MEDIATIZATION STUDIES 3/2019 DOI: 10.17951/ms.2019.3.95-114

DOROTA MARQUARDT

University of Economics in Katowice. Faculty of Informatics and Communication. Department of Communication Design and Analysis, Poland

dorota.marquardt@uekat.pl

ORCID: https://orcid.org/0000-0001-7434-2415

Linguistic Indicators in the Identification of Fake News

Abstract. The issue of fake news identification was approached from the corpus linguistics and discursive studies perspective. The texts of both actual and fake news have been analysed in search of dependences that would permit the increase of the ability to determine the probability of the given news being real or fake, taking into account the discursive characteristics of the particular texts.

Keywords: fake news; lie; discourse; genre; trust

1. Introduction

Spreading false or manipulated information is not a new phenomenon. It has its own, long history, however, it became truly common in the world of the half-truth. Internet media, the change in the manner of functioning of the media previously defined as traditional as well as the way in which politics and business is conducted made fake news part of our everyday lives. It is used both as disinformation and in political or media conflicts by means of rendering the inconvenient information untrue. Therefore, the following questions are worth posing: is there any way to identify such false news? If so, what mechanisms should one focus on? Are there any universal mechanisms?

The article focuses mainly on the identification of fake news, drawing the attention to the elements associated with the construction of such information, the used language (on the semantic, syntactic, pragmatic and discursive level) as well as the issues of trust towards the information source and the manner in which the information is distributed.

The research was divided into two fundamental parts. In the first one, the corpora of true and false news underwent a qualitative-quantitative comparative analysis. As the result of this analysis, the language dependencies, differences and similarities associated with the sentiment of the text, the sentence construction and the issues of the utterance

construction, including the text complexity level, were indicated. In the second part, a model was created that permitted the increase of the probability of recognizing the information as true or false on the basis of language indicators and the credibility of the source.

2. The state of research on fake news

2.1. The state of research in linguistics and IT

The research on false information flooding the contemporary media are conducted on many levels and from the viewpoints of multiple branches of science. In this article, the focus is placed mostly on the linguistic and IT aspects (with the elements of media studies[1]).

The amount of direct linguistic research on fake news is low (Newman et al. 2003). However, if we approach the issue in a broader sense, it transpires that false information appearing in the media is connected with the subject of lies in the language (Antas 2008). This articles utilizes the most important findings from this area. One should also keep in mind that all of the research results obtained on the basis of an English material can be directly translated into results for the Polish language.

The research papers on lies in the language focus on the construction of a sentence (a lie is characterized by more complex syntax structures with a lower amount of conveyed information) (Antas 2008). The studies on deception in the language show that the stories on a given event feature significantly less "I" statements and fewer references to the authors' own experiences in the fake news. However, these studies were associated with an individual's statement on their own action, whereas the gathered corpus is connected with the coverage and conveying of information that the author was not part of, and the news form excludes the use of first person singular statements. Moreover, the research shows that false texts were dominated by expressions of a negative sentiment (Newman et al. 2003; Hancock, Toma, Ellison 2007; Enos et al. 2015). Lying in language can also be approached from the point of view of the theory of speech acts (Austin 1961). John L. Austin distinguishes the truthfulness of a reference from its effectiveness (denotation from connotation), points to the problem of pretending to be a speech act and its illocution. Media researchers and cognitivists also draw attention to illocution as an important issue. Leonard Shedletsky emphasizes the functioning of disinformation ("bullshit" phenomenon) by writing that in order to explain this phenomenon it is necessary to focus on the intent or character of the speaker, on the audience (their values and beliefs) and on the text itself (Shedletsky 2018). When studying fake news, the computer scientists focus mostly on: the way the false information is distributed,

[1] Of course, this does not mean that the other approaches – especially educational, political science and sociological – are underappreciated.

the data on the sources of information, the combinations of various data that aid in the identification of fake news. It has been determined that false information spreads 6 times faster than the truth on Twitter (Vosoghi, Roy, Aral 2018). The IT specialists attempt to automate the process of fake news identification, study the manner of its distribution and test new solutions along with their efficiency (Conroy, Rubin, Chen 2015; Brewer, Goldthwaite Young, Morreale 2013).

2.2. Definitions of fake news

The English-language literature features, apart from the notion of fake news, terms such as "misleading information" (sometimes used interchangeably with the idea of misinformation) and "deception". Misinformation is defined as "false or inaccurate information, especially that which is deliberately intended to deceive" (Kumar, Geethakumari 2014). Deceptive news, on the other hand, is often divided into serious fabrications (untrue tabloid information), large-scale hoaxes (false information which leads to serious consequences for other parties – loss of property, etc.) and humorous fakes (false information aiming to amuse and not to mislead us) (Rubin, Chen, Conroy 2015). Fake news, however, is the most unclear notion. Some use it in relation to any information that is not entirely true[2] or a piece of information that can be swiftly verified (containing verifiable information, e.g. legal, financial, historical matters, etc.).

This article assumes the definition of fake news as untrue information, which can be verified (therefore, the definition does not include opinions, comments, etc.) which intends to mislead (regardless whether the misleading was to have significant consequences for the society, any group or person).

3. Methodology and the research material

3.1. Research material

Two text corpora were utilized during the analysis. The first one is the corpus of real news – 30 various news published by the Polish Press Agency (PAP – *Polska Agencja Prasowa*) was gathered, assuming that the probability of the news placed there being false is lower than in case of other publishers. The news pieces were gathered between 15 and 17 September 2018.

The second corpus is a collection of 30 fake news prepared by journalism and social communication students during their workshops on Internet journalism[3] in

[2] This is where the doubts associated with the definition of truth often appear.

[3] The students of journalism and social communication of the University of Economics in Katowice have prepared real and fake Internet news and then compared the manner in which

the 2017/2018 academic year in compliance with the rules of creating information published in the Internet.

3.2. Research proceedings

The analysis begins with the quantitative research of the real and fake texts. The sentiment of the article is defined (whether there are more positive or negative statements, what are the dominant values for the given article), a frequency list of the words in the given articles is created (on the basis of which the fundamental interpretative framework is established) and the amount of words in the sentences, the noun-to-verb ratio and the text complexity level (using the FOG index) are determined.

The results of the quantitative analyses become the basis of the qualitative analysis. What will be verified is the coherence of the interpretative framework and the convergence of the text and the title (Chopra, Jain, Sholar 2017). Moreover, the discursive analysis will also be conducted.

Next, the analysis will be complimented with conclusions on the sources of information and the authors of the given texts (Twitter entries), which were taken from the already conducted Twitter fake news research.

3.3. Methodology

The entirety of the qualitative analysis will be fitted into the cognitive paradigm. The interpretative framework theories evolving from the semantic framework will be used. The discursive analysis will also refer to the cognitive assumptions. The quantitative research will be conducted with the aid of the software created as part of the Clarin (clarin-pl.eu) project, the author's own software as well as the Jasnopis software.

4. Research results

4.1. Quantitative analysis

4.1.1. Keywords

Keywords have been determined for each article, which allowed to establish the article's subject and which will form the base of the interpretative framework re-creation. The keywords have been divided into two groups: verbs, which form the framework centre, and nouns, adjectives and other parts of speech which may play the role of

such information was constructed. They indicated the areas one should focus on when receiving information from the media to avoid considering the false information as real/true one.

arguments in the sentences created around the verbs. The noun-to-verb ratio as well as the mean number of words in a sentence have also been determined for each of the articles. The results of the analysis are available in Tables 1 and 2.

Table 1. Keywords in the true news corpus (ws.clarin-pl.eu/tfidf.shtml)

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"The Bialystok Commemoration Peloton on the anniversary of Soviet aggression on Poland"	Białystok (9), memory (7), Soviet (6), Poland (6), aggression (5)	be (6), connect (2), have (2), stop (2)	4.5	23
"At least one thousand people participated in the pro-European demonstration in Budapest"	European (4), Orbán (7), resolution (4)	be (2), call (2), accept (4)	3.34	18
"Persistence in the ecological development bears fruit in Suining, China"	ecological (10), city (9), Suining (9), development (7), China (6)	be (3), take place (2)	7.29	22
"The president thanked the farmers for their efforts during the harvest festival in Spała"	Polish (22), president (11), Poland (10), farmer (10)	be (13), thank (4), highlight (3), thank (3)[4], develop (3)	2.28	22
"The Eurovision song contest to be held in Tel Aviv next year"	contest (9), Eurovision (5), song (5), Israel (4)	announce (2), be (2)	6.47	22
"USA: the number of deaths following the tropical storm in Florence reaches 17"	state (6), Florence (5), north (5)	be (5), flood (2), cause (2), inform (2), evacuate (2)	3.64	23
"Polish jazz in Asia on the 100th anniversary of the Polish independence restoration"	Polish (11), program (9), jazz (7), China (6)	be (7), perform (3), play (3)	5.66	25
"32 rebels and 4 radio employees died in Yemen"	rebel (6), Hudajda (4), Yemen (4), coalition (4)	die (5), support (2), be (2)	5.96	26

[4] In the Polish language, two different words exist within the article (*dziękować* and *podziękować*) – they share the stem and are of similar meaning. However, they open different valence spots and create different sentence schemes.

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"Kaczyński: strong West Pomerania is in the Polish national interest"	Kaczyński (5), strong (4), very (4), Germany (4)	be (8), highlight (3), state (2), know (2), speak (2)	1.76	19
"Canadian Prime Minister on the Polish festival in Toronto: diversity is the source of strength"	Polish (20), Canada (14), festival (11), Poland (10), Toronto (10), Prime minister (9)	be (26), say (8), celebrate (5), highlight (3)	3.16	22
"Cardinal Bagnasco: Europe should follow the path towards unity"	Europe (13), unity (8), church (6), Bagnasco (5)	be (19), follow (3), say (3), highlight (3)	2.96	22
"Simon Yates triumphs in the Vuelta a Espana race"	stage (7), race (7), Spain (6)	be (8), win (3), win (2)[5], finish (2), bet (2), be victorious (2)	7.86	12
"Kornel Morawiecki: WiS to register the lists of regional council candidates in all voivodships"	Morawiecki (5), list (5), all (5)	register (4), add (3), search (3), want (2), say (2)	3.2	13
"Kremlin spokesperson: the individuals suspected of attacking Skripal are not connected with Putin"	Skripal (6), British (6), Russian (5), attack (5)	be (4), become (3), announce (3)	3.02	19
"Team from the Kielce University of Technology wins the Martian rover competition"	competition (11), team (10), group (9), rover (7), science (6)	be (11), prepare (4), become (4), highlight (4), must (3)	3.68	17
"Macedonia: several thousand participate in the demonstration supporting the EU and the NATO"	EU (5), NATO (5), country (5), name (5), Skopje (4), agreement (4)	appeal (2), conclude (2)	4.32	17
"The head of the Ministry of National Defence meets the World War II heroes in New York"	MoND (6), world (4), head (4), hero (4), American (4), war (4), Poland (4), meeting (4)	meet (3), inform (2), forget (2), fight (2), say (2)	3.56	22

[5] In the Polish language one of the words occurred in the perfect aspect, whereas the second one in the imperfect aspect.

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"National Readout of *The Spring to Come* by Stefan Żeromski begins"	readout (15), *The Spring to Come* (11), national (11), year (9), Duda (7), Żeromski (7), president (7), Poland (7)	be (15), say (5), speak (5), become (4), read (4)	3.91	19
"Prime Minister: it is the ruling camp that follows the constitution the most"	Constitution (6), work (5), ruling (5), Poland (4), camp (4), government (4), prime minister (4)	be (12), say (4), abide (3)	1.63	15
"Poland wins 3:2 with Romania in the Davis Cup"	match (6), Majchrzyk (5), Romanian (5), year (5), duel (4), Polish (4), Pole (4), game (4)	be (9), win (5), break (3), be victorious (2), beat (2), seal (2), remain (2)	3.28	20
"Experts: respiratory system diseases as the second most commons cause of death in the world"	lung (20), respiratory (20), disease (20), system (12), professor (9), death (8), health (8)	be (15), contribute (3), remind (3), highlight (3), conduct (2), say (2)	3.57	20
"Sasin: we never announced the 500 plus programme for pensioners"	pensioner (13), PiS (8), 500 plus (6), Sasin (6), Kaczyński (6), addition (5), minister (5)	be (21), say (8), announce (5), state (5)	1.68	15
"Head of SLD on the coalition with PO: I can talk about it, but on equal terms"	PiS (7), coalition (5), Czarzasty (5), election (4), SLD (4), party (4)	be (12), concern (4), talk (3)	1.89	15
"European community associated with head tumours joins forces to understand the rare cancer form"	neck (11), head (11), disease (9), cancer (9), patient (8)	can (4), find (3), be (3), be sick (2), survive (2), spread (2), find (2)[6]	8.22	23
"Syria accused Israel of conducting a missile attack on the Damascus airport"	Israeli (10), attack (10), Syria (9), Israel (7), target (5), missile (5)	accuse (2), be (2), use (2), conduct (2), involve (2), announce (2), arm (2), highlight (2)	3.64	20

[6] In the Polish language one of the words occurred in the perfect aspect, whereas the second one in the imperfect aspect.

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"Szczerski: an important Polish-American declaration – a potential effect of the USA's presidential visit"	president (12), visit (8), USA (6), American (6), Duda (5), Poland (5)	be (6), unify (3), move (2)	5.48	26
"Tymoteusz Bies as the best pianist of the Szymanowski International Music Competition"	Szymanowski (10), competition (8), pianist (6), music (6), Katowice (6), reward (5), Polish (5)	be (4), recognize (2)	6.77	17
"Trump: the Three Seas Initiative has great potential"	Three Seas (6), initiative (3), my (3), great (3), Trump (3), potential (3)	unify (1), highlight (1), organize (1), add (1), confirm (1)	6.67	40
"First local elections in years in Syria, candidates mostly from Assad's party"	vote (5), Assad (5), election (4), Damascus (4), Syria (4)	be (8), conduct (2)	4.16	24
"Turtle extinction may negatively influence the environment"	turtle (13), species (8), he (8)	be (10), may (6), decrease (2), spread (2), have (2), inhabit (2), threaten (2)	4.48	20

Source: Author's own study.

Table 2. Keywords in the fake news corpus

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"Beata Szydło attacked near Łazienki"	Szydło (5), Beata (3), hour (3), hotel (3)	become (3), want (3), attack (2), be (2), move (2)	1.97	12
"Biedronka shops open on trade Sundays!"	shop (8), Biedronka (7), open (5), own (4), Sunday (4), network (4)	be (14), announce (3), inform (2), attempt (2)	2.13	13
"Painful loss in the life of Sławomir"	one (3), Sławomir (3)	be (4), become (2), sell (2)	2.87	15
"Borys Szyc passed away"	apartment (4), hotel (4), hour (3), police (3), man (3), letter (3), Borys (3)	be (3), remain (2)	2.39	9

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"Ed Sheeran in Katowice!"	Katowice (4), inauguration (3), Ed Sheeran (3), university (3)	become (2), perform (2), be (2)	2.86	14
"Sunday riots followed by a crisis"	shop (3), situation (3), place (3)	have (2)	2.96	16
"Another offspring of the Polish most famous restaurateur"	man (4) Magda Gessler (3), day (2)	make (1), say (1), throw(1), add (1), confirm (1), remain silent (1), expect (1), confirm (1)[7], meet (1)	3.27	18
"Johnny Depp passed away!"	actor (5), end (3), Johnny Deep (3), apartment (3)	be (7), find (5), can (3), become (2), suspect (2)	2.21	10
"The greatest mystery of mankind solved!"	death (3), island (3), Michael Jackson (3), artist (3), centre (3)	find (2)	3.0	11
"Polish Post does not employ individuals under the age of 55"	work (7), post (8), co-partnership (4), Poland (2), employee (2), woman (2), new (2)	be (8), conduct (2)	2.29	19
"Crowd of disappointed fans! Maciej Musiał reveals the shocking truth!"	actor (5), partner (3)	be (5)	2.89	13
"University of Economics begins the construction of an airplane runway"	airplane (4), place (4), University of Economics (3), student (3), construction (3)	be (3), construct (2), fly (2)	2.58	15
"Jim Carrey passed away"	movie (4), Jim Carry (3), depression (2), suicide (2), death (2), end (2), actor (2), role (2), summer (2)	be (4), begin (3), commit (2)	1.58	14
"Compulsory military service returns. Military trainings begin next year!"	military (9), service (4), year (3), age (3), compulsory (3)	be (2), wake (2)	2.84	14

[7] In the Polish language one of the words occurred in the perfect aspect, whereas the second one in the imperfect aspect.

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"The drama of Polish families"	next (2), ban (2), many (2), my (2), Pole (2), trade (2)	be (5)	2.79	13
"Did we just see the ending of the *Game of Thrones* series?!"	series (6), ending (4), Martin (3), own (3), all (3), fan (3)	be (8), begin (3)	2.78	17
"Krzysztof Krawczyk passed away"	Krzysztof Krawczyk (6), heart (4), problem (4), one (3), artist (3)		4.29	12
"Poland is the world champion in football! Lewandowski's goal settled the matter!"	world (4), goal (3), Poland (3), championship (3), meeting (3), Lewandowski (3), ball (3)	be (3), beat (2), win (2)	3.25	16
"80-year-old woman assaulted in Katowice centre"	perpetrator (3), Katowice (3), money (2), assault (2), elder (2), place (2), boy (2), purse (2), police (2), woman (2), attack (2)	detain (2)	2.58	13
"Record reckless driver on the A4 motorway"	Police officer (3), police (3), own (2), A4 motorway (2), Rzeszów (2), registration (2), reckless driver (2), speed (2), number (2), vehicle (2)	be (4), have (3), reward (2), know (2)	1.96	21
"Silesian Juwenalia Student Carnival cancelled"	university (5), parliament (5), Silesian (4),	say (3), be (3), take place (2)	2.45	9
"Dinosaurs in the Silesia"	dinosaur (4), garden (3), allotment (2), Chorzów (2), discovery (2), find (2), remains (2), area (2), work (2), team (2), professor (2)	notify (2), encounter (2)	3.28	18
"Sudden heat wave – weather forecasters predict 30˚C"	Africa (3), temperature (2), phenomenon (2), tourism (2), movement (2), air (2)	be (6)	2.76	12
"White confusion"	situation (2), ground (2), supporter (2), environment (2), theory (2), snow (2), topic (2), system (2), scientific (2)	be (7), ask (2), comment (2)	2.08	14

Article	Nouns, adjectives, pronouns, etc.	Verbs	Noun-to-verb ratio	Mean sentence length (in words)
"Kamil Stoch has to return the gold medal won in the Olympics"	Polish (3), Kamil (2), official (2), representation (2), committee (2), medal (2), score (2), Olympic (2)	win (2), return (2), confirm (2)	2.57	17
"Kim Dzong Un stole an asteroid from Putin"	Putin (3), Russia (3), asteroid (2), artificial (2), president (2), Kim Dzong Un (2), surface (2)	have (2), present (2), be (2), reach (2)	3.36	13
"Lewandowski in Barcelona"	Lewandowski (4), Robert (3), Barcelona (3)	be (3), connect (3)	3.69	15
"Mariacka flooded with free alcohol!"	Mariacka (4), alcohol (4), student (3), all (3), August (3), beer (3), bar (3), free (3), opportunity (3)	be (4), take place (2)	3.16	17
"Metro in the Silesia – soon!"	Katowice (7), summer (4), work (4), metro (3), city (3), problem (3), traffic jam (3)	be (7), become (2), develop (2)	2.75	12
"Michael Jackson is alive!"	Michael Jackson (4), death (4), star (3), next (3), surgery (3)	be (4), live (3), experience (2)	2.3	17

Source: Author's own study.

When comparing the analyses of the true and fake news corpora one may notice a significant difference in the noun-to-verb ratio. In real news the ratio is significantly higher: an average of 4.27 in the true news corpus compared to 2.73 in the fake news corpus. When it comes to the amount of words in a sentence, the situation is similar. The true news corpus features an average of 20.6 words, whereas the fake news corpus – 14.3. The domination of verbs is visible in both corpora and this may be a result of the popularity of the use of passive voice in the informative style.

4.1.2. The analysis of the sentiment and text complexity level

The corpora of true and fake news were analysed using the software created as part of the Clarin project (https://ws.clarin-pl.eu/sentyment.shtml) – the sentiment analyser. The analysis allows to determine how many words in the text have a positive sentiment and how many are described as negative. Moreover, it permits the identification of emotions dominating in the text (positive and negative), where one word may be a conglomerate of various emotions. Therefore, the number of words with, e.g. a positive sentiment will not be equal to the number of positive emotions in the

given text (the number of emotions is usually higher). What was also determined is the number of words in the text to create the index of the saturation of the text with emotions depending on its length.

The texts were also analysed in terms of their difficulty (complexity) level, referring to, i.a., the Gunning readability index: FOG = 0.4×(LW/LZ) + 100×(LWT/LW)], where *LW* is the number of words in a text, *LWT* – the number of words having 4 or more syllables[8] and *LZ* – the number of sentences. The Jasnopis (http://jasnopis.pl/aplikacja) application was utilized in the analysis. It allows, i.a., the determination of the text difficulty level, scoring it on the basis of a scale of 1–7, where 1 is a very simple text (comprehensible for grade 1–3 pupils) and 7 is a very complex text, understandable for specialists in the given area, doctorate holders[9]. The results of the analysis are available in Tables 3 and 4.

Table 3. The results of the sentiment analysis and the difficulty class of true news texts

Article	Number of positives	Number of negatives	Dominant positive emotion	Dominant negative emotion	Number of words in the text	Text difficulty class
"The Bialystok Commemoration Peloton on the anniversary of Soviet aggression on Poland"	12	12	joy, happiness	sadness, hurt, unhappiness	245	5
"At least one thousand people participated in the pro-European demonstration in Budapest"	4	3	trust, usefulness	anger, uselessness	152	5
"Persistence in the ecological development bears fruit in Suining, China"	18	4	joy, usefulness	anger, uselessness, mistake	220	7
"The president thanked the farmers for their efforts during the harvest festival in Spała"	28	16	joy, usefulness	sadness, unhappiness	434	5
"The Eurovision song contest to be held in Tel Aviv next year"	7	2	joy, usefulness	sadness, unhappiness	126	6

[8] For the English language it was assumed that difficult words consist of 3 or more syllables; due to the characteristics of the Polish language, and after Bartosz Broda and others, it was assumed that it will be 4 syllables for the Polish language (Broda et al. 2010).

[9] Cf. the scale and the scoring manner – http://jasnopis.pl/aplikacja#.

Article	Number of positives	Number of negatives	Dominant positive emotion	Dominant negative emotion	Number of words in the text	Text difficulty class
"USA: the number of deaths following the tropical storm in Florence reaches 17"	15	13	joy, usefulness	sadness, unhappiness	185	5
"Polish jazz in Asia on the 100th anniversary of the Polish independence restoration"	19	6	joy, happiness	anger, uselessness, unhappiness	304	6
"32 rebels and 4 radio employees died in Yemen"	6	5	joy, usefulness	sadness, anger, hurt	170	5
"Kaczyński: strong West Pomerania is in the Polish national interest"	5	3	joy, happiness	sadness, unhappiness	143	3
"Canadian Prime Minister on the Polish festival in Toronto: diversity is the source of strength"	27	13	joy, usefulness, happiness	sadness, hurt	537	6
"Cardinal Bagnasco: Europe should follow the path towards unity"	23	2	joy, usefulness	anger, hurt	255	6
"Simon Yates triumphs in the Vuelta a Espana race"	26	5	joy, happiness	sadness, anger, unhappiness	318	6
"Kornel Morawiecki: WiS to register the lists of regional council candidates in all voivodships"	4	1	joy, trust, usefulness, happiness		84	4
"Kremlin spokesperson: the individuals suspected of attacking Skripal are not connected with Putin"	6	8	trust, usefulness	sadness, uselessness, hurt	212	5
"Team from the Kielce University of Technology wins the Martian rover competition"	43	10	joy, usefulness	antipathy, hurt	423	6
"Macedonia: several thousand participate in the demonstration supporting the EU and the NATO"	3	1	joy, usefulness, happiness	anger, fear, surprise with the unpredictable, uselessness, hurt, unhappiness	126	5

Article	Number of positives	Number of negatives	Dominant positive emotion	Dominant negative emotion	Number of words in the text	Text difficulty class
"The head of the Ministry of National Defence meets the World War II heroes in New York"	17	4	joy, usefulness	anger, hurt	146	4
"National Readout of *The Spring to Come* by Stefan Żeromski begins"	15	6	joy, usefulness	sadness, hurt	436	5
"Prime Minister: it is the ruling camp that follows the constitution the most"	6	6	trust, usefulness	sadness, hurt	156	3
"Poland wins 3:2 with Romania in the Davis Cup"	27	1	joy, happiness	anger, uselessness, hurt	267	5
"Experts: respiratory system diseases as the second most commons cause of death in the world"	35	29	joy, usefulness	sadness, unhappiness	503	6
"Sasin: we never announced the 500 plus programme for pensioners"	9	3	joy, usefulness	sadness, uselessness, unhappiness	244	3
"Head of SLD on the coalition with PO: I can talk about it, but on equal terms"	16	1	joy, usefulness	anger, uselessness	196	4
"European community associated with head tumours joins forces to understand the rare cancer form"	21	2	joy, usefulness	sadness, uselessness	358	7
"Syria accused Israel of conducting a missile attack on the Damascus airport"	6	10	joy/trust, usefulness	sadness, hurt	273	5
"Szczerski: an important Polish-American declaration – a potential effect of the USA's presidential visit"	14	11	joy, usefulness	sadness, hurt, unhappiness	205	7

Article	Number of positives	Number of negatives	Dominant positive emotion	Dominant negative emotion	Number of words in the text	Text difficulty class
"Tymoteusz Bies as the best pianist of the Szymanowski International Music Competition"	14	3	joy, usefulness	sadness, unhappiness	177	6
"Trump: the Three Seas Initiative has great potential"	7	0	joy, happiness		73	6
"First local elections in years in Syria, candidates mostly from Assad's party"	9	3	joy, usefulness	sadness, anger, antipathy, mistake	191	5
"Turtle extinction may negatively influence the environment"	11	11	trust, usefulness	anger, hurt	262	4

Source: Author's own study.

Table 4. The results of the sentiment analysis of the fake news texts

Article	Number of positives	Number of negatives	Dominant positive emotion	Dominant negative emotion	Number of words in the text	Text difficulty class
"Beata Szydło attacked near Łazienki"	2	4	joy, usefulness	anger, unhappiness	118	3
"Biedronka shops open on trade Sundays!"	9	1	joy, usefulness	anger, antipathy, unhappiness	147	4
"Painful loss in the life of Sławomir"	13	4	joy, happiness	anger, hurt, unhappiness	103	4
"Borys Szyc passed away"	2	4	joy, usefulness	sadness, uselessness	102	4
"Ed Sheeran in Katowice!"	13	1	joy, usefulness, happiness	anger, hurt, unhappiness	92	4
"Sunday riots followed by a crisis"	5	5	joy, usefulness	anger, uselessness	114	4
"Another offspring of the Polish most famous restaurateur"	6	4	joy, usefulness, happiness	anger, fear, unhappiness	85	4
"Johnny Depp passed away!"	7	8	joy, usefulness	sadness, unhappiness	124	3
"The greatest mystery of mankind solved!"	7	5	joy, happiness	sadness, unhappiness	107	3

Article	Number of positives	Number of negatives	Dominant positive emotion	Dominant negative emotion	Number of words in the text	Text difficulty class
"Polish Post does not employ individuals under the age of 55"	12	4	joy, usefulness	sadness, uselessness	163	4
"Crowd of disappointed fans! Maciej Musiał reveals the shocking truth!"	10	4	joy, usefulness	sadness, uselessness	125	4
"University of Economics begins the construction of an airplane runway"	5	0	trust, usefulness		83	4
"Jim Carrey passed away"	5	8	joy, usefulness	sadness, uselessness, hurt	133	3
"Compulsory military service returns. Military trainings begin next year!"	14	9	trust, usefulness	sadness, hurt	170	5
"The drama of Polish families"	3	1	joy, excitement due to something unexpected, usefulness/ happiness	uselessness, hurt	62	4
"Did we just see the ending of the *Game of Thrones* series?!"	6	7	joy, usefulness	sadness, uselessness	123	4
"Krzysztof Krawczyk passed away"	3	4	joy, usefulness	sadness, anger, fear, hurt	121	5
"Poland is the world champion in football! Lewandowski's goal settled the matter!"	10	3	joy, happiness	anger, uselessness	115	4
"80-year-old woman assaulted in Katowice centre"	1	14	joy, usefulness	anger, hurt, unhappiness	95	3
"Record reckless driver on the A4 motorway"	4	2	joy, trust, usefulness	anger, uselessness, mistake	107	5
"Silesian Juwenalia Student Carnival cancelled"	6	1	joy, usefulness	anger, mistake	93	4
"Dinosaurs in the Silesia"	6	4	trust, usefulness	anger, unhappiness	91	5

Article	Number of positives	Number of negatives	Dominant positive emotion	Dominant negative emotion	Number of words in the text	Text difficulty class
"Sudden heat wave – weather forecasters predict 30˚C"	5	2	joy, usefulness, happiness	fear, unhappiness	105	4
"White confusion"	3	4	joy, usefulness	anger, uselessness	139	4
"Kamil Stoch has to return the gold medal won in the Olympics"	5	2	joy, usefulness, happiness	sadness, anger, uselessness, unhappiness, hurt	103	5
"Kim Dzong Un stole an asteroid from Putin"	8	6	joy, happiness	sadness, mistake	94	4
"Lewandowski in Barcelona"	5	2	joy, happiness	fear, uselessness, unhappiness	81	4
"Mariacka flooded with free alcohol!"	9	0	joy, usefulness	sadness, anger, fear, uselessness, unhappiness, hurt	101	5
"Metro in the Silesia – soon!"	8	2	joy, usefulness	sadness, anger, unhappiness, hurt	122	3
"Michael Jackson is alive!"	8	10	joy, happiness	anger, unhappiness	128	4

Source: Author's own study.

The analysed text corpora have a different (although not diametrically different) difficulty levels: an average of 5.17 for the true and 4 for the fake news corpus. The fake news texts are generally understandable for individuals with secondary education or those having a large amount of life experience, whereas the true news texts are, in general, more difficult and comprehensible for educated individuals. The text difficulty difference is one point in the seven-point scale.

However, what is worth noting is the large differentiation of the text difficulty due to the subjects and the defined target audience. In the true news corpus, the information on the politics in the country was the most simple (usually 3 or 4 points), whereas the most difficult texts were associated with scientific (ecology, medicine) and legal matters. It is not possible to distinguish such dependencies in the fake news corpus. What is significant, the fake texts corpus did not feature any texts with the highest difficulty levels – 6 or 7 points on the difficulty scale.

In the elaborations associated with lies in the language, the negative sentiment notions are considered as the indicators of falsity. In the true news corpus, the negative notions dominated over the positive ones in two texts (one of the articles focused on the assassination of a former Russian agent and his daughter in Great Britain, the other – on war in Syria). There were also two articles with an identical number of positive and negative sentiment notions (one focusing on the anniversary of the Soviet aggression on Poland and the second one – on the effects of turtle extinction). On the other hand, in the fake news corpus there were seven texts where the negative notions dominated and one where the number of positives and negatives were equal.

4.2. Qualitative analysis

The conducted discursive analysis shows that the true news often features an individual's full name, preceded with the name of their position (sometimes only the name of the institution is available) and one of the statements such as: said, announced, highlighted, underlined, added, called, stated, wrote, thanked, advised, informed, confirmed, etc. Some of the conclusions from the analysis prove challenging to be used in the model directly, e.g. the length of the sentence, despite the fact that the corpus conclusion analysis indicates a different average for the real and fake news, is not a decisive criterion in a single text. A similar situation occurs in terms of the average text complexity level.

The conclusions on whether the text is true or false should also be associated with the declared genre as news, articles, etc. differ. Therefore, some of the variables should be adjusted to the statement genre. The conclusions drawn in this article apply to news only.

When re-creating the interpretative framework, it is worthwhile to commence with a simple comparison of the article title and the keywords. If the determined keywords are not clearly connected with the title or are connected with it in vague manner, one may assume that the given information is fake. Should one encounter such an article, it is worth to compare the sentiment of the subject with the sentiment of the whole article. If the title is unambiguously positive or negative and the text content presents an opposite sentiment, the probability of the text being false is higher – cf., e.g. the "Painful loss in the life of Sławomir" text, where 2 of the 7 words in the title resonate with a negative and none with a positive sentiment and where the entire text seems definitely positive (13:4)[10]. The article keywords feature the word "sell", which is not relevant to the framework mentioned in the title – the framework of loss, usually of a close person or a valuable item.

[10] It is similar in case of the following texts: "Sunday riots followed by a crisis" (lack of compliance between the title and the keywords, negative sentiment of the title, a 5:5 sentiment of the article), "Crowd of disappointed fans! Maciej Musiał reveals the shocking truth!" (low compliance of the keywords and the title, dominant presence of negative notions in the title, domination of positive notions in the text 10:4), "The drama of Polish families" (low compliance of the keywords and the title, dominant presence of negative notions in the title, domination of positive notions in the text 3:1).

5. The probability model of the recognition of the information as truthful

The conducted analyses prove that only the compilation of many factors associated with the text analysis and (as per the literature research) the determination of the source credibility may increase the probability with which one can decide whether the information in focus is true or fake. The model aiding in such distinctions has been presented below.

In order to increase the probability of the information being true or false, one should:

1. Analyse the text complexity level: if the article score is 6 or 7 points, the probability of the article's truthfulness increases.[11]

2. Verify the compliance of the keywords with the title (initially, a simple verification whether the keywords appear in the title; next, whether most important verbs permit, as per the Valence dictionary, constructions compliant with the interpretative framework appearing in the title): low or lack of compliance increases the probability that the given article is false, high compliance – that it is true.

3. Verify the sentiment of the subject and compare it with the sentiment of the whole article: if the sentiments are not in line, the probability that the news is fake increases.

4. Verify whether the text contains a source, usually an individual's full name along with their position and one of the verbs introducing an utterance. The presence of such formulae increases the probability that the news in focus is real.

5. Create a list of credible sources. If it contains Internet websites, newspapers, magazines, TV and radio stations, etc. then the credibility is defined on the basis of the quality of the conveyed information. In social media – it is defined on the basis of what the given user shared previously, whether they quote (forward, re-tweet) credible sources, how long they function in the given media, whether they possess a confirmed account, how many individuals follow the particular person, how many friends does he or she have and other factors dependent on the specific medium.

6. Conclusions

The analysis of the corpora of the gathered news shows that some of its elements may prove useful when creating a fake news identification model. The results presented in the article are associated with small corpuses – this research is exploratory, it will be broadened and the results will be verified on larger corpuses that are more diverse in terms of genre and the publication location (featuring both news and articles as well as posts/entries on the social media).

[11] One should note that not all media feature such articles. A lower index does not increase the probability that the given information is fake.

The analysis proved that the creation of a fake news identification model may be possible, however, it will not be a tool allowing to determine beyond all doubt, and in all cases, whether the given information is true or not. This is due to the fact that there are various actions, including the actions of the particular countries' institutions, that misinform, distort and create false information that is confirmed by the representatives of these institutions. In such cases, the journalists conveying the given information are almost certain that it is true – e.g. the fictional assassination of Arkady Babchenko. And in such cases, the model will not be of much assistance.

References

Antas J. (2008). *O kłamstwie i kłamaniu*. Universitas: Kraków.

Austin J.L. (1961). *Truth.* In: *Philosophical Papers*, J.L. Austin. Oxford University Press: Oxford, pp.117–133.

Brewer P.R., Goldthwaite Young D., Morreale M. (2013). The Impact of Real News about "Fake News": Intertextual Processes and Political Satire. *International Journal of Public Opinion Research*, Vol. 25(3).

Broda B. et al. (2010). Trudność tekstów o Funduszach Europejskich w świetle miar statystycznych. *Rozprawy Komisji Językowej Wrocławskiego Towarzystwa Naukowego*, Vol. 37.

Chopra S., Jain S., Sholar J.M. (2017). *Towards Automatic Identification of Fake News: Headline-Article Stance Detection with LSTM Attention Models*. Stanford CS224d Deep Learning for NLP final project.

Conroy N.J., Rubin V.L., Chen Y. (2015). *Automatic Deception Detection: Methods for Finding Fake News*. ASIS&T 2015, November 6–10, St. Louis, USA.

Enos F. et al. (2015). *Human Detection of Deceptive Speech*, https://pdfs.semanticscholar.org/c0d3/6377afe2d3f3adf85e1aa73e95798ca35e69.pdf, 25.06.2018.

Hancock J., Toma C., Ellison N. (2007). *Lying in Online Data Profiles*. CHI 2007, April 28 – May 3, 2007, San Jose, USA.

Kumar K., Geethakumari G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, Vol. 4(14), http://www.hcis-journal.com/content/4/1/14, 26.06.2018.

Newman M. et al. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, Vol. 29(5), pp. 665–675, DOI: 10.1177/0146167203251529.

Rubin V., Chen Y., Conroy N. (2015). *Deception detection for news: Three types of fakes*. ASIS&T 2015, November 6–10, St. Louis, USA.

Shedletsky L. (2018). Seeing bullshit rhetorically: Human encounters and cultural values. *Res Rhetorica*, Vol. 5(4).

Vosoughi S., Roy D., Aral S. (2018). The spread of true and false news online. *Science*, Vol. 359(6380), pp. 1146–1151.