lists of popular events. They employed LIWC, a commonly used sentiment analysis tool, to look at the linguistic characteristics of rumor spreading (Linguistic Inquiry and Word Count).

According to the findings, rumors feature more terms associated with skepticism and doubt, such as negation and supposition, and are less effective as discussion subjects. These findings show that the process of doubt affects users' perceptions of rumors and resulting in a variety of writing styles. People expressing hesitation in their tweets can be ascribed to the existence of a lot of negation in rumors. Being a statistical test, Kwon et al. (2013) confirm that rumors have a distinct writing style, providing empirical support for social and psychological ideas concerning rumor propagation.

Rumors have a larger percentage of terms relating to social and hearing. That is, terms relating to social relationships, such as 'friend,' 'buddy,' and 'neighborhood,' appear more frequently in rumor tweets. This would indicate that rumors are more likely to spread through social relationships as the major dissemination mechanism.

Zubiaga et al. (2018) use journalists to annotate rumors in real time and analyze 4K tweets connected to rumors. Their findings show that legitimate rumors are resolved faster than false rumors, and that users have a general predisposition to believe any unverified rumor. The latter, on the other hand, is less common among credible user accounts (e.g., reputable news outlets), which often offer information accompanied by proof. Thomson et al. investigate Twitter's activity in the aftermath of Japan's Fukushima Daiichi nuclear power plant tragedy. The writers sort the communications into categories based on the user, location, language, kind, and source's reliability. They observe that anonymous users and individuals who reside far away from the tragedy share more information from less trustworthy sources. Finally, Dang et al. investigate how Reddit users engage with rumors by looking at a prominent fake rumor (i.e., Obama is a Muslim). They categorize people into three groups: those who promote fake rumors, those

who reject false rumors, and those who make jokes about false rumors. They used a Naive Bayes classifier with an accuracy of 80% to identify these people, and discovered that more than half of the users laughed about the rumor, 25% denied the joke, and just 5% backed the claim.

Starbird et al. (2010) investigate and detect distinct sorts of expressed doubt in OSN postings, during the lifecycle of a rumor, The article collects 15 million tweets relating to two crisis occurrences in order to examine the degree of ambiguity in communications (Boston Bombings and Sydney Siege). They discovered that rumor-related tweets employ distinct language patterns. Their findings can be employed in future detection systems to efficiently and quickly detect rumors. Zubiaga et al. (2018) present a novel method for gathering and preparing datasets for the identification of misleading information. They suggest getting OSN data that will then be annotated by humans, rather than identifying rumors from breaking websites and then retrieving data from Open Storage Networks (OSN). They gather tweets from the 2014 Ferguson unrest incident as part of their analysis. They employ journalists as annotators, with the goal of labeling tweets and discussions. The journalists specifically annotated 1.1k tweets, which may be divided into 42 separate articles. According to their research, 24.6 % of tweets are rumors. Finally, Spiro et al. conduct a quantitative study of tweets on the Deepwater Horizon oil leak in 2010. They claim that as a result of the media coverage, the number of tweets about the catastrophe has grown. They also discovered that when retweets contain event-related phrases, they are more likely to be sent in a sequential order (Zannettou et al. 2019).

With the volume and pace of user-generated material on social media, detecting rumors is critical. Information may be spread through social media independently of the source's verification status or truth value. Rumors are fueled by the forwarding and sharing of material, as well as the absence of validation, because it allows for unrivaled interchange and broadcasting. However, when consumers are exposed to hazardous or unpleasant information, this may be detrimental. Furthermore, most social media platforms allow users to establish groups based on the same interests. Yet, such virtual alignments may result in the formation of echo chambers, in which members' own opinions are amplified and reinforced. Unconfirmed posts look more trustworthy in such echo chambers. When a member of a group hears a piece of information, they may believe it is accurate since it comes from their "own" people.

Various research has reported on automatic rumor identification in social media data, particularly on Twitter and Sina Weibo. Zubiaga et al. (2016) conducted a thorough survey in 2016. Zubiaga et al. (2016) conducted a comprehensive survey in 2018. The authors reviewed the current literature on the various sub-tasks of rumor resolution. To detect rumorous content in microblogs, several machine learning and deep learning models have been applied. Kumar and Sangwan (2019) investigated rumor detection using a range of machine learning techniques in 2018. The learning models for rumor identification and prediction were trained using a variety of features, including text-based, user-based, and network-based features. Deep learning algorithms have recently been utilized to identify rumors in the textual modality. Superlative results have been reported for RNN, attention-based RNN, CNN with RNN, and LSTM with RNN. Jin et al. (2017) have presented multimodal rumor detection using LSTM and RNN with attention. Zubiaga et al. (2016) proposed the sequential classifer model, CRF. This study proposes creating a hybrid learning model that aims to combine deep neural models with machine learning approaches that are optimal.

In times of crisis, rumors abound. Rumors proliferate in the virtual social environment due to the situation's unpredictability and significance, as well as a lack of knowledge. The usage of country-specific information published in the original language adds to the linguistic challenges of identifying rumors (Lotfi et al., 2021)

4.6.4 Lotfi et al. (2021)’s technique – Recognizing rumors

An innovative technique for recognizing rumor-based talks of diverse global events such as real-world emergencies and breaking news on Twitter is examined in Lotfi et al.’s (2021) study. Three areas of information transmission are investigated in this study: the language style used to communicate rumors, the characteristics of the people participating in the distribution of information, and structural traits. The characteristics of a reply tree and a user graph are structural features. In order to improve the efficiency of rumor discussion identification, structural characteristics were extracted as additional features. The additional characteristics are successful in identifying rumors, and the suggested approach is superior to existing methods, as the F1-score improved by 4% in the experiments. The suggested strategy was tested using Twitter data gathered during five breaking news stories. The conversation is the system's input, and the characteristics retrieved from it include the discussion's linguistic content, the identities of the persons engaging in the chat, and the structural elements. These characteristics are then in-

putted to a model that has been trained on manually annotated chats to determine whether or not a communication is a rumor.

Several of the chosen characteristics show considerable differences between the two groupings. The linguistic characteristics revealed how individuals reacted to rumors. In rumor chats, users use more question marks in reply tweets. In rumor dialogues, the source tweets are lengthier and contain fewer positive feelings. Furthermore, when exposed to rumor-related information, users are more likely to do an insight action (e.g., think, know), employ sad terms (e.g., crying, grief, sad) in a phrase, and take a hearing action (e.g., listen, hearing). Furthermore, as compared to non-rumor-based talks, people in rumor conversations have a lot of followers and their homepage has a URL.

6.7 Clickbait

Clickbait are sources that deliver generally trustworthy or dubious factual material but purposefully employ exaggerated, deceptive, and unverified headlines and thumbnails (Rehm, 2018; Szpakowski, 2018) to get visitors to explore the intended Web page (Ghanem et al., 2019). The objective is to boost traffic for the sake of profit, fame, or sensationalization (Pujahari and Sisodia, 2019; Zannettou et al., 2018). The substance seldom fulfills the reader's attention once they arrive (EAVI, 2018). Clickbait is a marketing technique for attracting consumers' attention. Clickbait, such as sensational headlines or breaking news, is frequently used to direct users to advertisements. More advertisement clicks equals more money (Chen et al. 2015a).

Clickbaits are phrases that are intended to draw a user's attention to a web page whose content falls well short of their expectations when they click on the link (Spicer, 2018). Many people find clickbaits irritating, and as a result, they only spend a brief amount of time on such sites. Higher clicks, on the other hand, translate into more cash for content providers, as the commercial component of employing online adverts is strongly dependent on web traffic (Conroy et al., 2015).

Clickbait is a type fake material that uses linguistic titles to entice viewers but fails to deliver on its promises. Chen et al. (2015) investigated possible strategies for programmatic clickbait discovery by integrating syntax and semantics with pictures and newsreader behavior under non-textual signals. They investigated the significance of these indicators in detecting false news, but were unable to put their research into practice. Bourgonje et al. (2017) suggested a clickbait detection technique that evaluates headline

relevance for article bodies. They used the data set supplied by the organizers of the inaugural fake news challenge on stance detection and applied a logistic regression classifier to reach a considerable accuracy of 89.59 %. Rashkin et al. (2017) examined the language of actual news to satire, hoaxes, and propaganda, and determined the characteristics of fake text and other internet sources for fact-checking (Verma et al., 2021).

To be considered clickbait, an article must contain the following characteristics: i) short text, ii) a media attachment, such as image or video and iii) the link to the publisher's article (Kiesel et al., 2019). To attract more readers, the majority of social media publishers employ click bait pieces to some level. Journalistic ethics, on the other hand, are opposed to these practices since they utilize unethical methods to mislead readers (Potthast et al., 2017).

For a variety of reasons, people may spread false information on social media. One of them is to promote reading, which can be done easily with clickbait. Clickbait is a deceptive advertising with a hyperlink attached. Its purpose is to entice them to click on the link and read what's within (Anna Escher, 2016). These ads entice consumers with appealing titles but offer nothing in the way of useful information. Clickbait attracts a significant number of people. (Monther et al.)

In the form of a tool that filters and detects sites carrying fake news, Aldwairi and Alwahedi (2018) developed a method to safeguard consumers against clickbait. They took into account a number of variables when classifying a website as a source of fake news. The program navigates a web page's content, examines the syntactical structure of the links, and looks for terms that might be deceptive. Before viewing the web page, the user is alerted. In addition, the program scans the links for terms linked with the title and compares them to a set of criteria. It also looks for punctuation marks on the page, such as question and exclamation marks, to see whether it is clickbait. Additionally, they also investigated the bounce rate, which is the proportion of users that leave a website after seeing a certain page. When the bounce rate was high, the content was flagged as a potentially deceptive source of information (Khan et al. 2021).

The initial stage was to find a reliable clickbait database, after which the properties were computed and the data files for WEKA were created. As a result, they crawled the internet for clickbait URLs. They concentrated on social media websites like Facebook, Forex, and Reddit, which are more likely to include false news or clickbait advertising or content. After collecting URLs in a file, a python script calculated the characteristics

from the title and content of the web sites in the second stage. Finally, the characteristics were retrieved from the site pages. Keywords, titles that begin with a number, all capitals words, question and exclamation marks, whether the user left the page quickly, and material linked to the title are among the characteristics. To validate the answer, they had to employ WEKA machine learning. They used the script below to extract the parameters needed to funiculate WEKA because WEKA requires specifically formatted input. Using a logistic classifier, the experimental findings reveal 99.4 % accuracy.

The syntactical structure of the links that direct viewers to these sites serve as a starting point. For example, when a user types in a set of search phrases with the goal of locating web pages that contain information connected to those terms, the tool will launch and go through the sites that the search engine has obtained before delivering them to the user. The plug-in will detect sites with links that include language that may mislead the reader, such as those with a lot of exaggeration and slang language. Such web pages will be labeled as possible fake news sources, and the user will be alerted before choosing to click on one of them. The user will be more capable to grasp the decision if the linkages and their syntactical structure are shown (Aldwairi & Alsaadi, 2017).

In addition, the program will employ the quantity of words linked with the phrasing used in the headings of the websites to determine whether ones contain misleading information. A baseline of, say, eight words will be used to classify a web page as having valid information, with those whose links contain more than the threshold number of words being identified as probable sources of fake news. The logic behind this strategy is based on the fact that clickbaits, on average, contain far longer words than non-clickbaits (Lewis, 2011). As a result, it's likely that the tool will utilize the phrase as a criterion to determine if a headline is likely to be clickbait.

The technology will track how punctuation marks are employed in web sites, in addition to the syntactic characteristics of headlines related with apparent clickbaits. The model will highlight websites with a lot of exclamation points and question marks in their headlines. Links to such websites will be flagged as possible clickbaits. For example, a trustworthy website may feature a headline like “Donald Trump Wins the US Presidential Race!”. A clickbait, on the other hand, might be constructed like this: “Guess what???? Donald Trump is the Next US President!!!!!!!!!”. In this example, the program would classify the first as non-clickbait and the second as a possible link to deceptive material (Aldwairi and Alwahedi, 2018).

6.7.1 Technology that detects clickbait

Several research employ machine learning approaches to detect clickbait on the internet. Chen et al. (2015) recommends employing SVMs and Naive Bayes to solve the problem. In addition, Chakraborty et al. (2019) proposes using SVM and a browser add-on to provide users with a system for news articles. For identifying clickbait tweets, Potthast et al. (2017) suggest using Random Forest. Zannettou et al. (2019), on the other hand, employ deep learning approaches to detect clickbait on YouTube. They present a semi-supervised model based on variational auto encoders in particular (deep learning). According to their findings, they can recognize clickbaits with reasonable accuracy, and YouTube's recommendation engine does not include clickbait videos in its recommendations.

The FakeFlow model separates it into N segments (with an accuracy of 85% in detecting bogus news). The system then uses both word embeddings and other affective elements such as emotions, hyperbolic words, and so on to capture the document's emotional flow. To determine whether a document is fake or real, the model learns to pay attention to the flow of emotive information across it. Ghanem et al. (2021) employ a list of 350 hyperbolic terms (Chakraborty et al., 2016), which are words having a strong positive or negative connotation (e.g., terrifying, breathtakingly, soul-stirring, etc.). These eye-catching words were collected from clickbait news headlines by the writers.

The emotional characteristics in the news text present a clear illustration of how false news stories attempt to manipulate the viewer. It appears that the presence of fear, grief, and surprise emotions at the start of the story drew attention to this section. On the other hand, towards the conclusion of the piece, it is clear that such unpleasant feelings do not exist, whereas emotions like delight and anticipation do (Ghanem et al. 2021).

6.8 Misleading texts

Misleading information is used to frame an issue or a person. It is when the material is not supported by the headlines, images, or captions. Separate pieces of source data may be accurate, but they are presented incorrectly (context/content).

The study of Dilmon and Rakefet (2009) included 48 native Hebrew speakers aged 20 to 4), all of whom had completed at least a high school education. The experimental population was diverse, with around 48 % males and 52 % females. Only 21% of the subjects had completed high school, whereas 79% had completed some form of higher

education. Sixty percent of people said they are religious, while forty percent said they are not. The individuals all agreed to take part in the study voluntarily. A list of 43 verbal criteria for inspection was created from the information obtained from them in the following areas: morphology (the tense system, pronouns, and persons), syntax (word order, conjunctions, etc.), semantics (registers, obscurity, generality, etc.), discourse (distraction techniques, misrepresentation, persuasion, etc.), and speech prosody (pauses, repetitions, exclamations, and so on). They were asked to tell two stories from their past, one of them true (‘‘the story from the past’’) and the other totally false (‘‘the fabricated story’’). MANOVA analyses with repeated measurements were used to test the writers’ prediction that variations in numerous linguistic criteria would be identified between the discourses of truth and deceit (for each linguistic criterion, the differences between the four stories were examined).

The removal of any term that might make it possible to confront the misleading text with reality or show that what is told is incompatible with the facts, the withholding of "loaded" details from the discourse (details that might reveal the invention), and the use of words that add no new information to the story, are all examples of *concealment*. The use of words with several meanings and that do not offer a distinct account of the event is referred to as *vagueness*. Narratives with fabricated scripts depict something that did not exist as if it did, or something that took on a different shape than it did in reality. The creator of such a text must ensure that no one can identify any indications that the text is inaccurate. By employing generic language or expressing things that cannot be verified, the speaker hides the specifics that could be scrutinized and obscures the environment that the tale reflects (Chapman, in press; Johnson and Raye, 1981; Dulaney, 1982).

- The criterion *"number of specific terms"* and *"number of generalized words"* showed a significant difference [$F(3,14) = 32.70$, $p < 0.001$] between the four tales. When the cause of the disparities was investigated, it was discovered that the created account used the fewest specific terms, yet there was no significant difference between it and the original story. The usage of certain terms was more prevalent in the day tale than in the story of the past and the invented story, with the life story making the most extensive use of these phrases. The fabricated story had the highest relative amount of generalized terms, according Dilmon and Rakefet (2009) analysis. Although there was no significant difference between the two stories, the narrative from the past had a lower number. The story of a day had the smallest amount of generic terms, while the life story had the fewest.

Narratives using invented scripts include fewer past tense verbs and more present and future tense verbs. A person attempting to deceive can give a narrative that appears to be totally genuine on the surface, but in linguistic terms, they use verbs in the present or future tenses, implying that the story has not yet occurred or has not occurred at all. Another motive for the speaker's choice of present or future verbs is to avoid an investigation to validate what has been said. When they talk about something that is happening right now or hasn't happened yet, the situation can't be investigated. When challenged with their fabrication, they might claim that they never mentioned that these things actually happened.

- A significant difference [$F(3,14) = 19.76$, $p < 0.001$] was identified between the four stories for the criterion *"number of verbs in the past tense"***.** The rate of use of the past tense in the life story was similar to that in the day story, and this rate was higher than the rate of use in the past story and the invented story, both of whose values are identical. The outcome is the opposite in the present and future tenses.
- The criteria *"number of verbs in the present"* revealed a significant difference between the tales [$F(3,14) = 12.74$, $p < 0.01$], as did the criterion *"number of verbs in the future"*: [$F(3,14) = 20.58$, p 0.001]. When the source of the disparities was investigated, it was discovered that the rate of use of verbs in the future in the past tale was the same as in the manufactured story, and was higher than the rate of use in the life story and the story of a day, both of which produced comparable findings.

Negative words are negation words (no, there is not, can't), as well as words having a negative semantic burden, such as failure, hit, attack, evil, and so on. The person's unfavorable attitude toward what he is saying is evidenced by the great amount of negative terms (Sovran, 2000:79–81). This conclusion demonstrates the importance of upbringing and social customs, which forbids deception by the speaker and in their language (Fraser, 1994). It is instilled in a person from a young age that deception is harmful and should be avoided. This education does not ensure that a person will always be completely honest, but it does have an impact on their attitude toward deception. The individual's mindset is reflected in the words he or she chooses to employ. Nothing about the selection of these particular phrases affects the story in any way in terms of substance, yet there is a semantic difference between the words used while inventing and those used in other situations. This discrepancy might imply something the speaker doesn't want to admit: their dissatisfaction with what they are saying (Dilmon, Rakefet. 2009).

- The criteria *"number of words relating to the other"*, comparing the four stories, showed a significant difference [$F(3,14) = 37.40$, $p < 0.001$]. The criteria of *"amount of negative words"* was also shown to be significantly different

amongst the four stories [$F(3,14) = 47.88$, $p < 0.001$]. The number of words referring to the other and the number of negative words in the day narrative were lower than their numbers in the past story and the invented story, and their lowest value was discovered in the life story, according to the cause of the discrepancies.

6.9 Fake Reviews

A Fake review is considered any (positive, neutral, or negative) review that is not based on an actual consumer's honest and impartial opinion, or that does not accurately reflect a consumer's true experience with a product, service, or business (Valant, 2015). Fake reviews jeopardize the former's reliability and have an impact on the customer purchasing process.

Over the last two decades, online reviews of products and services have grown increasingly crucial. Customer purchase decisions are influenced by review score and number of reviews (Maslowska et al., 2017). According to Kumar et al. (2018), up to 90% of buyers read reviews before making a purchase, and a product's conversion rate might rise by up to 270 % as it collects reviews. Reviews may boost conversion rates by 380 % for high-ticket items (Askalidis and Malthouse, 2016).

Slight language differences can have a significant influence on readers' understanding and appraisal of textual material, according to communication and psycholinguistic research (Toma& Hancock, 2012; Van Laer, Edson Escalas, Ludwig, & Van Den Hende, 2019). Liars employ less self-references (e.g., I, me, my), fewer third-person references (e.g., he, she), and more negative emotive phrases (e.g., walk, go) to portray a convincing tale, according to Newman et al. (2003). They created a text-based computer model to discern between authentic and fake abortion accounts. Fuller, Biros, and Wilson (2009) employed a variety of linguistic indicators as inputs for the classification model (such as lexical density, first-person pronouns, pleasantness, and verb quantity) and reached a classification accuracy of roughly 74%. Quantity, specificity, affect and uncertainty were used to categorize language signals by Fuller et al. (2009). To identify fraud in online dating profiles, Toma and Hancock (2012) used a linguistics method. They discovered that liars were methodical in achieving their communication objectives. Perceivers believe that persons who write specific and crisp self-descriptions that are lengthier in length are more trustworthy. The research' findings show that verbal indicators to deception are generally evaluated haphazardly, are context-specific, and are

not theoretically supported (Matsumoto &Hwang, 2014). As a result, there is a need to understand the theoretical characteristics of language features that cause customers to assess a misleading review (Ansari et al., 2021).

Both consumers and businesses benefit from online customer reviews. Companies are aware that online reviews and suggestions influence product sales (Cui, Lui, &Guo, 2012; Goh, Heng, & Lin, 2013; Riquelme, Román, &Iacobucci, 2016; Saboo, Kumar, &Ramani, 2016) and company stock prices across product categories(Tirunillai&Tellis, 2012). The reliability of internet reviews contributes to their influence: these evaluations reflect genuine customers sharing their thoughts on their purchase and consuming experiences.Furthermore, unlike traditional advertising, where customers are aware that they are being persuaded (Friestad & Wright, 1994), internet reviews are submitted by users who have no expectation of reward; hence, their ratings and reviews are considered genuine and unbiased. However, fraudulent reviews exist, such as when firms pay certain customers to leave favorable ratings for their company and negative reviews for competitors.

6.9.1 Online hotel reviews

Moon et al. (2021) have found a few prominent markers of fraudulent messaging in online customer evaluations in their research thus far. Fake reviews are more extreme (i.e., more positive or negative) than real reviews.

False negative reviewers overproduced negative emotion phrases like "terrible" or "disappointed" in comparison to true evaluations, according to Ott et al. (2013), just as fake positive reviewers overproduced positive emotion terms like "elegant" or "luxurious", findings on which they concluded after investigating the reviews of 20 hotels. The findings show that n-gram–based Support Vector Machine (SVM) classifiers outperform untrained human judges in detecting negative misleading opinion spam in a balanced sample. The technique is that they train their model on all reviews for 16 hotels and test it on all reviews for the remaining 4 hotels for each cross-validation iteration (Ott et al., 2013).

According to Kronrod, Lee and Gordeliy (2017), authentic reviews employed much more past tense verbs, distinctive terms (e.g., the words "keyboard" or "typing" for a computer), and concrete, specific nouns than fictitious evaluations. These researchers

put their theories to the test by employing automatic text analysis methods on real and false reviews provided by volunteers for the aim of their research Kronrod et al., 2017).

Finally, honest reviews included more "spatial" features, such as "bathroom" and "location" for hotels, as opposed to fraudulent reviews, which tend to describe abstract notions like why or with whom the client visited the hotel (Li et al., 2013; Ott et al., 2011; Ott et al., 2013).

The survey-based text-categorization technique of Moon et al. (2021) enables for a thorough capture of varied language usage patterns of false and authentic reviews, rather than merely examining some features of distinct language usage patterns (e.g., excessive usage of extreme terms by Moon, Kim, &Bergey, 2019). As a consequence, they empirically demonstrate that their technique is more effective than existing computational linguistics practices at detecting fake reviews (Li et al., 2014). Importantly, when compared to standard computational linguistics approaches, their empirical study demonstrates three specific sources of excellence for the approach: (1) review corpus selection, (2) term dictionary, and (3) decision rule. Then, using review content analysis, Moon et al. (2021) uncover a number of characteristics linked to producing fake reviews (Burgoon& Qin, 2006). More specifically, compared to authentic evaluations, fake reviews are related with emotional exaggeration, a lack of specific information, an emphasis on the present and future rather than the past, and a certain pronoun use pattern (more first-person and less third-person pronouns).

Even when people are instructed to submit false reviews by dishonest firms, the circumstances are likely to have an impact on how the fake reviews are written. Academics and practitioners working to minimize biases induced by fake reviews might benefit from understanding such latent motives and features. This research reveals which categories are particularly important to hotel evaluations using the LIWC lexicon of words and categories, which is designed to be abstract. The LIWC categories are used as possible predictors of false reviews to determine whether content characteristics in uploaded reviews are indicative of fake reviews.

The Emotional Tone category is defined as positive emotion – negative emotion (Cohn,Mehl, &Pennebaker, 2004). Moon et al.'s (2021) findings on emotional tone are in line with Vrij's (2000) observation that falsehoods are more likely to utilize negative emotion terms. Fake reviewers are aggressive, employing exclamation marks and power phrases, and their tone is more negative than favorable.

The use of *first-person pronouns* is common, according to empirical findings, suggesting that the writer was attempting to persuade the reader that he or she had actually made the transaction being recounted. Furthermore, *third-person pronouns* were used less frequently in false reviews in review data, maybe because building a story through third-person involvement needs more creativity and is thus more cognitively taxing. This discovery takes us to the topic of cognitive heuristics. In conclusion, the outcomes of cognitive efficiency efforts demonstrate an increased usage of (1) quote marks, (2) less geographical features and (3) more frequent depictions of work-related material.

The final two categories represent the idea that false reviews are more focused on the present and future than real ones. They can speculate that when reviewers are intended to assess purchases, such as hotel visits that have already occurred in the past, they will presumably write in the past tense (Kronrod et al., 2017). However, because writing a false review necessitates making things up, authors may mistakenly write in the present and future tenses by forgetting to write in the past tense as if the visit had occurred.

Furthermore, because positive evaluations might have a favorable impact on future hotel revenue (Dellarocas, 2003), fake reviews could be written not just to represent previous experiences, but also to communicate current or future thoughts about the hotel (e.g., I see a bright future of the hotel). As a result, subtleties in tense usage may aid in the detection of fraudulent reviews. While Kronrod et al. (2017) found that legitimate reviews used more past tense verbs, Moon et al. (2021) suggests that the use of present and future tense verbs might be an indicator of a fraudulent review.

The All-Terms model (74,6%) outperforms MKB's (Moon, Kim, &Bergey, 2019) Extreme Terms model (56,6%) in terms of overall prediction accuracy. Given that the suggested All Terms procedure's major goal is to detect fake reviews, the writers' model better prediction accuracy over MKB is a testament to the procedure's uniqueness.

6.9.2 Linguistic characteristics of Fake Reviews

According to Neisari et al. (2021), some studies have employed textual content and linguistics aspects, alone or in combination with other factors, to assess the validity of the review. Language patterns, commonly used terms, word definitions, and term frequency are only a few of the textual content aspects. Textual content alone can generally detect false reviews with a modest level of accuracy, such as 75% (Kohonen & Mäkisara,1989). Spam identifiers that use linguistic traits are frequently deceived by clever

spammers who strive to produce opinions similar to genuine ones. Furthermore, textual properties are domain-specific, making it challenging to develop a unified cross-domain verification technique. For example, terms like "delicious" or "tasty" that are used to describe restaurants cannot be used to describe vehicle repair businesses. As a result, additional characteristics are frequently used in conjunction with contextual analysis to improve detection power and, as a result, prediction accuracy (Neisari et al., 2021).

Psycholinguistic deception detection is a problem related to fake news identification that looks into things like lying words, false beliefs, computer-mediated deceit in role-playing games, and so on (Li et al. 2014). Fraudulent reviews, on the other hand, exhibit significantly different dynamics, indicating that the characteristics used to identify lies aren't as successful in detecting fake reviews.

Ansari et al. (2021) take on a linguistic-based approach based on speech-act theory and extending the Ludwig et al. (2016) framework to the setting of e-commerce reviews. This method takes a closer step to understanding how the language presentation of online reviews affects a customer's impression of the review's deception. The findings are based on the perspectives of 202 consumers (online shoppers) who were polled via a sample of 120 smartphone reviews from Flipkart, India's top e-commerce site. The findings of the study show that customers' perceptions of reviews using reference cues (with more personal pronouns) are deemed dishonest. This research demonstrates that deception employs less self-references (first-person singular pronouns such as I, me, and my) and other references (third-person pronouns, such as he, she, and they).

When opposed to fake reviews based on imagination, authentic reviews based on genuine experiences tend to be more detailed. After all, producing false evaluations necessitates recounting events that did not occur in fact and conveying non-existent opinions (Newman et al., 2003). As a result, fake reviews are frequently found insufficient in terms of specificity. As a result, although genuine evaluations may contain extensive objective information, fake reviews may contain a lot of imprecise and non-content phrases with insufficient data (Hancock et al., 2008; Vrij et al. 2000). Informativeness, perceptual details, contextual details, lexical variety, and the usage of function words may all be used to determine the level of depth in reviews (Banerjee & Chua, 2014; Ott et al., 2011; Yoo & Gretzel, 2009).

Cognition indications are language cues that may be exposed as a result of carelessness when creating fake reviews. Individuals who participate in fake behavior are frequently

stimulated physically and mentally, which can be difficult to disguise (Zuckerman et al., 1981). Despite best efforts, arousal often leaks through in the form of verbal signals that can be used to identify deceit (Vrij et al. 2000). Writing false reviews, for example, is considered intellectually challenging (Newman et al., 2003). Fake reviews, given their difficulties, may contain more evidence of cognition than genuine submissions (Pasupathi, 2007).The use of discrepancy words such as "should" and "may", fillers such as "you know" and "like", tentative words such as "perhaps" and "guess", causal words such as "because" and "hence", insight words such as "think" and "consider", motion words such as "arrive" and "go", and exclusion words such as "without" and "except", could all be used to determine cognitive indicators in reviews (Boals & Klein, 2005; Newman et al., 2003; Pasupathi, 2007).

Because reviews have been shown to impact a company's current image and future revenues (Banerjee & Chua, 2014), fraudulent reviews might include less past tense and more present and future tense to influence a hotel's current and future reputation (Tausczik & Pennebaker, 2010). Fake reviews might also employ more causal, insight, and motion words than real reviews, but less exclusion words (Boals & Klein, 2005; Newman et al., 2003; Tausczik & Pennebaker, 2010).

Prior research in personality and psychology (e.g., Newman et al., 2003) has shown that deception/lying often involves more use of personal pronouns (e.g., "us") and associated actions (e.g., "went," "feel") towards specific targets with the goal of incorrect projection (lying or faking), which often involves more use of positive sentiments and emotion words (e.g., "nice," "deal," "comfort," "helpful" etc.) (Mukherjee et al., 2013).

6.9.3 Studies on online reviews websites

"Dianping", China's version of Yelp (a restaurant rating website), has developed a mechanism to detect false reviews. It has been demonstrated that the system's accuracy is quite high, implying that when the algorithm detects a fraudulent review, it is nearly certainly a fake review. There are two factors that strengthen this faith in its accuracy. To begin with, Dianping has a staff of specialist assessors tasked with assessing its detection system. They carefully analyze a random sample of discovered fake reviews every week based on all the data they collected (e.g., reviews, side information, IP addresses, click data, etc). Secondly, the fact that Dianping sends an email to its reviewer with reasoning for each discovered fake review results in an even stronger piece of

proof. Dianping, on the other hand, has no idea what the actual recall of their system is because no one knows how many fake reviews there are. The system's high accuracy and unknown recall imply that the fraudulent reviews it detects are virtually likely fake, but the remaining reviews may not be entirely real, i.e., they may contain numerous fake reviews that Dianping's system is unable to detect (Li et al. 2014).

Wang et al. (2021) performed two researches that looked at Yelp.com restaurant and lodging reviews. The authors used linguistic inquiry and word count (LIWC) 2015 to code the review contents and used logistic regression to assess the hypotheses. Emotional signals are associated to review fakeness in a positive way, meaning that fake reviews contain more emotional cues than real evaluations. Affective expressions in communication are referred to as emotional cues (DePaulo et al., 2003). According to interpersonal deception theory, deceivers exhibit more emotional expression than truth-tellers, such as fear and remorse (Porter et al., 2012; Vrij et al., 2019). These nonstrategic emotions are deception detection leakage indications (DePaulo et al., 2003). Deceivers, on the other hand, are better prepared in an online setting than they are in a face-to-face encounter. Review fabricators have plenty of time and can accurately mimic the written sentiment of honest reviewers (Johnson, 2007). Furthermore, in the online writing context, deceivers' quick replies are not visible. As a result, non-strategic emotions and leakage cues are unlikely to have a significant role. In the internet world, deceivers may be more prone to deploy strategic emotions. Review fabricators, for example, employ more emotive language to entice and persuade potential customers (DePaulo et al., 2003). When promoting firms, reviewers tend to overstate the benefits by using more favorable sentiments. Reviewers tend to assume more negative feelings while repressing competition in order to harm their reputation. Emotional signals can be used to tell the difference between fraudulent and genuine reviews (Li et al., 2020). Reviews that contain a larger %age of emotive language are more likely to be false than genuine (Wang et al., 2021)

6.9.4 Linguistic signals and emotion

Li et al. (2020) suggests four language signals connected to a reviewer's psychological processes (i.e., affective, cognitive, social, and perceptual) and investigate their links with false reviews, as well as the impact of time distance and reviewer location on these reviews. The findings of a logistic regression study (using SPSS) of 43,496 Yelp.com

reviews reveal that affective, social, and perceptual signals are highly associated to fake reviews, with time distance and reviewer location having significant effects.

Lying is frequently linked to unpleasant emotional states including shame, remorse, and fear (Toma and Hancock, 2012). Individuals who create fraudulent internet reviews, on the other hand, may utilize both positive and negative emotional expressions more frequently than those who publish real ones, because online reviews are primarily written language.

People utilize social signals in their online messages, either consciously or subconsciously, to indicate their sociability with third parties other than themselves, a sense of belonging to certain groups of people, and social identity, and these social cues help people form and sustain social connections (Pennebaker et al., 2003). According to several research, third-person pronouns (e.g., s/he or her/his) are employed more frequently in misleading content in an online context than first-person pronouns (e.g., I or my) (Hancock et al., 2007; Van Swol and Braun, 2014).

To make a false review appear more legitimate, the deceiver may strive to add more comprehensive perceptual expressions (Burgoon et al., 2001). However, it has been suggested that giving perceptual details may be more challenging for someone who fabricates fake evaluations or opinions (Vrij et al., 2000). Simply stating that something "tastes nice" is less persuasive than providing a full description of the dish's flavor, such as "the color and plating were exquisite and the texture was soft and fresh." The latter is impossible to express without having eaten a real meal at the restaurant; hence, deceivers who wish to make their phony review appear genuine will find it difficult to include perceptual language in their review (Li et al., 2020). As a result, perceptual signals are linked to fake reviews in a negative way.

6.9.5 Coh-Metrix

Coh-Metrix is a complex linguistic analysis tool (Graesser et al., 2011) that might be used instead of the widely used LIWC since it examines language at a higher level. When trained and evaluated with Coh-Metrix output, binominal regressions provide better detection accuracy. With an overall detection rate of 76.6 %, the program recognized 82.4 % of fraudulent reviews and 66 % of authentic reviews for positive evaluations. The detection accuracy for unfavorable reviews was 66.8%, with 53.4 % of actual reviews and 83.9 % of false reviews accurately recognized. It has also shown a total of 29

indices that, when integrated in a binominal logarithmic regression, gave an 81 % detection accuracy, with 89.7% deception detection and 58 percent truth detection accuracy. Although the combined algorithm increased the identification of fake reviews, it had a significant false negative error rate (42%).

The output of the Coh-Metrix provided more generic results. True messages were often larger (number of paragraphs and sentences per paragraph), simpler to read, and more exact in terms of terminology (hypernymy), according to Plotkina et al. (2020); but they lacked temporal cohesion. True messages also had fewer adjectives and more connectives (logical and adversative). Positive and negative online evaluations had diverse language patterns, according to Coh-Metrix; yet, while the reviews of various valences did not share all of the same cues, they did not contradict each other.On the basis of this information, it was decided to divide the positive and negative evaluations and examine them independently. Only one element of the reasons to produce a false review differed significantly between the control and reputational review (social motivation) groups: syntactic simplicity (Plotkina et al., 2020).

- *In terms of comprehensibility*, real and fake review titles differed considerably in terms of two structural features: the number of words and the percentage of lengthy words. Longer titles were utilized in authentic reviews, while fewer lengthy words were used. In addition, three structural elements of real and fictional evaluations differed significantly: characters per word, words per sentence, and percentage of lengthy words. Genuine reviews had fewer words per character, fewer lengthy words, but longer sentences.
- *In terms of three informativeness variables*: nouns, conjunctions, and pronouns, the titles of authentic and fictional reviews differed considerably in terms of specificity. When compared to fake entries, authentic evaluations had more nouns, conjunctions (e.g. but), but fewer pronouns. The DBPM analysis was unable to uncover any pronouns that were significantly different. Authentic reviews have much more spatial words than fictional evaluations in terms of contextual data. In addition, the four informativeness factors of nouns, articles, verbs, and pronouns differed considerably between authentic and fictional evaluations. The number of nouns in authentic evaluations was much higher, while the number of articles, verbs, and pronouns such as "my" was significantly lower. Authentic reviews had more spatial words than fictional reviews in terms of contextual data.
- *In terms of exaggeration*, the affectiveness of actual and fictional review headlines differed dramatically. Positive emotion words were prominent in authentic evaluations, whereas negative emotion terms were scarce. Furthermore, as compared to fictional posts, actual evaluations were substantially less emotional and included fewer words such as "poor."

- *In terms of carelessness*, the use of self-references in the form of first-person singular words differed considerably between real and fictional review titles. When compared to fictional posts, authentic evaluations included less first-person singular terms. Genuine reviews had much fewer modal verbs and filler words than fictional posts in terms of ambiguity words. The DBPM analysis, however, was unable to find any statistically different first-person singular words, modal verbs, or filler words. In addition, the use of self-references in the form of first-person singular terms differed considerably between real and fictional assessments. In comparison to fictional posts, real reviews had less first-person singular terms like "I." When it came to cognitive words, the former had more exclusion words like "but."

This study linguistically examined a dataset of 1,800 hotel evaluations (900 authentic1900 fake), which were assessed using 83 factors according to the suggested framework. C4.5, JRip, logistic regression, random forest, and support vector machine were used to classify authentic and fraudulent reviews using average probability voting among five frequently used supervised learning algorithms: C4.5, JRip, logistic regression, random forest, and support vector machine (Ghose&Ipeirotis, 2011; Ott et al., 2011; Zhou, Burgoon, Twitchell, Qin, & Nunamaker Jr, 2004). Using independent samples t-tests, the feature-selected variables were subsequently examined to discover a filtered set of linguistic factors that varied between authentic and fictional reviews. With an accuracy of 77.28 %, the suggested language framework functioned admirably (Banerjee et al., 2017).

Banerjee et al. (2014) used two publicly accessible secondary opinion spam datasets to empirically evaluate the suggested linguistic framework. The total number of reviews included in the research was 1,600, evenly spread across 20 of Chicago's most popular hotels. The Linguistic Inquiry and Word Count (LIWC) method, an automated text analysis tool, may be used to compute linguistic indicators of textual content.

Genuine reviews got lower readability scores for all three variables, namely, FOG, CLI (linguistic complexity), and ARI (ease of reading), when compared to misleading ones. The former appeared to be less complicated and more straightforward to read than the later. Genuine reviews appeared to have more adjectives, articles, and nouns in their POS tags, whereas misleading reviews appeared to have more prepositions, adverbs, verbs, and pronouns. Deceptive reviews proved to be more heavily inflated with self-references past tense, function words, and perceptual words than genuine evaluations in terms of writing style.

The only POS tag that proved to be a significant predictor of review genre was verbs. The negative link suggested that the larger a review's share of verbs, the less likely it was to be authentic. As a result, misleading evaluations appear to have more verbs than real reviews.

When compared to authentic evaluations, deceptive reviews tended to be more heavily embellished with self-references perceptual terms, and function words. The usage of the past tense, on the other hand, was positively related to the dependent variable. The greater the amount of past tense in a review, the more likely it was to be authentic. In comparison to fraudulent evaluations, genuine reviews seemed to be richer in the past tense (Banerjee et al., 2017).

6.10 Trolling

Trolling is defined as the act of purposefully uploading provocative or controversial information to an online community in order to provoke readers or disrupt dialogue. The term "troll" is now often used to describe someone who harasses or insults people online (Wardle et al., 2018).

A troll is a user who creates the appearance of genuinely wanting to be a part of the group in question, including declaring or transmitting pseudo-sincere goals, but whose true intention(s) is/are to disrupt and/or increase disagreement in conversation (Hardaker, 2015). Trolling is defined as a special sort of malevolent online behavior aimed at disrupting interactions and general online discourse, aggravating conversational partners, and luring them into unproductive discussion (Coles & West, 2016). For a long time, trolling was associated with antisocial behavior in gaming communities and fringe discussion forums like "4chan" (Kirman et al., 2012; Samory & Peserico, 2017).Trolling rapidly spread on social media, where trolls weren't just "amusing themselves by upsetting other users," but were actively spreading false information as part of a state-sponsored effort to sway public opinion on political candidates, public health issues like vaccination, and social justice issues (Broniatowski et al., 2018; Llewellyn et al., 2018; Stewart et al., 2018; Zannettou et al. 2019). Badawy et al., 2019 define trolling as “users who exhibit a clear intent to deceive or create conflict with the goal to manipulate the public opinion on a polarized topic and cause distrust in the socio-political system”. Trolling has progressed from the use of foul language to the spread of misinformation, rumors, and fake news (Benkler et al., 2018).

6.10.1 The tools of trolls

It is obvious that there is a relationship and mutual connection between conspiracy theories and trolls. “Linguistic Cues to Deception: Identifying Political Trolls on Social Media” measures how deceitful trolls' tweets compare to authentic users using misleading language cues. Trolls employ misleading language to deceive people into believing the information they offer, as deception often comprises words and information purposefully delivered to establish a false conclusion (Buller et al., 1994). When it comes to evaluating communications, they get on social media, people have a tendency to be truth-biased (Levine, Park, and McCornack, 1999). As a result, human deception detection accuracy is just slightly better than chance (Frank and Feeley, 2003).

When the subject of debate is a contentious disputed matter, such as politics, this issue becomes even more evident, because internet consumers are confronted to much more political information published by ordinary people than before. According to Bakshyet al. (2015), 13% of Facebook users that declare their political stance publish political news. Furthermore, many posts might not be created by people at all. For example, troll accounts and social bots tried to sway the 2016 presidential election in the United States by sending out fraudulent tweets, or "fake news," in support or opposition to specific parties (Pennycook and Rand, 2018). Before the 2016 election, this fraudulent, fabricated information was shared with a wealth of Americans on social media (Twitter and Facebook). Super-users are occasionally targeted by bots that add responses and mentions to postings (De Beer and Matthee, 2020). Such behaviors can be used to encourage others to spread false information (Shao et al. 2018).

Trolls are user identities whose main aim is to display conflict and deceit, according to “Linguistic Cues to Deception: Identifying Political Trolls on Social Media”. Their goal in the 2016 elections was to destabilize the democratic process and ignite distrust in the political system. The Russian government allegedly financed these trolls in order to influence debates about political topics in order to create conflict and hatred among various groups (Gerber and Zavisca, 2016). Deception in presidential politics in the United States has gotten increasingly common over the last several decades, according to Stanley Renshon (Borenstein, 2016). It is critical to recognize communication that is intended to deceive and mislead in order to resist the corrosive impact of online political manipulation. However, until recently, the problem of automated identification of false material had received little attention. This report fills in the gaps by conducting an em-

pirical assessment of Russian trolls' misleading language in their attempts to sway US elections. This might lead to better techniques for detecting disinformation spread by fake accounts on Twitter.

6.10.2 Linguistic characteristics of trolling

Addawood et al. (2019) compiled a dataset of 13 million election-related messages published on Twitter in 2016 by over a million different individuals for their study. About the dataset, it contains accounts linked to the confirmed Russian trolls, as well as individuals who shared postings on a range of topics concerning the 2016 elections over the same time period. To investigate the manipulation of public opinion, researchers selected and analyzed 49 linguistic cues as possible indications of misleading language, concluding that the quantity of hashtags and retweets are the most crucial troll evidence. With a mean F1 score of 82 percent and recall of 88 percent, this study found that false linguistic cues can aid in effectively identifying trolls.

Addawood et al. (2019) employ misleading language indicators to match the tweets of deceptive trolls to those of authentic users. Trolls employ misleading language to deceive people into believing the information they offer, as deception often comprises words and information purposefully delivered to establish a conclusion that was incorrect (Buller et al., 1994).

When it comes to evaluating communications, they get on social media, people have a tendency to be truth-biased (Levine, Park, and McCornack, 1999). As a result, human deception detection accuracy is just slightly better than chance (Frank and Feeley, 2003).

- *Words used:* Deception was associated to uncertainty. Trolls utilize less hedges, model verbs and modifiers than non-trolls, contradicting the premise. Other verbal signs indicating doubt, such as the usage of quotes and questions, were, on the other hand, much greater in trolls than in non-trolls. In addition, trolls employ less subjective language than non-trolls. Trolls are less certain, that leads to more deceit, because subjectivity is employed to communicate opinions and judgements (Banfield, 1982; Wiebe, 2000).
- *Self-Reference:* Trolls make far fewer references to themselves and others than non-trolls. In addition, trolls employ far less generic phrases and indefinite articles than non-trolls, contradicting the theory that they utilize a more broad narrative to separate themselves from the deceit.

- *Causal phrases:* Trolls employ a smaller number of discourse markers. In the same way, using causal phrases adds complexity and specifics to a tale while also increasing the risk of self-contradiction. It is discovered that trolls utilize fewer causality words, such as "because," and less sense keywords.
- *Emotion:* Furthermore, trolls write with substantially less emotion than non-trolls, according to the research. This contrasts prior research that showed deceivers use more emotive language (Zhou et al., 2004; Burgoon et al., 2003). Another sign of specificity is the usage of relativity terms; it appears that trolls use less relativity words than humans, reflecting earlier research (Perez-Rosas et al., 2017).
- *Size of words and phrases:* Trolls are shown to have less complicated, shorter tweets than non-trolls, as well as less sophisticated terms (with fewer than six letters). In comparison to non-trolls, they utilize far more words in each phrase and more punctuation. Trolls, it was hypothesized, use fewer words and phrases to communicate their apprehension. Trolls used considerably fewer nouns, verbs, adverbs, and prepositions in their tweets, confirming previous study on deceit (Burgoon et al., 2003). Trolls, on the other hand, utilized substantially more words in total than non-trolls. Trolls have a larger word count than non-trolls, although these words are not essential components of speech like nouns and verbs.
- *Additional characteristics:* Trolls employed linguistic cues that were extremely compelling. For example, it was discovered that trolls utilized much more URLs in their tweets than non-trolls (Tan et al., 2016; Khazaei, Lu, and Mercer, 2017). Trolls also employ fewer function words, as evidenced by earlier research (Khazaei, Lu, and Mercer, 2017). Trolls also employ much less present-focused terms than non-trolls, despite the fact that the usage of present tense has been proven to be a part of non-persuasive remarks (Xiao, 2018). Tweets that are less reward-oriented are regarded to be more convincing (Xiao, 2018). Trolls employ much less reward-focused terms than non-trolls in this data.Trolls utilized much more hashtags than non-persuasive tweets, confirming the hypothesis that persuasive tweets contain more hashtags.

Trolls have much lower moral standards than non-trolls, according to the findings of this study. This supports the theory that employing fewer moral cues in the text indicates the user is attempting to deceive. The misleading language traits were captured by the theory-driven linguistic analysis. Troll accounts, for example, are shown to employ far more persuasive linguistic signals and significantly less nuanced and specific language. They utilized these linguistic signals to create a classifier that accurately identified trolls (average F1 score is 82% and recall is 88%). While metadata variables were extremely unique and predictive of troll accounts, some language features, particularly ones linked to information complexity and persuasion, were also predictive of troll ac-

counts. The use of more hashtags, tweets, and retweets is linked to a higher risk of becoming a troll, as does the use of fewer nouns and publishing shorter tweets with fewer words (Addawood et al.2019).

6.10.3 Other Studies

Miao et al. (2020) compared the results of three machine learning approaches when applied to tweets using stylometric information. They discovered that troll tweets are more likely to utilize a bigger number of digits than normal tweets after analyzing them in terms of these high-importance traits. In a character n-grams model, '2016' is discovered to have significant influence on classifying troll tweets and regular tweets. Troll tweets are also often shorter (in terms of both the amount of words and the number of sentences) than normal tweets. Conjunctions and pronouns, on the other hand, are less common in troll tweets than in normal tweets. They also discovered that stop words are more common in typical tweets, which they attribute to the large quantity of pronouns and conjunctions.

6.11 Pseudoscience

Pseudoscience is defined as information that makes questionable or incorrect claims about legitimate scientific investigations. Experts are frequently coming in contrast with this type of information (EAVI, 2018). It encourages erroneous notions like metaphysics and naturalistic fallacies (Guacho et al., 2018). The actors use scientific legitimacy in order to get money or fame (Forstrop, 2005).

Pseudoscience has a huge impact on people's lives, causing readily avoidable calamities. Vaccination, cancer, nutrition and smoking are all well-studied areas in medicine and healthcare that include misinformation (Albarracin et al., 2018; Jolley and Douglas, 2014; Syed-Abdul et al., 2013; Wang et al., 2019). Since the COVID19 outbreak, it was propagated that mortality rates are overstated and hence there is no necessity to follow lockdown rules or different social distancing practices, that could have contributed to the spread of the virus (Lynas, 2020). In addition, misinformation can have a harmful influence on environmental policy.

7 Results

7.1 RQ1: Which are the categories of linguistic features?

Table 1: Linguistic features categories

Category	Example	Dimension	Author
Lexical	> Syllables, Words, Sentences > Big words, Syllables per word, Sentence length	>Quantity > Complexity	Burgoon et al. (2003)
Grammatical	Rate of adjectives and adverbs	Specificity & Expressiveness	Burgoon et al. (2003)
Lexical	Word length, Preposition	Complexity	Newman et al. (2003)
Grammatical	Pronouns, First & Third person, Articles, Negation	Linguistic Dimension	Newman et al. (2003)
Lexical	> Words, Sentences, Noun phrases > Sentence length, word length, noun phrase length	> Quantity > Complexity	Zhou et al. (2004)
Grammatical	> Verbs > Modal verbs > Passive voice	> Quantity > Uncertainty > No immediacy	Zhou et al. (2004)
Sentiment-based	> General terms, self-reference > Emotion words, redudancy > Spatio- temporal information, perceptual > Positive, negative words	> No immediacy > Expressivity > Specificity > Affect	Zhou et al. (2004)
Grammatical	> Personal pronoun, Negation > Proper noun, Adverb, Stative verb, Comparative and superlative adjective, Conjunction > Passive voice > Modal verbs	> Linguistic Dimension > Complexity > No immediacy > Uncertainty	Mahyoob et al. (2020)
Lexical	Sentence length	Complexity	Mahyoob et al. (2020)
Syntactic	> Question mark > Quotation	> Uncertainty > Complexity	Mahyoob et al. (2020)
Lexical	Text length	Complexity	Deng et al. (2021)
Sentiment-based	Emotional cues	Expressivity	Deng et al. (2021)
Digital	Hashtag, URL	Persuasion	Deng et al. (2021)
Sentiment-based	Persuasive, Comparative, Emotional (positive and negative) and Uncertainty words	Persuasion, Expressivity, Uncertainty	Zhou et al. (2021)
Lexical	> Word and sentence length & Number of syllables > Number of words, Number of sentences	> Complexity > Quantity	Verma et al. (2021)
Sentiment-based	Words about emotions, behaviours	Psychological Processes	Verma et al. (2021)

After our extended research, theories of linguistics from 2003 to 2021 show that more popular are lexical features, then grammatical and sentiment-based. Only in the research of Mahyood et al. (2020) we see some syntactic features and in Deng et al. (2021) survey is containing some digital linguistic features. The most noted dimension within the bibliography is Complexity and then Quantity and Uncertainty.

7.2 RQ2: What linguistic features do we meet for disinformation and/or how these are used by fake news detection tools?

Table 2: Linguistic features in different types of disinformation

Type of disinformation	Linguistics			Paper	Practical Implementation/ Model	Domain	Accuracy
	Category	Feature	Description				
Conspiracy Theories	Grammatical	> Smaller usage of the third singular (i.e. she or he) and the first plural person (i.e. we) > Focus more on present than past or future	> Linguistic Dimension > More Relativity	Detection of Conspiracy Propagators Using Psycho-Linguistic Characteristics (Giachanou et al., 2021)	Convolutional Neural Network (CNN)/ ConspiDetector	Social Media (Twitter)	74%
	Sentiment-based	> More words about religion, swear words, assent words (e.g. agree, yup, okey) > Less personal concerns, usage in causation (because, effect, hence), usage regarding discrepancy (should, would, could) and tentative (e.g. maybe, perhaps)	> Psychological Processes	Detection of Conspiracy Propagators Using Psycho-Linguistic Characteristics (Giachanou et al., 2021)	Convolutional Neural Network (CNN)/ ConspiDetector	Social Media (Twitter)	74%
	Sentiment-based	> Higher rate on negative sentiment, fear, anger, and disgust	> Psychological Processes	Thought I'd Share First' and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study (Gerts et al., 2021)	Sentiment Analysis/ AFINN, NRC	Health (Tweets about COVID-19)	
Hoaxes	Syntactic	> Less complex language that is easier to comprehend with simple sentences	> Low Complexity	Detecting Hoaxes, Frauds, and Deception in Writing Style Online (S. Afroz et al., 2012)	Statistical methods/ Using a large feature set	Politics, technology, etc	97%
	Lexical	> Fewer average syllables per word > Shorter sentences	> Low Complexity > Low Complexity	Detecting Hoaxes, Frauds, and Deception in Writing Style Online (S. Afroz et al., 2012)	Statistical methods/ Using a large feature set	Politics, technology, etc	97%
Rumors	Sentiment-based	> Words related to social relationships (e.g., family, mate) are more frequently used & More words related to skepticism and doubts (eg negation, speculation) & Higher fraction of words related to social and hear (eg friend, buddy, neighborhood	> More Psychological Processes	Aspects of Rumor Spreading on a Microblog Network (Kwon et al., 2013)	Sentiment Analysis/ LIWC	Social (popular events tweets)	>
	Sentiment-based	> Fewer positive emotions & Use of sad words (eg crying, grief, sad)	> Psychological Processes	Rumor Conversations Detection in Twitter through Extraction of Structural Features (Lotfi et al., 2021)	Machine learning techniques/ Model for manually annotated conversations	News from all over the world (Twitter)	> 4%
	Syntactic	> More question marks	> More Uncertainty	Rumor Conversations Detection in Twitter through Extraction of Structural Features (Lotfi et al., 2021)	learning techniques/ Model for manually annotated conversations	News from all over the world (Twitter)	> 4%
Clickbait	Sentiment-based	> Hyperbole and slang phrases	> More Psychological Processes	Detecting Fake News in Social Media Networks (Aldwairi et al., 2018)	Machine Learning Techniques/ WEKA	Social Media ads & articles (Facebook, Forex, Reddit)	99%
	Sentiment-based	> Fear, sadness, and surprise emotions words	> Psychological Processes	FakeFlow: Fake News Detection by Modeling the Flow of Affective Information (Ghanem et al., 2021)	Fake news detection model/ FakeFlow	News Headlines	85%
	Lexical	> Longer words	> High Complexity	Detecting Fake News in Social Media Networks (Aldwairi et al., 2018)	Machine Learning Techniques/ WEKA	Social Media ads & articles (Facebook, Forex, Reddit)	99%

	Syntactic	> Large amount of exclamation marks and question marks	> High Expressiveness	Detecting Fake News in Social Media Networks (Aldwairi et al., 2018)	Machine Learning Techniques/ WEKA	Social Media ads & articles (Facebook, Forex, Reddit)	99%
Misleading Connection	Grammatical	> More use of present or future verbs	> More Relativity	Between thinking and speaking—Linguistic tools for detecting a fabrication (Dilmon, 2009)	Statistican Methods/ MANOVA	Personal Stories	
	Sentiment-based	> Large number of negative words > Use of words that don't contribute any new information to the story > Use of ambiguous words	> Psychological Processes (High Persuasion, Concealment, Vagueness)	Between thinking and speaking—Linguistic tools for detecting a fabrication (Dilmon, 2009)	Statistican Methods/ MANOVA	Personal Stories	
Fake Reviews	Sentiment-based	> Overproduced negative emotion terms & general words & lack of specific detail & power terms	> More Psychological Processes	Content Analysis of Fake Consumer Reviews by Survey-Based Text Categorization (Moon et al., 2021)	n-gram–based & Sentiment Analysis/ Support Vector Machine (SVM) Classifier & LWIC	Hospitality (Hotel reviews)	75%
	Sentiment-based	> More discrepancy words (eg should, may), fillers (eg you know, like), tentative words (eg perhaps, guess), casual words (eg because, hence), insight words (eg think, consider), motion words (arrive, go) and exclusion words (eg without, excect)	> More Psychological Processes	Customer Perception of the Deceptiveness of Online Product Reviews: A Speech Act Theory Perspective (Ansari et al., 2021)	Linguistic-based approach grounded in the speech-act theory	Technology (smartphone reviews)	
	Sentiment-based	> More associated actions (eg went, feel) & More use of positive sentiments and emotion words (eg nice, deal, comfort, helpful)	> More Psychological Processes	What Yelp Fake Review Filter Might Be Doing? (Mukherjee et al., 2011)	Personality and Psychology research	Bibliography	
	Sentiment-based	> Greater emotional expression such as nervousness and guilt and higher proportion of emotional words	> More Psychological Processes	Detecting Fake Hospitality Reviews through the Interplay of Emotional Cues, Cognitive Cues and Review Valence (Wang et al., 2021)	Linguistic Analysis/ LWIC	Hospitality (Hotel reviews)	
	Sentiment-based	> More positive and negative emotional expressions & More detailed perceptual expressions (to make a fake review look more authentic)	> More Psychological Processes	Unveiling the Cloak of Deviance: Linguistic Cues for Psychological Processes in Fake Online Reviews (Li et al., 2020)	Logistic regression analysis (using SPSS)	Businesses (Yelp.com)	
	Sentiment-based	> More precise words	> More Psychological Processes	Illusions of Truth—Experimental Insights into Human and Algorithmic Detections of Fake Online Reviews (Plotkina et al., 2020)	Linguistic analysis/ Coh-Metrix		90%
	Sentiment-based	> Less spatial, positive emotion words > More filler words	> Less Psychological Processes > Less Psychological Processes	Don't be deceived: Using linguistic analysis to learn how to discern online review authenticity (Banerjee et al., 2017)	Supervised learning algorithms/ C4.5, JRip, logistic regression, random forest, support vector machine	Hospitality (Hotel reviews)	77%
	Sentiment-based	> More function and perceptual words	> More Psychological Processes	A Linguistic Framework to Distinguish between Genuine and Deceptive Online Reviews (Banerjee et al., 2014)	Linguistic Analysis/ LWIC	Hospitality (Hotel reviews)	
	Grammatical	> Less past sense verbs (more present and future language) > More first-person(pronouns) & less third-person pronouns	> Less Relativity > Low linguistic dimension	Content Analysis of Fake Consumer Reviews by Survey-Based Text Categorization (Moon et al., 2021)	n-gram–based & Sentiment Analysis/ Support Vector Machine (SVM) Classifier & LWIC	Hospitality (Hotel reviews)	75%

	Grammatical	> Fewer first-person singular pronoun (such as I, me, my) and third-person pronouns (such as he, she, they) > Fewer past tense and more present and future tense	> Linguistic dimension > Less Relativity	Customer Perception of the Deceptiveness of Online Product Reviews: A Speech Act Theory Perspective (Ansari et al., 2021)	Linguistic-based approach grounded in the speech-act theory	Technology (smartphone reviews)	
	Grammatical	> More personal pronouns (eg "us")	> Linguistic dimension	What Yelp Fake Review Filter Might Be Doing? (Mukherjee et al., 2011)	Personality & Psychology re-search	Bibliography	
	Grammatical	> More third-person pronouns (e.g., s/he or her/his)	> Linguistic dimension	Unveiling the Cloak of Deviance: Linguistic Cues for Psychological Pr o-cesses in Fake Online Reviews (Li et al., 2020)	Logistic regres sion analy-sis (using SPSS)	Businesses (Yelp.com)	
	Grammatical	> More connectives (logical and adversative)	> High Complexity > Less Complexity	Illusions of Truth—Experimental In sights into Human and Algorithmic Detections of Fake Online Reviews (Plotkina et al., 2020)	Linguistic anal ysis/ Coh-Metrix		90%
	Grammatical	> Fewer nouns and conjunctions, but more pronouns and adjectives > More first-person singular words > More modal verbs	> Complexity > Linguistic Dimension > Uncertainty	Don't be deceived: Using linguistic analysis to learn how to discern on-line review authenticity (Banerjee et al., 2017)	Supervised learning algor-ithms/ C4.5, JRip, logistic regression, random forest, support vector machine	Hospitality (Hotel reviews)	77%
	Grammatical	> Fewer adjectives, articles and nouns > More prepositions, adverbs, verbs and pronouns > More self-references > Fewer past tense	> Low Complexity > High Complexity > Linguistic Dimension > Less Relativity	A Linguistic Framework to Distin-guish between Genuine and Dece-ptive Online Reviews (Banerjee et al., 2014)	Linguistic Analysis/ LWIC	Hospitality (Hotel reviews)	
	Lexical	> Less number of paragraphs and sentences	> Low Complexity	Illusions of Truth—Experimental In sights into Human and Algorithmic Detections of Fake Online Reviews (Plotkina et al., 2020)	Linguistic anal ysis/ Coh-Metrix		90%
	Lexical	> More characters per word > Short titles but with long words & Longer words but shorter sentences	> High Complexity > Complexity	Don't be deceived: Using linguistic a-nalysis to learn how to discern on-line review authenticity (Banerjee et al., 2017)	Supervised lea rning algor-ithms/ C4.5, JRip, logistic regression, random forest, support vector machine	Hospitality (Hotel reviews)	77%
	Syntactic	> Exclamation marks	> Expressiveness	Content Analysis of Fake Consumer Reviews by Survey-Based Text Categorization (Moon et al., 2021)	n-gram–based & Sentiment Analysis/ Sup-port Vector Machine (SVM) Classifier & LWIC	Hospitality (Hotel reviews)	75%
Trolling	Grammatical	> Fewer modifiers, hedges and modal verbs >Fewer indefinite articles > Fewer discourse markers > Fewer function and present-focused words	> Uncertainty > No-immediacy > Specificity > Persuasion	Linguistic Cues to Deception: Identifying Political Trolls on Social Media (Addawood et al., 2019)	Linguistic cues Analysis	Politics on social media (Twitter)	82%

	Grammatical	> Fewer conjunctions and pronouns	> Low Complexity	Detecting Troll Tweets in a Bilingual Corpus (Miao et al., 2020)	3 machine learning methods/ Comparative performance	Social Media (Twitter)	
	Sentiment-based	> Less subjective language > Fewer general terms > Use of causation words (eg because) & fewer sense terms and relativity words & less emotion > Fewer reward-focused words	> Uncertainty > No-immediacy > Specificity > Persuasion	Linguistic Cues to Deception: Identifying Political Trolls on Social Media (Addawood et al., 2019)	Linguistic cues Analysis	Politics on social media (Twitter)	82%
	Syntactic	> Use of questions > Less complex words & quotations > More punctuation	> Uncertainty > Low Complexity > High Complexity	Linguistic Cues to Deception: Identifying Political Trolls on Social Media (Addawood et al., 2019)	Linguistic cues Analysis	Politics on social media (Twitter)	82%
	Lexical	> More words (but not important parts of speech) > More words per sentence	> Big Quantity > High Complexity	Linguistic Cues to Deception: Identifying Political Trolls on Social Media (Addawood et al., 2019)	Linguistic cues Analysis	Politics on social media (Twitter)	82%
	Lexical	> Fewer words > Shorter sentences	> Low Quantity > Low Compexity	Detecting Troll Tweets in a Bilingual Corpus (Miao et al., 2020)	3 machine learning methods/ Comparative performance	Social Media (Twitter)	
	Digital	> More hashtags & Use of links, URLs	> Persuasion	Linguistic Cues to Deception: Identifying Political Trolls on Social Media (Addawood et al., 2019)	Linguistic cues Analysis	Politics on social media (Twitter)	82%
Fabricated	Lexical	> Fewer words, paragraphs (stories) > Longer paragraphs (stories) & More number of words per paragraph (stories) & More words (headlines)	> Low Complexity > High Complexity	Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017), 2018	AI Techniques/ Natural language processing (NLP)	Politics (digital articles)	
	Syntactic	> More punctuation (headlines)	> High Complexity	Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017), 2018	AI Techniques/ Natural language processing (NLP)	Politics (digital articles)	
	Sentiment-based	> More positive and negative affect & more emotiveness (headlines) & more informal words > More slang, swear and affective words (stories)	> More Psychological Processes	Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017), 2018	AI Techniques/ Natural language processing (NLP)	Politics (digital articles)	
	Grammatical	> More demonstratives like pronouns and unspecific (headlines) & less verifiable facts like specific names (headlines)	> Uncertainty	Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017), 2018	AI Techniques/ Natural language processing (NLP)	Politics (digital articles)	
Biased or one-sided	Sentiment-based	> Emotionally driven language (especially negative emotion)	> Psychological Processes	What Does Fake Look Like? A Review of the Literature on Intentional Deception in the News and on Social Media (Damstra et al., 2021)	Linguistic Analysis/ Tweets	Politics	
Pseudoscience							

All the linguistic features from our research were divided into four basic categories: Lexical, Grammatical, Syntactical and Sentiment-based and the types of disinformation that contained all the above categories were Fake reviews, Trolling and Fabricated. On the other hand, biased or one-sided had only one category, sentiment-based, which was missing only in hoaxes (18 mentions). The next most "popular" category is grammatical (12) and the last ones are lexical (7) and syntactical (6). Fake reviews and Trolling have the biggest variety in different linguistic features.

Between linguistic features, there are a lot of similarities and correspondence between the different types of disinformation. The extensive use of present and future tense over past tense find agreement between misleading, fake reviews and conspiracy theories but the last one doesn't use future tense verbs that much like the others. More positive emotions showed deception for conspiracy theories and one of the research projects about fake reviews and fabricated, while more negative for rumors, misleading, biased/one-sided and one of the researches about fake reviews and fabricated. In general, the expression of emotions (positive or negative) showed deception for clickbait, fake reviews, trolling, biased/one-sided and fabricated. The lack of detailed, complex words and the big appearance of informal words showed deception for hoaxes, fake reviews, trolling and fabricated. Longer sentences with a big number of words show deception for trolling and hoaxes. More long words (with more syllables and/or characters) is a sign of deception for hoaxes, clickbait and fake review. More punctuation indicated disinformation for rumors, clickbait, trolling and fabricated but with the research that was done for trolling it discovered as well as the opposite (fewer quotation marks). The hyperbole use of slang phrases signifies deception for clickbait and fabricated. Conspiracy theories and fake reviews use more third and first-person, though for fake reviews there are researches that support the opposite. Fake reviews and trolls agree with the lack of conjunctions in their texts. Fake reviews and fabricated used to contain fewer paragraphs. It seems that all the disinformation types have similarities but especially fake reviews with conspiracy theories, fabricated, trolling and clickbait, also fabricated with clickbait.

As for the tense, trolls use fewer present than the others. The use of causation words is a sign of disinformation for trolling, but for conspiracy theories is the opposite. Variety of discrepancy words shows deception for fake reviews when the shortage of these words shows the same for conspiracy theories. Exactly the same is happening for tentative

words between these two types of disinformation. There is a lack of compatibility between fake reviews and trolls in the case of general words; the use of more general words translates as deception in fake reviews, while less using in trolls gives the same. The same about these two types is happening for function words. Fake reviews tend to have shorter sentences, but trolls longer sentences. Another difference between fake reviews and trolling is about the existence of pronouns, which is more in fake reviews and fabricated and less in trolling. Moreover, fake reviews tend to give more modal verbs, while trolls have fewer. About the number of words, it seems that trolling and fabricated disagree ultimately and that means that trolling has fewer words per paragraph (stories) while fabricated has more words in headlines but less in general and so do trolls. Conspiracy theories and fabricated contain more swear words. It is obvious that trolling and fake reviews have remarkable differences and this is logical because fake reviewers pretend to be true, but trolls show that they are fake. There are few differences between again trolling and conspiracy theories and also fabricated, while fabricated has distinction with conspiracy theories too.

About *conspiracy theories*, we found 2 papers with linguistic features from the SLR, which were published the same year (2021). Most features are sentiment-based and define psychological processes. Both data are drowned from Twitter and especially in the paper of Gerts et al. (2021) information belongs to the health domain. The detection models that have been used for conspiracy theories detection were CNN, ConspiDetector (which gave 74% accuracy) and Sentiment Analysis with AFINN and NRC (no accuracy percent available). These two sources agree on negative sentiment words like swear words.

As for *hoaxes*, we find one and only paper that refers to specific linguistic features and all the features (syntactical, lexical) showed low complexity of the text. In the same way researches about misleading connection are few. In our case the only paper we find was from 2009. The linguistic features are grammatical and sentiment-based and were detected with MANOVA Statistican method via people personal stories.

The researches that showed particular linguistic features that detect *rumors* were two, one from 2013 with the use of LIWC (Sentiment Analysis) and the other one, more recent, from 2021 with machine learning techniques use. Both of them drew data from Twitter and found sentiment-based linguistics, but the one is more positive oriented and the other one more negative.

Clickbait also is referred to 2 papers from our extensive research with the recent one (2021) using the FakeFlow detection model with 85% and the other using machine learning WEKA with 99,40% accuracy, which is the highest percentage from all the researches of this thesis. The two pieces of research agree only in sentiment-based linguistics and especially in the hyperbole of phrases. Also, both drew their data from news articles, but 2018 one especially from social media like Facebook, Forex and Reddit.

Most of the research material about linguistic features in specific types of disinformation was about *fake reviews*, 7 different papers and that is the reason there are many differences inside the fake reviews detection researches. For example, Banerjee et al. (2017) disagree with the plethora of positive emotion and filler words is deceptive reviews. Also, about the previous research, we noticed with surprise that almost the same authors in 2017 disagree with their 2014 research about the adjectives, in 2017 more adjectives are a deception sign, but in 2014 fewer adjectives give the same results. In this case, we are sticking with their 2017 results, which are the newest. While all researchers agree with more first and third pronouns as deceptive reviews are Ansari et al. (2021) have exact different opinions. Most linguitic features are sentiment-based and define psychological processes. In the research of deception in reviews most popular technique used was linguistic-based and the model was LWIC. Although, the biggest accuracy gave another linguistic model, Coh Metrix (89.70%). The data for the researches that were used for deception detection was mainly from the hospitality domain, hotel reviews.

The surveys that were found about *trolling* were two, one from 2019 with linguistic cues analysis and 82% accuracy score and the other from 2020 with three different learning methods and comparative performance. Our research about linguistic features in trolling found more grammatical and lexical features that shows complexity and belong to the social media (Twitter) domain. Moreover, this type of disinformative text contains digital features like hashtags, links and URLs. The two papers for trolling have different opinions about the number of words. Addawood et al. (2019) insist that more words show deception, but Miao et al. (2020) claim the opposite i.e., fewer words are for trolling. Also, it is noteworthy that in research of 2019 findings show that more punctuation demonstrates deception and especially question marks, but in the same paper, fewer quotations (which are part of punctuation) also show deception.

For the *fabricated* type of disinformation, we found only one paper the only with AI Techniques and NLP (natural language processing) model, but there isn't an accuracy percent available. The domain of the data taken is from Politics (news sites). In this type of disinformation, we meet all the categories of linguistic features, which show mainly complexity (lexical, syntactical), uncertainty (grammatical) and psychological processes (sentiment-based). In this research, the features differentiate in stories and headlines so fewer words for the stories demonstrate deception and more words in headlines show the same.

For *biased/ one-sided* news the results of the SLR lead us to one recent paper (2021) that declares that this type of disinformation comes with emotionally driven language which is especially negative and this is, of course, sentiment-based linguistics and define psychological processes. The researchers' pieces of information come from linguistic analysis in Twitter again and the tweets were about politics. Unfortunately, accuracy percent is not available.

For the differentiation of the disinformation types, we base our research on Kapantai et al. (2021) paper so we have to say that unfortunately, we could not find research worthy material about *Pseudoscience*. In future research one of the first things to do is to make extended research only about pseudoscience and linguistic features that show deception.

8 Conclusion

In the systematic literature review for the use of linguistic features and the systems that use them is found that we have positive results in detection of fake news, deceivers, conspiracy propagators, fabricated news, trolls, etc. The automatic systems in many times outperform the manual, human way to detect the deceiving. It is believed that the best way is the combination of both when it is possible. The need for constant alert and in time, revealing the deception, is urgent in the era of continuous and enlarging battle of truth and lie.

The separation of fake and true through linguistic features is an ocean of information. However, the division into types of disinformation (rumors, clickbait, etc) makes the task of detection tools easier, but also of people who want to detect fraud in different types. For example, it is important for marketing professionals to know if the reviews for their business are fake in order to be treated properly.

There is of course the possibility and the risk that the present research will be studied by individuals or organizations that systematically and professionally create fake content so knowing the elements that suggest each type of disinformation to adapt their texts to look real and achieve the goal (e.g., to fight their competition) with minimum effort and the result is quite accurate.

Something that is more important than finding correspondence with the types of disinformation is finding the source of deception because the importance of fake news created by a specific organization that has opposite political or economic interests is different from dealing with a misguided citizen with ideologies, that is, with distorted reality due to perceptions. Also, it is another thing for an organization to deceive a company and this would have a legal interest because if it is found from the source that a competitor misinforms the business public, the company can take legal action against it. In addition, we may simply be dealing with a malicious citizen who is having fun causing problems without thinking about the consequences of his actions.

Through our research we understand the depth of this issue, information reaches the point of drowning in an ocean of lies or half-truths. We are in the age of information

and we are about to cancel it completely. If this data is involved with AI tools and is ultimately wrong then disaster can come as well, because rubbish only rubbish can give. That's why we have to check the sources thoroughly to make sure that the information comes from a reliable scientific body.

More important is the way you convey something, but even more crucial is the content. If the content is random and unconvincing, we can conclude that it is not something organized. In this database companies like Google, which keeps records of everything (and this is its most basic function), can easily with a flashback to find out who put forward an argument, who first claimed something fake and find out by whom or which group started something.

Distinguishing true from fake is not an easy story and to achieve it effectively you need all the above mentioning means. In addition, someone can conclude from this analysis that when someone enters such data can be characterized.

9 Limitations and Future work

During our research in various scientific databases for relevant with this topic documents, it was sure clearer that there is a research gap between disinformation and linguistic features. Information that we could gather was more general about linguistic features that we can meet in fake news. It was difficult to find material for all the different types of disinformation. As the compass for the division of the different types of disinformation was the paper of Kapantai et al. (2021) the original plan was to find linguistic features that define all these categories, but we were unable to find data for Imposter, which can be explained because disinformation is not related to the content of information only by definition. Furthermore, we could not locate noteworthy scientific studies on the fascinating issue of Pseudoscience and the distinctive linguistic traits we observe in this type of deception. So, one of the first things to accomplish in future research is to focus on Pseudoscience and linguistic features that indicate deception.

It has been discussed that a questionnaire will be an interesting idea as a validation process of our SLR results that are contained summarized in Table 2. Our questionnaire survey will include this table and some examples of deceptive text, one of each disinformation category (fake review, rumor, clickbait, etc) without being named so participants will categorize each paragraph according to the linguistics features that they notice based on our research results. In this way, we can verify our results in order to apply our model in further scientific research by using linguistic features to detect disinformation in specific types. Unfortunately, in this case, the creation of questionnaires was not possible due to the minimal time we had at our disposal for such a large topic and in order to realize this survey properly the sample we would need would be very large. We are very willing to continue our research in the future in this direction, as we now have a strong basis that was difficult to establish. Therefore, given the great literature research that has been done, we will start directly in a more practical way and focus on the survey.

In addition, during the reviewing of our bibliography we discovered an interesting keyword, the “style-based approach AND fake news”, but from the beginning, our research

was accomplished with specific keywords and that is why this wasn't included. In future work about the same topic, we will definitely add this phrase in our research keywords which will probably give us engaging content as a result.

Last but not least as we mention above, we think that more crucial is the creation of a model that would find in some way the source of disinformation by using for example multiple meta-data about the news source and author, like social media information dissemination aspects and by using Deep Learning methods with larger datasets as Gravanis et al. (2019) proposes. Giachanou et al (2021) about future research found it fascinating to look at the profile of conspiracy theorists in different nations using the geolocation data accessible in tweets and look at how their findings might be applied to increase the efficacy of false news detection systems and intervention techniques. In this way, the duty of detecting fake news would not be based only on the content, but it could also increase the blockage of their spread through social networks.

10 Bibliography

Addawood, Aseel, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. “Linguistic Cues to Deception: Identifying Political Trolls on Social Media.” Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019 (Icwsm):15–25.

Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. 2012. “Detecting Hoaxes, Frauds, and Deception in Writing Style Online.” Proceedings - IEEE Symposium on Security and Privacy 461–75. doi: 10.1109/SP.2012.34.

Aldwairi, Monther, and Ali Alwahedi. 2018. “Detecting Fake News in Social Media Networks.” Procedia Computer Science 141:215–22. doi: 10.1016/j.procs.2018.10.171.

Aliev, Rafik Aziz, Janusz Kacprzyk, and Witold Pedrycz. 2020. Advances in Intelligent Systems and Computing 1323 11th World Conference “ Intelligent System for Industrial Automation .”

Ansari, Sana, and Sumeet Gupta. 2021. “Customer Perception of the Deceptiveness of Online Product Reviews: A Speech Act Theory Perspective.” International Journal of Information Management 57(November 2020):102286. doi: 10.1016/j.ijinfomgt.2020.102286.

Asubiaro, Toluwase Victor, and Victoria L. Rubin. 2018. “Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017).” Proceedings of the Association for Information Science and Technology 55(1):747–50. doi: 10.1002/pra2.2018.14505501100.

Banerjee, Snehasish, and Alton Y. K. Chua. 2014. “A Linguistic Framework to Distinguish between Genuine and Deceptive Online Reviews.” Lecture Notes in Engineering and Computer Science 2209(January):501–6.

Banerjee, Snehasish, Alton Y. K. Chua, and Jung Jae Kim. 2015. “Using Supervised Learning to Classify Authentic and Fake Online Reviews.” ACM IMCOM 2015 - Proceedings. doi: 10.1145/2701126.2701130.

Choudhary, Anshika, and Anuja Arora. 2021. “Linguistic Feature Based Learning Model for Fake News Detection and Classification.” Expert Systems with Applications 169(August 2020):114171. doi: 10.1016/j.eswa.2020.114171.

Clarke, Jonathan, Hailiang Chen, Ding Du, and Yu Jeffrey Hu. 2021. “Fake News, Investor Attention, and Market Reaction.” Information Systems Research 32(1):35–52. doi: 10.1287/isre.2019.0910.

Damstra, Alyt, Hajo G. Boomgaarden, Elena Broda, Elina Lindgren, Jesper Strömbäck, Yariv Tsfati, and Rens Vliegenthart. 2021. “What Does Fake Look Like? A Review of the Literature on Intentional Deception in the News and on Social Media.” Journalism Studies 0(0):1–17. doi: 10.1080/1461670X.2021.1979423.

De Beer, Dylan & Matthee, Machdel. (2021). Approaches to Identify Fake News: A Systematic Literature Review. 10.1007/978-3-030-49264-9_2.

De Pablo, Alvaro, Oscar Araque, and Carlos A. Iglesias. 2020. “Radical Text Detection Based on Stylometry.” ICISSP 2020 - Proceedings of the 6th International Conference on Information Systems Security and Privacy 524–31. doi: 10.5220/0008971205240531.

Deng, Qi, Yun Wang, Michel Rod, and Shaobo Ji. 2021. “Speak to Head and Heart: The Effects of Linguistic Features on B2B Brand Engagement on Social Media.” Industrial Marketing Management 99(July):1–15. doi: 10.1016/j.indmarman.2021.09.005.

Dilmon, Rakefet. 2009. “Between Thinking and Speaking-Linguistic Tools for Detecting a Fabrication.” Journal of Pragmatics 41(6):1152–70. doi: 10.1016/j.pragma.2008.09.032.

Dwivedi, Sanjeev M., and Sunil B. Wankhade. 2021. “Survey on Fake News Detection Techniques.” Advances in Intelligent Systems and Computing 1200 AISC:342–48.

Etaiwi, Wael, and Arafat Awajan. 2017. “The Effects of Features Selection Methods on Spam Review Detection Performance.” Proceedings - 2017 International Conference on New Trends in Computing Sciences, ICTCS 2017 2018-Janua(2):116–20. doi: 10.1109/ICTCS.2017.50.

Fallis, Don. 2015. “What Is Disinformation?” Library Trends 63(3):401–26. doi: 10.1353/lib.2015.0014.

Gerts, Dax, Courtney D. Shelley, Nidhi Parikh, Travis Pitts, Chrysm Watson Ross, Geoffrey Fairchild, Nidia Yadria Vaquera Chavez, and Ashlynn R. Daughton. 2021.

"'Thought I'd Share First' and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study." JMIR Public Health and Surveillance 7(4):1–17. doi: 10.2196/26527.

Ghanem, Bilal, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. "FakeFlow: Fake News Detection by Modeling the Flow of Affective Information." EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference 679–89. doi: 10.18653/v1/2021.eacl-main.56.

Giachanou, Anastasia, Bilal Ghanem, and Paolo Rosso. 2021. "Detection of Conspiracy Propagators Using Psycho-Linguistic Characteristics." Journal of Information Science. doi: 10.1177/0165551520985486.

Gravanis, Georgios, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. "Behind the Cues: A Benchmarking Study for Fake News Detection." Expert Systems with Applications 128:201–13. doi: 10.1016/j.eswa.2019.03.036.

GYONGYI, Z., H. GARCIAMOLINA, and J. PEDERSEN. 2004. "Combating Web Spam with TrustRank." Proceedings 2004 VLDB Conference 576–87. doi: 10.1016/b978-012088469-8/50052-8.

Huseynova, F. (2021). Analysis of Influence of Imagination, Fantasy, Exaggeration, and Hoax on a Level of Lie Under Linguistic Information.

Jachim, Peter, Filipo Sharevski, and Paige Treebridge. 2020. "TrollHunter [Evader]: Automated Detection [Evasion] of Twitter Trolls during the COVID-19 Pandemic." PervasiveHealth: Pervasive Computing Technologies for Healthcare 59–75. doi: 10.1145/3442167.3442169.

Jiang, Shan, and Christo Wilson. 2018. "Linguistic Signals under Misinformation and Fact-Checking." Proceedings of the ACM on Human-Computer Interaction 2(CSCW):1–23. doi: 10.1145/3274351.

Kapantai, Eleni, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2021. "A Systematic Literature Review on Disinformation: Toward a Unified Taxonomical Framework." New Media and Society 23(5):1301–26. doi: 10.1177/1461444820959296.

Kasseropoulos and Tjortjis. (2021). "An Approach Utilizing Linguistic Features for Fake News Detection. In: Maglogiannis I., Macintyre J., Iliadis L. (eds) Artificial

Intelligence Applications and Innovations." AIAI 2021. IFIP Advances in Information and Communication Technology, vol 627. Springer, Cham. doi: 10.1007/978-3-030-79150-6_51

Kennedy, Stefan, Niall Walsh, Kirils Sloka, Andrew McCarren, and Jennifer Foster. 2019. "Fact or Factitious? Contextualized Opinion Spam Detection." ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop 344–50. doi: 10.18653/v1/p19-2048.

Khan, Tanveer, Antonis Michalas, and Adnan Akhunzada. 2021. "Fake News Outbreak 2021: Can We Stop the Viral Spread?" Journal of Network and Computer Applications 190(June):103112. doi: 10.1016/j.jnca.2021.103112.

Kronrod, Ann, Jeffrey K. Lee, and Ivan Gordeliy. 2017. "Detecting Fictitious Consumer Reviews: A Theory-Driven Approach Combining Automated Text Analysis and Experimental Design." Marketing Science Institute Working Paper Series, Report No. 17-124 (17):1–61.

Kumar, Akshi, M. P. S. Bhatia, and Saurabh Raj Sangwan. 2021. "Rumour Detection Using Deep Learning and Filter-Wrapper Feature Selection in Benchmark Twitter Dataset." Multimedia Tools and Applications (0123456789). doi: 10.1007/s11042-021-11340-x.

Kwon, Sejeong, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. "Aspects of Rumor Spreading on a Microblog Network." Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8238 LNCS:299–308. doi: 10.1007/978-3-319-03260-3_26.

Lahlou, Yasmine, Sanaa El Fkihi, and Rdouan Faizi. 2019. "Automatic Detection of Fake News on Online Platforms: A Survey." ICSSD 2019 - International Conference on Smart Systems and Data Science 3–6. doi: 10.1109/ICSSD47982.2019.9002823.

Lakhtionova, Liudmyla, Nataliia Muranova, Oleksandr Bugaiov, Alla Ozeran, and Svitlana Kalabukhova. 2021. Balance Sheet (Statement of Financial Position) Transformation in the Light of New Digital Technology: Ukrainian Experience. Vol. 136.

Li, Huayi, Zhiyuan Chen, Bing Liu, Xiaokai Wei, and Jidong Shao. 2014. "Spotting Fake Reviews via Collective Positive-Unlabeled Learning." Proceedings - IEEE International Conference on Data Mining, ICDM 2015-Janua(January):899–904. doi: 10.1109/ICDM.2014.47.

Li, Jianing, and Min Hsin Su. 2020. “Real Talk About Fake News: Identity Language and Disconnected Networks of the US Public’s ‘Fake News’ Discourse on Twitter.” Social Media and Society 6(2). doi: 10.1177/2056305120916841.

Li, Lin, Kyung Young Lee, Minwoo Lee, and Sung Byung Yang. 2020. “Unveiling the Cloak of Deviance: Linguistic Cues for Psychological Processes in Fake Online Reviews.” International Journal of Hospitality Management 87(September 2019):102468. doi: 10.1016/j.ijhm.2020.102468.

Liu, Xiaozhong. 2013. “Full-Text Citation Analysis : A New Method to Enhance.” Journal of the American Society for Information Science and Technology 64(July):1852–63. doi: 10.1002/asi.

Long, Si Hong, and Mohd Pouzi Bin Hamzah. 2021. “Fake News Detection.” Lecture Notes in Electrical Engineering 724(2):295–303. doi: 10.1007/978-981-33-4069-5_25.

Lotfi, Serveh, Mitra Mirzarezaee, Mehdi Hosseinzadeh, and Vahid Seydi. 2021. “Rumor Conversations Detection in Twitter through Extraction of Structural Features.” Information Technology and Management 22(4):265–79. doi: 10.1007/s10799-021-00335-7.

Mahyoob, Mohammad, Jeehaan Algaraady, and Musaad Alrahaili. 2020. “Linguistic-Based Detection of Fake News in Social Media.” International Journal of English Linguistics 11(1):99. doi: 10.5539/ijel.v11n1p99.

Meibauer, Jörg. 2018. “The Linguistics of Lying.” Annual Review of Linguistics 4:357–75. doi: 10.1146/annurev-linguistics-011817-045634.

Meinert, Judith, Milad Mirbabaie, Sebastian Dungs, and Ahmet Aker. 2018. “Is It Really Fake? – Towards an Understanding of Fake News in Social Media Communication.” Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10913 LNCS:484–97. doi: 10.1007/978-3-319-91521-0_35.

Memon, Shahan Ali, and Kathleen M. Carley. 2020. “Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset.” CEUR Workshop Proceedings 2699.

Miao, Lin, Mark Last, and Marina Litvak. 2020. “Detecting Troll Tweets in a Bilingual Corpus.” LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings (May):6247–54.

Moon, Sangkil, Moon Yong Kim, and Dawn Iacobucci. 2021. “Content Analysis of Fake Consumer Reviews by Survey-Based Text Categorization.” International Journal of Research in Marketing 38(2):343–64. doi: 10.1016/j.ijresmar.2020.08.001.

Mu, Yida, and Nikolaos Aletras. 2020. “Identifying Twitter Users Who Repost Unreliable News Sources with Linguistic Information.” PeerJ Computer Science 6:1–18. doi: 10.7717/peerj-cs.325.

Mukherjee, Arjun, Vivek Venkataraman, … B. Liu-Seventh international AAAI, and Undefined 2013. 2011. “What Yelp Fake Review Filter Might Be Doing?” Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media 409–18.

Murphy, Julie, Anthony Keane, and Aurelia Power. 2020. “Computational Propaganda: Targeted Advertising and the Perception of Truth.” European Conference on Information Warfare and Security, ECCWS 2020-June:491–500. doi: 10.34190/EWS.20.503.

Neisari, Ashraf, Luis Rueda, and Sherif Saad. 2021. “Spam Review Detection Using Self-Organizing Maps and Convolutional Neural Networks.” Computers and Security 106:102274. doi: 10.1016/j.cose.2021.102274.

Ott, Myle, Claire Cardie, and Jeffrey T. Hancock. 2013. “Negative Deceptive Opinion Spam.” NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference (June):497–501.

Pennycook, Gordon, and David G. Rand. 2019. “Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality.” Proceedings of the National Academy of Sciences of the United States of America 116(7):2521–26. doi: 10.1073/pnas.1806781116.

Plotkina, Daria, Andreas Munzel, and Jessie Pallud. 2020. “Illusions of Truth—Experimental Insights into Human and Algorithmic Detections of Fake Online Reviews.” Journal of Business Research 109(December):511–23. doi: 10.1016/j.jbusres.2018.12.009.

Qazvinian, Vahed, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. “Qazvinian et Al. - 2011 - Rumor Has It Identifying Misinformation in Microblogs(2).” Conference on Empirical Methods in Natural Language Processing 1589–99.

Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. “Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-

Checking.” EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings 2931–37. doi: 10.18653/v1/d17-1317.

Sepehri, Amir, David M. Markowitz, and Rod Duclos. 2021. “The Location of Maximum Emotion in Deceptive and Truthful Texts.” Social Psychological and Personality Science 12(6):996–1004. doi: 10.1177/1948550620949730.

Shabani, Shaban, and Maria Sokhn. 2018. “Hybrid Machine-Crowd Approach for Fake News Detection.” Proceedings - 4th IEEE International Conference on Collaboration and Internet Computing, CIC 2018 299–306. doi: 10.1109/CIC.2018.00048.

Shevchenko, Larysa, Dmytro Syzonov, Olga Pliasun, and Volodymyr Shmatko. 2021. “Media Literacy Research during COVID-19 Pandemic: Social Network Screening.” International Journal of Media and Information Literacy 6(1):219–30. doi: 10.13187/IJMIL.2021.1.219.

Shrestha and Spezzano F. 2021. “Textual Characteristics of News Title and Body to Detect Fake News: A Reproducibility Study.” In: Hiemstra D., Moens MF., Mothe J., Perego R., Potthast M., Sebastiani F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12657. Springer, Cham. doi: 10.1007/978-3-030-72240-1_9.

Thu, Pyae Phyo, and Than Nwe Aung. 2018. “Implementation of Emotional Features on Satire Detection.” International Journal of Networked and Distributed Computing 6(2):78–87. doi: 10.2991/ijndc.2018.6.2.3.

Verma, Pawan Kumar, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. “WELFake: Word Embedding over Linguistic Features for Fake News Detection.” IEEE Transactions on Computational Social Systems 8(4):881–93. doi: 10.1109/TCSS.2021.3068519.

Visentin, Marco, Annamaria Tuan, and Giandomenico Di Domenico. 2021. “Words Matter : How Privacy Concerns and Conspiracy Theories Spread on Twitter.” (June). doi: 10.1002/mar.21542.

Volkova, Svitlana, and Jin Yea Jang. 2018. “Misleading or Falsification: Inferring Deceptive Strategies and Types in Online News and Social Media.” The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018 575–83. doi: 10.1145/3184558.3188728.

Volkova, Svitlana, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. “Linguistic Model for Fake News on Twitter.” 647–53.

Vosoughi, Soroush, Mostafa ‘Neo’ Mohsenvand, and Deb Roy. 2017. “Rumor Gauge: Predicting the Veracity of Rumors on Twitter.” ACM Transactions on Knowledge Discovery from Data 11(4). doi: 10.1145/3070644.

Wang, Erin Yirun, Lawrence Hoc Nang Fong, and Rob Law. 2021. “Detecting Fake Hospitality Reviews through the Interplay of Emotional Cues, Cognitive Cues and Review Valence.” International Journal of Contemporary Hospitality Management. doi: 10.1108/IJCHM-04-2021-0473.

Wu, Lianwei, Yuan Rao, Cong Zhang, Yongqiang Zhao, and Ambreen Nazir. 2021. “Category-Controlled Encoder-Decoder for Fake News Detection.” IEEE Transactions on Knowledge and Data Engineering (November). doi: 10.1109/TKDE.2021.3103833.

Ye, Junting, and Steven Skiena. 2019. “MediaRank: Computational Ranking of Online News Sources.” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2469–77. doi: 10.1145/3292500.3330751.

Younus, Arjumand, and M. Atif Qureshi. 2020. “Combining BERT with Contextual Linguistic Features for Identification of Propaganda Spans in News Articles.” Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020 5864–66. doi: 10.1109/BigData50022.2020.9378432.

Zannettou, Savvas, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. “The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans.” Journal of Data and Information Quality 11(3):1–26. doi: 10.1145/3309699.

Zhou, Cheng, Kai Li, and Yanhong Lu. 2021. “Linguistic Characteristics and the Dissemination of Misinformation in Social Media: The Moderating Effect of Information Richness.” Information Processing and Management 58(6):102679. doi: 10.1016/j.ipm.2021.102679.

Zhou, Zhixuan, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. “Fake News Detection via NLP Is Vulnerable to Adversarial Attacks.” ICAART 2019 - Proceedings of the 11th International Conference on Agents and Artificial Intelligence 2:794–800.doi:10.5220/00075663079408.