Journal of Optimization in Industrial Engineering
Vol.15, Issue 1, Winter & Spring 2022, 301-309
DOI:10.22094/JOIE.2021.1945486.1913

Research Article

Uncertainty Identification in Microblogs

Fairouz Zendaoui [a],*, Walid Khaled Hidouci [a], Saeed Rouhani [b]

[a] *Laboratoire de la Communication dans les Systèmes Informatiques, Ecole Nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. http://www.esi.dz*

[b] *Department of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran*

Received 10 September 2021; Revised 18 November 2021; Accepted 15 December 2021

Abstract

Microblogging, like Twitter, has grown in popularity as a medium for human expression, allowing users to quickly create content on breaking news, public events, or products. The vast volume of microblogging data is a valuable and timely source of public attitudes, views, and opinions on a variety of issues. Users express themselves freely with varied degrees of uncertainty, which makes using microblogs as a data source a tedious task that necessitates the consideration of this factor. We're talking about the uncertainty expressed in microblogs, not the one relative to the claimed information factuality. This aspect of microblogging that we are addressing has gotten little attention, despite the fact that it is critical to understand the degree of ambiguity with which users plan to share information. This topic is particularly relevant to research projects involving information retrieval or investigation in microblogs. In this paper, we present a state of the art on the identification of uncertainty in microblogs with the aim of identifying this issue and describing the current knowledge through the study of similar or related work. We mainly concluded that, to adapt to the characteristics of social media, it is necessary to identify the uncertainty based on the contextual uncertain semantics rather than the traditional cue-phrases, and considering multiple sub-classes could provide more information for research on handing uncertainty in social media texts.

Keywords*:* Uncertainty Identification; Microblogs; Semantics; Tweets Classification; Social Media.

1. Introduction

Microblogging is a new type of online communication in which individuals write brief postings about their everyday lives, discuss their thoughts, express their feelings, and exchange information. There are a variety of microblogging sites accessible these days, and they are utilized for different purposes. Some are mostly used for chat, while others are used for image and video sharing, while yet others are utilized for formal and official purposes. Twitter is one of the most outspoken platforms for sharing feelings and opinions on a wide range of topics, including sports, entertainment, religion, politics, and many more. Microblogging has grown in popularity as one of the most popular social networking platforms today, making it a potentially vast data source that is garnering the interest of academics working in the fields of knowledge discovery and data mining (Hua et al., 2012). The data may be utilized to extract a wealth of useful information that can be used to design future technologies.

The automatic extraction of information from noisy sources has opened up new possibilities for querying and analyzing data, particularly in today's era of social media domination, where microblogs like Twitter are probably the best source of current information. The amount of data accessible on these platforms is enormous, but the majority of it is unstructured, redundant, and ambiguous, making extracting information from it considerably more difficult (Kumar et al., 2017). Several works have focused on extracting information from microblogs for several purposes such as: opinion investigation (Rouhani & Abedin, 2019), rumor detection (Feng et al., 2021), information search and retrieval (Basu et al., 2020), user behavior tracking (Yan et al., 2013), prediction (Jordan et al., 2019), sentiment analysis (Kothandan & Murugesan, 2021), data modelling (Rajabi et al., 2020) and marketing management (Ghelichkhan et al., 2020).

Many social media-based applications require uncertainty text detection because an increasing number of users use social media platforms as a source of information to generate or deduce interpretations based on them (Wei et al., 2013). Uncertainty identification is a fundamental semantic processing task in many Natural Language Processing techniques and applications, such as question answering, information extraction, and so on, because it determines the quality of information in terms of factuality. Because social media writings are typically produced in a wildly inconsistent manner, factuality becomes a significant problem (Li et al., 2018).

The majority of current uncertainty annotation work is based on the CoNLL-2010 task of uncertainty identification (Farkas et al., 2010), in which the uncertainty texts were annotated using cues. Because uncertainty cues are typically found in uncertain

*Corresponding author Email address: f_zendaoui@esi.dz statements (Li et al., 2018). Existing techniques that rely on lexical cues, on the other hand, suffer considerably from the informal or word-of-mouth nature of social media, in which cue phrases are frequently presented in poor form or even missing from sentences. When compared to the current uncertainty corpus, the expression of doubt in social media is quite different from that in formal language, in the sense that when making uncertain remarks, users frequently ask questions or link to external information (Wei et al., 2013). Szarvas et al. (2012) proposed an interesting classification of semantic uncertainty phenomena, which has now been expanded in some research to include uncertainty in microblogs.

As microblogs have become widely used for sharing knowledge and experience and also a source of information in which users seek to find truths, we are particularly interested in works that have dealt with the detection of uncertainty in microblogs expressions. In this paper we present a state of the art on the identification of uncertainty in microblogs with the aim of identifying this issue and describing the current knowledge through the study of similar or related work.

This rest of this paper is organized as follows. In section 2, we introduce the notion of microblogging as a mass communication medium to frame the background of the study. In section 3, we approach the identification of uncertainty, in particular, semantic uncertainty, and present some related works. We focus in section 4 on works dealing with the uncertainty in microblogs and we bring out the different types of uncertainty defined in this context. We conclude the paper in section 5.

2. Microblogging

2.1. Background

One of the most famous Web 2.0 representatives and typically named in terms of Social Media are Blogs and Microblogs. Both are used for communication, knowledge management and publishing. Weblog is a combination of the words Web and Log, and it refers to an information system that keeps track of diary entries online. Microblogs are short blog postings that run from 140 to 250 characters in length (Ebner, 2018). Microblogging, according to Java et al. (2007), is a small-scale type of blogging made up of short, concise messages that is used by both consumers and companies to share news, publish status updates, and continue dialogues. Both blogs and microblogs use a chronological sequence of entries, with the most recent entries appearing first. Long form blog entries, on the other hand, appear to be considerably slower and plodding in comparison to the speed with which information may come and go on microblogs like Twitter, where microblogging communities promote rapid word of mouth of both positive and bad content (Owyang, 2008).

2.2. Microblogs functioning

There are other alternative platforms, such as Tumblr, Yammer, and even Instagram, but Twitter is without a doubt the most well-known. Anyone with a Twitter account may send messages, known as Tweets. The simple rule is that each Tweet must be no more than 140 characters long. By simply following other users and vice versa, each user may create friends. Followers are those who are interested in another user (Ebner, 2018).

As example a tweet might be: “RT @beatdoebeli: >>Personal Smartphones in Primary School: Devices for a PLE?<< http://www.xxx.com #pdf #PLE #iPhone”. A hash (#) sign combined with a specific tag, known as a Hashtag (#pdf #iPhone), is an extremely potent combo. This may be viewed as a form of social-meta-data created and provided by the users. For example, it's simple to search or filter for certain Hashtags, making it simple to keep track of an ongoing event. It is possible to address another user publicly by using a @ symbol and a username. Finally, each public tweet can be retweeted by a user's followers. In other words, retweeting is the ability to reach a large number of people in a short amount of time — spreading the word in twitter lingo (Ebner, 2018).

2.3. Mass Communication through microblogging

Many individuals on Twitter and other platforms make messages to no one in particular. In other words, they are broadcasting messages to an unidentified audience. This finding is consistent with the fundamental notion of mass communication. As a result, microblogs may be considered a form of mass media, as well as a form of mass communication. Tweets are messages delivered by different communicators (users) to different receivers (followers and/or those who read microblogs but are not identifiable as followers of a person) via a medium (application). Microblog apps are used by individuals all over the globe and may be accessed by mobile, client, or online interface. Unlike traditional media, messages on a Microblog timeline display in real time and are not delayed in any way. In the same way as conventional mass media is bound by public opinion, beliefs, norms, and values, a Twitter user is bound by public opinion, believes, norms, and values in his utterances. Without any preconditions, Twitter allows users to join debates in real time. Both spontaneous and long-term social ties may evolve as a result of this engaged kind of communication (Ebner, 2018).

3. Uncertainty Identification

To investigate the identification of uncertainty in microblogs, we must first explain our meaning of the term uncertainty. Uncertainty may be viewed as a lack of information in its broadest sense: the receiver of the information (i.e., the listener or reader) cannot be positive of some bits of information (Vincze, 2014). In this way, uncertainty differentiates from factuality and negation: in the former, the hearer/reader is confident that the information is correct, whereas in the latter, he is positive that the information is incorrect. Uncertainty arises in computer science because of incomplete observability, nondeterminism, or both (Russell & Norvig, 2002). The concept of modality is frequently associated with uncertainty in linguistic theories: epistemic modality encodes how much certainty or evidence a speaker has for the proposition expressed by his utterance (Palmer, 2001), or it refers to a possible state of the world in which the given proposition holds (Palmer, 2001). (Kiefer, 2005). The common thread running through all of the techniques is that when there is ambiguity, the truth value/reliability of the statement cannot be determined since another piece of information is absent. Thus, in our perspective, uncertain propositions are ones whose truth value or dependability cannot be assessed owing to a lack of knowledge (Vincze, 2014). We are more interested in the expression of doubt in microblogs than in the uncertainty connected to the stated information factuality in our research. Previous work on uncertainty classification centered on categorizing sentences into uncertain or definite groups. Existing methods mostly rely on supervised methods (Light et al., 2004; Medlock & Briscoe, 2007; Medlock, 2008; Szarvas, 2008) that use an annotated corpus with various types of characteristics such as Part-Of-Speech (POS) tags, stems, n-grams, and so on (Wei et al., 2013).

3.1. Corpora annotated for uncertainty

In recent years, uncertainty has received a lot of attention in NLP applications in different domains, such as biology (Medlock & Briscoe, 2007; Kim et al., 2008; Settles et al., 2008; Shatkay et al., 2008; Vincze et al., 2008; Nawaz et al., 2010), medicine (Uzuner et al., 2009), news media (Saur & Pustejovsky, 2009; Wilson, 2008; Rubin et al (Wei et al., 2013). Some of these traits are briefly summarized here (Vincze, 2014):

- The Genia Event corpus (Kim et al., 2008), which uses negation and two forms of uncertainty to annotate biological events (9,372 sentences).

- The BioScope corpus (Vincze et al., 2008), which contains three types of biomedical texts – radiological reports, biological full papers, and abstracts from the Genia corpus – annotated for both negation and hedging keywords and their linguistic scopes (20,924 sentences).

- • The WikiWeasel corpus served as the CoNLL-2010 Shared Task's training and assessment database (Farkas et al., 2010). It has 20,745 sentences and is annotated for weasel hints (4,718 of which are uncertain).

- The FactBank database (Saur & Pustejovsky, 2009) includes annotations for events, sources, and factuality, among other things, with four categories of factuality distinguished.

Table 1 highlights the characteristics of each corpus based on its original annotation and terminology (Vincze, 2014).

Table 1

Features of the corpora

	BioScope	**Genia Event**	**WikiWeasel**	**FactBank**
keyword	•		•	
Target word		•		•
event		•		•
scope	•			
negation	•	•		•
speculation	•		•	
probable		•		•
doubtful		•		
possible				•
underspecified				•
concept of source			•	•
weasel			•	

Despite the fact that these corpora are all labeled for uncertainty, each one interprets the phrase in a slightly different way. The phrases hedge, speculation, factuality, polarity, weasel, and uncertainty are used in the above corpora to describe the phenomena uncertainty, whereas propositions might be uncertain, speculative, probable, possible, or doubtful. The definitions used in the four corpora were supplied by Vincze (2014), and discrepancies and similarities were discussed. There are significant commonalities but also discrepancies in the meaning of the aforementioned concepts, as evidenced by publicly accessible annotation rules, which may be linked to domain- and genre-specific elements of the texts. However, finding a common ground among the many terminology and notions of uncertainty would be better. The phrase uncertainty may be used as an umbrella term encompassing phenomena at the semantic level, based on corpus data and annotation principles.

3.2. Semantic uncertainty

Various concepts and terms associated with uncertainty phenomena are used. Modality is usually associated with uncertainty (Palmer, 2001), but the terms factuality (Saurí & Pustejovsky, 2012), veridicality (de Marneffe et al., 2012), evidentiality (Aikhenvald, 2004) and commitment (Diab et al., 2009) are also used. They are all similar but somewhat different language problems that fall under the semantic uncertainty category. At the semantic level, propositions might be uncertain, meaning that their truth value cannot be established solely based on the speaker's mental state. Szarvas et al. (2012) propose a categorization for semantic uncertainty. We use the word uncertainty here in the same way as Szarvas et al. (2012) used in their attempt to provide a coherent framework for the aforementioned phenomena: "uncertain propositions are those [...] whose truth value or dependability cannot be evaluated due to a lack of knowledge" (Vincze, 2014). Truth conditional semantics can be used to define semantically uncertain statements. Given the speaker's current mental state, they cannot be assigned a truth value, i.e., it is impossible to say whether they are true or untrue (Vincze, 2014). Semantic level uncertainty may be divided into ***epistemic*** and ***hypothetical*** categories, according to Szarvas et al. (2012). While instances of hypothetical uncertainty can be true, false, or uncertain, epistemically uncertain propositions are definitely uncertain – in terms of possible worlds, hypothetical propositions allow for the proposition to be false in the actual world, but in the case of epistemic uncertainty, the factuality of the proposition is unknown.

It is recognized that the assertion is neither true nor false in the situation of epistemic uncertainty: That describes a possible world in which the proposition holds, but this possible world does not correspond to the speaker's actual reality. In other words, the proposal is certain to be uncertain. Epistemic uncertainty is linked to epistemic modality: a phrase is epistemically uncertain if we can't tell whether it's true or untrue right now based on our world knowledge (thus the name) (Kiefer, 2005). At the same time, the source of an epistemically uncertain assertion cannot claim the uncertain notion while being confident of its opposite (Szarvas et al., 2012).

In the case of hypothetical uncertainty, neither the truth value of the propositions nor the likelihood of their occurrence can be assessed. Propositions under inquiry are an example of such statements: the truth value of the proposition under investigation cannot be determined until additional examination. Hypotheses can also be categorized as instances of conditionals. It's also typical in these two sorts of uncertain propositions for the speaker to say them while knowing (for others or even for himself) that the contrary is true, therefore they're known as paradoxical uncertainties (Vincze, 2014; Szarvas et al., 2012).

Non-epistemic types of modality are also linked to hypothetical uncertainty. The speaker's beliefs and hypotheses – which others may know to be true or incorrect in the current condition of the world – are expressed in the doxastic modality. The fundamental goal of deontic modality is necessity (duties, obligations, orders), dispositional modality is determined by the person's dispositions (i.e. physical abilities), and circumstantial modality is defined by external circumstances. Wishes, intents, goals, and aspirations are all associated to the Buletic modality. Dynamic modality is an umbrella term covering deontic, dispositional, circumstantial, and buletic modality (Kiefer, 2005; Vincze, 2014; Szarvas et al., 2012).

To summarize, uncertainty can be classified into two types: epistemic and hypothetical (Kiefer, 2005). Possible and Probable are the two sub-classes of Epistemic. Investigation, Condition, Doxastic, and Dynamic are the four sub-classes of Hypothetical. The categorization is detailed below (Kiefer, 2005):

- **A. Epistemic**: On the basis of our world knowledge, we cannot decide at the moment whether the statement is true or false). There are two sub-classes Possible and Probable.
- **B. Hypothetical**: This type of uncertainty includes four sub-classes:
  - **Doxastic**: Expresses the speaker’s beliefs and hypotheses.
  - **Investigation**: Proposition under investigation.
  - **Condition**: Proposition under condition.
  - **Dynamic**: Contains deontic, dispositional, circumstantial and buletic modality.

Below, we present some examples of instances provided by Vincze (2014). In his paper, the latter have done more examples illustrating the test-based classification of propositions with the cues.

- a) Epistemic: It may be raining.
- b) Hypothetical:
  - Dynamic: I have to go.
  - Doxastic: He believes that the Earth is flat.
  - Investigation: We examined the role of NF-kappa B in protein activation.
  - Condition: If it rains, we’ll stay in.

4. Uncertainty Identification in Microblogs

There are now a large number of social networks available on the Internet, and they are constantly expanding and evolving, both broadly and deeply. The number of social network users in the globe exceeded 3.8 billion people at the beginning of 2020, a 9% percent growth over 2019, and this number will continue to rise, boosted by the trend of individuals actively transitioning to use social networks from mobile devices (Bessarab et al., 2021).

With the rise of social media, there are an increasing number of text contents comprising a big number of informal or word-of-mouth terms. In terms of factuality, the quality of information on social media has become a big issue (Wei et al., 2013). In general, only true information with a high level of credibility has value for usage. However, the majority of social media expressions are released with a hypothesis and episteme that cannot be determined if they are true or incorrect at the moment. Previous research looked at the existence of uncertain statements and found that they were common on social networking sites such as Facebook, Twitter, and Sina Weibo. According to a twitter dataset statistic (Wei et al., 2013), at least 18.91% of tweets have an uncertainty characteristic. As a result, the quality of information in social media has become a real concern for a variety of social media-related activities, including rumor detection (Qazvinian et al., 2011) and credibility assessments (Jaworski et al., 2014). As a result, users must be able to identify uncertainty in order to synthesize data and generate a trustworthy interpretation (Li et al., 2018).

However, uncertainty identification in the context of social media is rarely investigated. Furthermore, unlike biology studies and Wikipedia entries, social media messages are typically brief and informal. Many cue phrases are presented in substandard shape or even omitted from sentences due to the word count constraint and casual speech. The ambiguous semantics will be provided implicitly by the entire sentence rather than explicitly by the cue words in this situation (Li et al., 2018). As a result, applying current out-of-domain corpora to the context of social media is ineffective.

Furthermore, when compared to the existing uncertainty corpus, the expression of uncertainty in social media differs from that in formal language in the sense that when making uncertain claims, users frequently ask questions or refer to external information. However, neither of the uncertainty expressions can be described using the literature's known categories of uncertainty. As a result, in the context of social media, a separate uncertainty categorization method is required (Wei et al., 2013).

Wei et al. (2013) conducted an empirical study on uncertainty identification in social media texts that took into account features other than plain text, such as the number of tweets and their relationships, and Vincze (2014) proposed using lexical, morphological, syntactic, semantic, and discourse-based features in a supervised classifier (Han et al., 2019). Based on supervised sequence labeling algorithms, Al-Sabbagh et al. (2015) provided a unified framework for identifying and extracting uncertainty cues, holders, and scopes in Arabic tweets.

Deep learning has also been utilized in the detection of uncertainty. On the CoNLL-2010 benchmark datasets, Adel & Schutze (2016) proposed an attention architecture for uncertainty detection that used a CNN or RNN with an attention mechanism to produce state-of-the-art results. An external lexicon of seed cue words or phrases was also integrated into the word embedding, and their model performed well on English datasets with this external information. However, because of the lengthy phrases that regularly appear on social media in both English and Chinese, and since the usage of a CNN or RNN results in a loss of semantics, this model is unsatisfactory for social media writings (Han et al., 2019). Han et al. (2019) proposed an approach with three major differences from previous work: (1) their model only uses word embedding, with no additional knowledge, external systems, or cue words; (2) their proposed neural networks use LSTM and the attention mechanism to generate the semantic focus, which can represent well the long sentences and substandard expressions typical of social media texts; and (3) they are the first to construct a unified experimental dataset.

4.1. Semantic uncertainties in microblogs

This section of the study focuses on similar investigations that used Szarvas et al. (2012)'s semantic categorization of uncertainty to detect uncertainties in microblogs.

First, we'll go through the work of Wei et al (2013). They investigated the effectiveness of several categories of characteristics from the social media context in an empirical study of uncertainty identification on a dataset of tweets. They came up with the following three key observations:

- No tweets of the type Investigation have been identified. They observed that while posting tweets, users rarely utilize terms like "examine" or "test" (indicative words of the Investigation category). When they have done that, the assertion should be regarded as extremely certain.
- Individuals regularly ask clarifying questions about certain issues, expressing their hesitation.
- People are more likely to send messages including external information (for example, a tale from a friend), which implies doubt.

They developed a variation of uncertainty types in the social media context based on these findings, removing the category of Investigation and adding the categories of ***Question*** and ***External*** under Hypothetical. Their suggested technique is based on Kiefer's (2005) work, which was earlier developed by Szarvas et al. to normalize uncertainty corpora in other genres (2012). However, they did not test these expanded schemas for specific genres because even the most basic one (Kiefer, 2005) was shown to be inadequate in the domain of social media.

As a result, Wei et al. (2013) proposed a new classification method for identifying uncertainty in social media and created the first uncertainty corpus using tweets. They ran uncertainty identification tests on the created dataset to see how well different types of features, such as ***content-based***, ***user-based***, and ***Twitter-specific*** features, performed. The findings demonstrate that the three types of social media-specific features can enhance the uncertainty detection. Furthermore, among the three, content-based features improve the most, and the existence of uncertain cues contributes the most to content-based features.

Another study was conducted by Li et al. (2018), who focused on annotating social media writings in Chinese with the purpose of identifying uncertainty. They are the first to annotate a Chinese social media corpus for an uncertainty detection challenge, the UIR Uncertainty Corpus (UUC).

The annotation technique used in this study is based on uncertainty categories done by Wei et al. (2013). However, due to differences in linguistic habits or usage between English and Chinese, several alterations were made to their annotation scheme. They noticed that the epistemic sub-classes of Possible and Probable in Chinese uncertainty expressions are very similar when compared to the Chinese uncertainty corpora. These two sub-classes were grouped into a single class called ***Possible***. Furthermore, they found that in Chinese social media, there were no expressions belonging to the Dynamic sub-class in any of the microblogs. As a result, the Dynamic sub-class is likewise removed. On the contrary, they observed that there were many words in Sina microblogs conveying the uncertainty statement with predicted semantics. They created a variation of uncertainty sub-class in Chinese social media corpus based on the above data by deleting the Dynamic type and adding the sub-class of ***Hope***.

The uncertainty scheme was updated and redefined into six sub-classes, as follows:

- **Question**: Sentences are used to ask for information or to test someone’s knowledge.
- **External**: Speaker repeat exactly what another person has said or written.
- **Doxastic**: Expresses the speaker’s beliefs and hypotheses.
- **Hope**: Expresses the speaker's desire for something to occur or be true, as well as his or her belief that it is feasible or plausible.
- **Possible**: Expresses something it can be done or achieved.
- **Condition**: Proposition is subject to change.

Each tweet must be labeled as either uncertain or certain in order to complete the annotation objective. Moreover, rather than the existence of uncertain cues, uncertainty statements should be recognized in terms of judgements about the author's intended meaning or implicit semantics. The sub-classes must also be designated in accordance with the system of notation. The study revealed that the aforementioned classification improved the corpus's detail and made it more valuable for future NLP applications.

4.2. Comparison and Discussion

By analyzing the only available works that performed tweets classification by semantic uncertainty types, we find that this research domain is very context-dependent, especially, on the community, its way of thinking, its way of speaking and its proper language. Between just two different languages (English and Chinese), there were many differences in the expression of uncertainty which resulted in the latter two different adaptations of the same semantic uncertainty classification model (given by Szarvas et al. (2012)), by maintaining and / or eliminating types from the model as shown in the table 2.

Table 2

Comparison between two annotation schemes for uncertainty done respectively by Wei et al. (2013) and Li et al. (2018)

Semantic Uncertainty Types	**The annotation scheme for uncertainty**	
	Wei et al. (2013)	Li et al. (2018)
Epistemic		
Possible	•	Both merged in one Possible type
Probable	•	
Hypothetical		
Doxastic	•	•
Investigation		
Condition	•	•
Dynamic	•	
Question	•	•
External	•	•
Hope		•

Table 3 summarizes and compares the results obtained in the two works. Both adopted the same annotation process using two annotators trained to annotate independently for certain / uncertain types and one annotator for resolving conflict labels. The annotation results obtained by Wei et al. (2013) showed that 19.52% of tweets are labeled as uncertain. Question is the uncertainty category with most tweets (52,69%), followed by External (22,46%) and Probable (13,93%). Regarding results of Li et al. (2018), 27.56% of microblogs are labeled as uncertain where Question is also the uncertainty category with the most microblogs (53,02%), followed by Possible (26,85%). In both cases, more than half of the uncertainties represent questions. This distribution of the results is really impressive and calls into question the works which have not considered all these categories of uncertainty, especially Question type, and which have confused them with the rest of the expressions.

Table 3

Comparison between the annotation results obtained respectively by Wei et al. (2013) and Li et al. (2018)

Annotation Results	**Wei et al. (2013)**	**Li et al. (2018)**
Data set source	Twitter	Sina Weibo
Annotation process	Two annotators trained to annotate independently for certain/uncertain types and one annotator for resolving conflict labels.	
Uncertainty assertions	Based on the semantic rather than the cue-phrases.	
Labeling multiplicity	Multilabel allowed for sub-classes	Single label annotation
Total number of microblogs	4743	40168
Uncertain microblogs	19,52 %	27,56 %
Uncertain category statistics	**Question** (**52,69**%) External (22,46%) Probable (13,93%)	**Question** (**53,02**%) Possible (26,85%) Condition (9,29%)

	Condition (7,66%)	Hope (5,52%)
	Doxastic (5,18%)	Doxastic (4,24%)
	Dynamic (2,26%)	External (1,05%)
	Possible (1,72%)	

During the annotation process, the two works did not employ cues to detect assertions of uncertainty. Instead, they relied on the implicit semantics in microblogs. According to Li et al. (2018), 22.28% of uncertain statements in their social media dataset do not contain cue-words. This is a significant amount of data, and it suggests that annotation based purely on detecting these cues may overlook a significant number of uncertain microblogs. We also observe one key benefit of such classifications from the current investigation. Once the dynamic type has been differentiated from the others, it's interesting to explore just the belief types to create sub-types of this class. This suggestion paves the way to studies in fields such as religious beliefs and orientations, alternative medicine based on herbal medicine and traditional practices, election campaign predictions, marketing and customer preferences, and so on. Finally, we can claim that the classification of uncertainty in microblogs, and especially the classification of semantic uncertainties, which we have highlighted in this research, is yet underexplored.

5. Conclusion

With the rise of social media, there are an increasing number of text contents comprising a big number of informal or word-of-mouth terms. In terms of factuality, the quality of information on social media has become a big issue. In comparison to current uncertainty corpora, social media authors can write in any style they choose. Humans use language to communicate their thoughts, opinions, and judgments. The author's confidence in their statement can be shown through the author's manner of expression. In fields like health, finance, engineering, and many others, where mistakes can lead to misleading outcomes, knowing the level of confidence of a claim is critical. Although uncertainty corpora exist in several domains, there is no standard uncertain corpus in social media texts, and this aspect of microblogging has received little attention. Uncertainty is expressed differently on social media than it is in formal language. This issue should be addressed in every investigative activity, including research and information retrieval. To adapt to the peculiarities of social media, it is required to recognize uncertainty based on contextual uncertain semantics rather than conventional cues, and taking into account the sub-classes might give additional information for study on dealing with uncertainty in social media texts. As a suggestion taken from this review paper, we propose to consider the types of semantic uncertainties when analyzing sentiments in microblogs such as Twitter and when exploring opinions in general. Advances in AI and DL could be exploited to build efficient classification models.

References

Adel, H., & Schütze, H. (2016). Exploring different dimensions of attention for uncertainty detection. *arXiv preprint arXiv:1612.06549*.

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press, Oxford.

Al-Sabbagh, R., Girju, R., & Diesner, J. (2015, April). A unified framework to identify and extract uncertainty cues, holders, and scopes in one fell-swoop. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 310-334). Springer, Cham.

Basu, M., Ghosh, K., & Ghosh, S. (2020). Information Retrieval from Microblogs During Disasters: In the Light of IRMiDis Task. *SN Computer Science*, *1*(1), 1-10.

Bessarab, A., Mitchuk, O., Baranetska, A., Kodatska, N., Kvasnytsia, O., & Mykytiv, G. (2021). Social Networks as a Phenomenon of the Information Society. *Journal of Optimization in Industrial Engineering*, *14*(1), 35-42.

De Marneffe, M. C., Manning, C. D., & Potts, C. (2012). Did it happen? The pragmatic complexity of veridicality assessment. *Computational linguistics*, *38*(2), 301-333.

Diab, M., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V., & Guo, W. (2009, August). Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)* (pp. 68-73).

Díaz, N. P. C. (2013, September). Detecting negated and uncertain information in biomedical and review texts. In *Proceedings of the Student Research Workshop associated with RANLP 2013* (pp. 45-50).

Ebner, M. (2018). Microblogs. In *The SAGE Encyclopedia of the Internet* (pp. 640-641). Sage Publications, Inc..

Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010, July). The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the fourteenth conference on computational natural language learning–Shared task* (pp. 1-12).

Feng, R. J., Zhang, H. J., Pan, W. M., Zhou, Z. Y., & Li, Y. J. (2021). A New Method of Microblog Rumor Detection Based on Transformer Model. In *Artificial Intelligence in China* (pp. 531-537). Springer, Singapore.

Ghelichkhan, A., Nematizadeh, S., Saeednia, H. R., & Nourbakhsh, S. K. (2020). Optimal Use of Social Media From the Perspective of Brand Equity in Startups with a Data Approach. *Journal of Optimization in Industrial Engineering*, *13*(2), 149-163.

Han, X., Li, B., & Wang, Z. (2019). An attention-based neural framework for uncertainty identification on social media texts. *Tsinghua Science and Technology*, *25*(1), 117-126.

Hua, W., Huynh, D. T., Hosseini, S., Lu, J., & Zhou, X. (2012). Information Extraction From Microblogs: A Survey. *Int. J. Softw. Informatics*, *6*(4), 495-522.

Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65).

Jaworski, W., Rejmund, E., & Wierzbicki, A. (2014, August). Credibility Microscope: relating Web page credibility evaluations to their textual content. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (Vol. 1, pp. 297-302). IEEE.

Jordan, S. E., Hovet, S. E., Fung, I. C. H., Liang, H., Fu, K. W., & Tse, Z. T. H. (2019). Using Twitter for public health surveillance from monitoring and prediction to public response. *Data*, *4*(1), 6.

Kiefer, F. (2005). Lehetoseg es szuksegszeruseg [Possibility and necessity]. *Tinta Kiadó, Budapest*.

Kim, J. D., Ohta, T., & Tsujii, J. I. (2008). Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, *9*(1), 1-25.

Konstantinova, N., De Sousa, S. C., Díaz, N. P. C., López, M. J. M., Taboada, M., & Mitkov, R. (2012, May). A review corpus annotated for negation, speculation and their scope. In *Lrec* (pp. 3190-3195).

Kothandan, J., & Murugesan, P. (2021). ML based social media data emotion analyzer and sentiment classifier with enriched preprocessor. *Journal of Information Technology Management*, *13*(Special Issue: Big Data Analytics and Management in Internet of Things), 6-20.

Kumar, M., Garg, A., Munjal, A., & AkanshaTanwar, A. (2017). Twitter Based Information Extraction. *International Journal of New Technology and Research*, *3*(3).

Li, B., Xiang, J., Chen, L., Han, X., Yu, X., Xu, R., ... & Wong, K. F. (2018, May). The UIR Uncertainty Corpus for Chinese: Annotating Chinese Microblog Corpus for Uncertainty Identification from Social Media. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Li, B., Xiang, J., Chen, L., Han, X., Yu, X., Xu, R., ... & Wong, K. F. (2018, May). The UIR Uncertainty Corpus for Chinese: Annotating Chinese Microblog Corpus for Uncertainty Identification from Social Media. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Light, M., Qiu, X. Y., & Srinivasan, P. (2004). The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 workshop: linking biological literature, ontologies and databases* (pp. 17-24).

Medlock, B. (2008). Exploring hedge identification in biomedical literature. *Journal of biomedical informatics*, *41*(4), 636-654.

Medlock, B., & Briscoe, T. (2007, June). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 992-999).

Nawaz, R., Thompson, P., & Ananiadou, S. (2010, July). Evaluating a meta-knowledge annotation scheme for bio-events. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing* (pp. 69-77).

Palmer, F. R. (2001). *Mood and modality*. Cambridge university press.

Qazvinian, V., Rosengren, E., Radev, D., & Mei, Q. (2011, July). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1589-1599).

Owyang, J. (2008) Retweet: The infectious power of word of mouth. http://www.webstrategist.com/blog/2008/11/23/retweet-the-infectious-power-of-the-word-of-mouth/ (last access June 2021).

Rajabi, F., Saghaei, A., & Sadinejad, S. (2020). Monitoring of social network and change detection by applying statistical process: ERGM. *Journal of Optimization in Industrial Engineering*, *13*(1), 131-143.

Rouhani, S., & Abedin, E. (2019). Crypto-currencies narrated on tweets: a sentiment analysis approach. *International Journal of Ethics and Systems*.

Rubin, V. L., Liddy, E. D., & Kando, N. (2006). Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications* (pp. 61-76). Springer, Dordrecht.

Rubin, V. L. (2010). Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, *46*(5), 533-540.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Saurí, R., & Pustejovsky, J. (2009). FactBank: a corpus annotated with event factuality. *Language resources and evaluation*, *43*(3), 227-268.

Saurí, R., & Pustejovsky, J. (2012). Are you sure that this happened? assessing the factuality degree of events in text. *Computational linguistics*, *38*(2), 261-299.

Settles, B., Craven, M., & Friedland, L. (2008, December). Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning* (Vol. 1).

Shatkay, H., Pan, F., Rzhetsky, A., & Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, *24*(18), 2086-2093.

Szarvas, G. (2008, June). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of acl-08: HLT* (pp. 281-289).

Szarvas, G., Vincze, V., Farkas, R., Móra, G., & Gurevych, I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, *38*(2), 335-367.

Uzuner, Ö., Zhang, X., & Sibanda, T. (2009). Machine learning and rule-based approaches to assertion classification. *Journal of the American Medical Informatics Association*, *16*(1), 109-115.

Vincze, V. (2014). Uncertainty detection in natural language texts. *PhD, University of Szeged*, 141.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, *9*(11), 1-9.

Yan, Q., Wu, L., & Zheng, L. (2013). Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and Its Applications*, *392*(7), 1712-1723.

Wei, Z., Chen, J., Gao, W., Li, B., Zhou, L., & He, Y., et al. (2013). An Empirical Study on Uncertainty Identification in Social Media Context. *Meeting of the Association for Computational Linguistics* (pp.58-62).

Wilson, T. A. (2008). *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*. University of Pittsburgh.

This article can be cited:
Zendaoui, F., Hidouci, W., Rouhani, S. (2022). Uncertainty Identification in Microblogs. *Journal of Optimization in Industrial Engineering*, 15(1), 301-309.

http://www.qjie.ir/article_687477.html
DOI: 10.22094/joie.2021.1945486.1913