Annotation of Messages from Social Media for Influencer Detection

Kévin Deturck, Damien Nouvel, Namrata Patel, Frédérique Segond

Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 85–90
Marseille, 24 June 2022

Annotation of Messages from Social Media for Influencer Detection

Kévin Deturck, Damien Nouvel, Namrata Patel, Frédérique Segond
Inalco Ertim, Inalco Ertim, Université Montpellier 3, Inria Minatec – Inalco Ertim
2 r. de Lille 75007 Paris, 2 r. de Lille 75007 Paris, Rte de Mende 34090 Montpellier, 17 av. des Martyrs 38000 Grenoble
{kevin.deturck, damien.nouvel, frederique.segond}@inalco.fr
namrata.patel@univ-montp3.fr

Abstract

To develop an influencer detection system, we designed an influence model based on the analysis of conversations in the "Change My View" debate forum. This led us to identify enunciative features (argumentation, emotion expression, view change, ...) related to influence between participants. In this paper, we present the annotation campaign we conducted to build up a reference corpus on these enunciative features. The annotation task was to identify in social media posts the text segments that corresponded to each enunciative feature. The posts to be annotated were extracted from two social media: the "Change My View" debate forum, with discussions on various topics, and Twitter, with posts from users identified as supporters of ISIS (Islamic State of Iraq and Syria). Over a thousand posts have been double or triple annotated throughout five annotation sessions gathering a total of 27 annotators. Some of the sessions involved the same annotators, which allowed us to analyse the evolution of their annotation work. Most of the sessions resulted in a reconciliation phase between the annotators, allowing for discussion and iterative improvement of the guidelines. We measured and analysed inter-annotator agreements over the course of the sessions, which allowed us to validate our iterative approach.

Keywords: annotation, influencer, social media

1. Introduction

1.1 Research problem: influencer detection

An influencer is defined in sociology as a person having the power to change peoples' views or behaviour simply by interacting with them (Katz and Lazarsfeld, 2017). Social psychology analyses such an impact by describing interpersonal interactions as a set of stimuli that can lead to a psychological change in everyone involved (Turner and Oakes, 1986). We define the process of influence by interactions initiated by an influencer, leading to the production of new opinions or actions among the targeted individuals.

Recent years have seen an increasing interest in influencer detection as it helps identify key users within a large inter-personal network. Influential users are likely to express their ideas with a greater impact than other individuals, as seen in political (Katz, 1957), commercial (Trusov et al., 2010) or terrorist recruitment contexts (Fernandez et al., 2018).

Interpersonal interactions being the vehicle of influence, we choose social media as a ripe field of observation as they are inherent to its very structure. The development of social media has boosted research on many issues pertaining to artificial intelligence and its impact on society; the detection of influence being one of them.

1.2 Annotation requirements: development of an influencer detection system

Our study is centred around an influence model designed to characterise the process of influence (Deturck, 2021). As computational linguists, we follow our predisposition to analyse the textual content of conversations. Our goal is to detect the linguistic markers of influence we identified by analysing conversations in the "Change My View" debate forum[1]. The markers reflect the specific discourse of both (1) the influencers, initiating the influence process, and (2) the individuals reacting to the influencers.

To develop an influencer detection system based on our model, we needed reference data to (1) develop linguistic rules, train models by learning and (2) evaluate the different modules of the system. As our model features original linguistic markers, we had to produce the corresponding reference data by supervising human annotators through successive annotation sessions.

Our annotation task corresponds to the *unitizing* type (Krippendorff, 1995). A unitizing annotation consists in extracting units by segmenting a text and categorizing the resulting segments. In our case, it is a matter of identifying, in social media messages, the text segments that correspond to one of our linguistic markers of influence.

The task is particularly difficult because annotators must identify both the relevant text boundaries and the corresponding category. In addition to that, the text segments are not necessary nor usually on sentence boundaries, they can be sub-sentence or super-sentence level spans.

The annotation task is also particularly difficult because it requires the identification of linguistic markers which involve interpretation of statements: on the one hand, each annotator must manage to do this interpretation work, which is complex, and, on the other hand, we must achieve consistent annotation across through the interpretations of the different annotators to build a reference corpus.

The rest of the article is organised as follows: in section 2 we introduce our influence model, in section 3 we present the annotation schema, in section 4 we describe the data, in section 5, we present and analyse the results of the annotation campaign, then we conclude in section 7.

[1] https://www.reddit.com/r/changemyview/

2. Influence Model

The model we present in this section is in line with works in social psychology, such as the one by Mason et al. (2007), and communication science, for example the one by Dillard and Wilson (2014). It describes influence as a process with source individuals impacting the minds of target individuals through the exchange of messages.

Our model contains three components: the *stimulus* and *stimulation* components correspond to a theoretical framework in social psychology, described by Turner and Oakes (1986), which gives an individual's social environment as a carrier of *stimuli* that can *stimulate* (or modify) the psychological state of the individuals in it. The *decision* component relates to the decision-making process, particularly studied in social psychology, as in the work by Ajzen (1996); the impact on decision-making is the conclusion of the influence process in our model.

3. Annotation Schema

3.1 *Stimulus* Linguistic Markers

3.1.1 Claim

A claim is a type of expression by which an individual delivers a description as factual, i.e. an assertion of what is allegedly a fact in the world (Saurí and Pustejovsky, 2012). A claim can be factual only in appearance, i.e. it can make a concrete description with certainty without it being true.

Example: "*#ISIS has showered Ayn al-Asad airbase"*, in a tweet from the "pro-Islamic State" dataset used for the annotation campaign (cf. section 4.1).

3.1.2 Pedagogy

The linguistic marker *Pedagogy* is the statement of an individual who guides other individuals in their understanding of the world or their behaviour in the world. This type of discourse is based on advices and explanations. Pedagogy had already been identified by Dillard and Wilson (2014) as having a link to influence.

Example: "*Turn it off so they can stay in the darkness of their misguidance."*, in a tweet from the "pro-Islamic State" dataset.

3.1.3 Argumentation

Argumentation is a type of discourse that consists of supporting the truthfulness of a statement with one or more logically articulated arguments (Eckle-Kohler et al., 2015). Example: "*It appears that ISIS are the best diplomats on Earth since they work for Iran, America, Turkey, Saudi and Israel"*, from the "pro-Islamic State" dataset used for the annotation campaign (cf. section 4.1).

3.2 *Stimulation* Linguistic Markers

3.2.1 Understanding

Understanding is manifested in the discourse of an individual reporting on the reasoning they have managed to produce through a message. This type of expression links to research in social psychology which considers the process of understanding a message as an important factor for the impact of communication (Wyer and Shrum, 2015).

Example: "*Yours was the first comment to make me understand how changing the definition would render the word useless"*, a participant in the "Change My View" forum

3.2.2 Information

Information acquisition appears in any utterance where the enunciator indicates receiving new information. Information acquisition corresponds to a stimulation of the intellect (Hidi and Baird, 1986).

Example: "*I realised that i was misinformed when it came to Duty to Retreat laws*", a participant in the "Change My View" forum

3.2.3 Affectation

Any reaction relating the experience of a feeling or emotion by the enunciator, in relation to the enunciation situation. The influence of affect on decision making is a research topic (Binali et al., 2010).

Example: "*You gave me some hope for the oils"*, a participant in the "Change My View" forum

3.2.4 Agreement

An utterance in which the speaker posits an equivalence between his or her viewpoint or actions and the viewpoint or actions of others, to whatever degree. Agreement is studied in relation to individuals' decision making (Germesin and Wilson, 2009).

Example: "*I do agree that the left has similar issues"*, a participant in the "Change My View" forum

3.3 *Decision* Linguistic Marker: Change of Mind

Change of mind is the purpose of an influencing action in the "Change My View" forum. We identify the expression of a change of mind with any statement in which the speaker indicates a questioning or evolution of his or her opinion, to whatever degree.

Example: "*I won't continue with the position I stated I'm my last comment"*, a participant in the "Change My View" forum

4. Methodology

4.1 Material

An annotation guide was designed to drive and facilitate the annotation process. It is a 24-page PDF document that provides definitions supplemented with examples and counter-examples for each of the markers (Deturck, 2021). This document was revised after each annotation session, based on post-annotation meetings between and with the annotators, to iteratively refine the marker definitions, an *agile* corpus annotation (Voormann and Gut, 2008).

For the variety of our reference corpus, we used two complementary data sources in English: the "Change My View" debate forum, in which the authors must elaborate on their views, and a corpus of tweets, constrained to a limited number of characters, posted by individuals categorized as supporters of the "Islamic State of Iraq and

Syria" (ISIS) organisation[2]; we used the latter only for the *stimuli* markers as it was not designed to provide reactions to the pro-ISIS' tweets.

We partitioned the data to distribute it among annotator groups and thus maximise the quantity of messages annotated during a session by including several groups. To simplify the annotation and thus promote its quality, we made sure that each dataset contains only one kind of message ("Change My View" or Twitter).

We sized each dataset so that it could be processed by a single annotator in a maximum of two hours, which is the duration imposed for session. We empirically estimated the annotation time for a single message according to its textual genre (a tweet or a forum post): 45 seconds for a tweet and 80 seconds for a forum post. This led us to create "Twitter" datasets containing 100 messages and "Forum" datasets containing 80 messages.

We used Gate software (Cunningham et al., 2013) as an annotation tool. This software provides a graphical interface for selecting portions of text and assigning a label, which allowed us to use it as is for our "unitizing" annotation task.

4.2 Annotators and Sessions

We organised five annotation sessions with non-native English speaking NLP students as annotators (see Table 1). It is not a concern that for all annotators English is not a native language, they can still understand enough the documents to correctly annotate it. In each session, the annotators were divided into groups of two or three and each group was given one dataset to annotate.

Session	Annotators	Datasets	Markers
Session 1	7 duos	5 "Twitter", 2 "Forum"	Claim
Session 2	5 duos, 1 trio	4 "Twitter", 2 "Forum"	Stimuli
Session 3	2 duos	2 "Twitter", 1 "Forum"	Stimuli
Session 4	2 duos	1 "Twitter", 1 "Forum"	Stimuli
Session 5	2 duos	2 "Forum"	Stimulation, Decision

Table 1: Annotation session configurations

The annotators in sessions 1 and 2 were completely different, whereas sessions 3 to 5 were held with four annotators who had already worked in session 2. Each group in a session annotated a different dataset; in session 3, the group that annotated the "Twitter" dataset had time to annotate one more while the other group was annotating a "Forum" dataset.

With the same objective of simplifying the task and thus improving the annotation quality as for the choice of one genre per dataset, we had annotated a subset of the markers per session.

Most of the sessions are focused on the markers used by influencers, *claims* and more broadly *stimuli*. For these sessions, "forum" datasets contain only messages from participants who are not the initial authors of discussions: in the "Change My View" debate forum, discussions are initiated by participants who expose their point of view on a topic of their choice, then, the other participants have to change the initial participants' mind and be influencers.

Session 1 focuses on *claims* because the annotation guide was written only for this marker at this point in the campaign.

Session 5 is the only session dedicated to *reaction* markers (*Stimulation* and *Decision*). Only forum messages are used for this session as it is the only resource that presents the reactions to the messages. Also, we selected for the datasets only the messages of the initial authors of discussions because, in the "Change my view" forum, they are the ones that must be influenced by other participants in a discussion, what we want to detect in their reactions.

At the end of each annotation session, we organised a *reconciliation* phase: each group of annotators discussed their disagreements (the text segments they did not annotate identically) to reach a single annotation set that could be used in the gold corpus. Finally, the conflicts were discussed together in a final phase, allowing us to update the annotation guide for future sessions.

As our annotation task is particularly subjective, we think that this reconciliation process, as it integrates different judgements, allows to achieve a relative objectivity and thus a better reference (Bonin et al., 2020). We can nevertheless question the limits of this objectivity, which may only be local, reconciliation leading to overtraining (Hovy and Lavid, 2010), limited in our case by the small number of sessions shared by the same annotators.

5. Results

5.1 Quantitative Synthesis

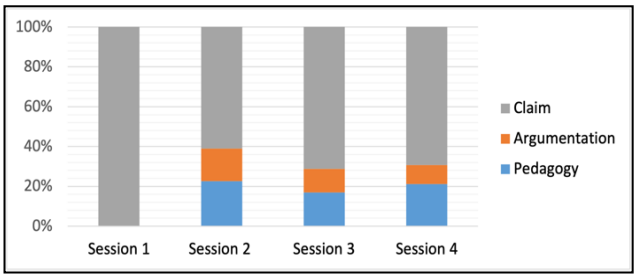

Figure 2: Pro-ISIS tweet annotation distribution

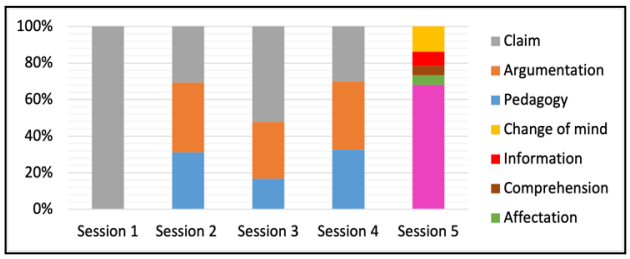

Figure 3: "Change My View" forum annotation distribution

[2] https://www.kaggle.com/datasets/fifthtribe/how-isis-uses-twitter

We compare the volumes of annotated marker types between the Twitter and forum datasets, respectively represented in figures 2 and 3.

To characterise the *stimuli* across the two textual genres, we can notice that forum messages contain, for two out of three sessions (sessions 2 and 4), a majority of argumentation. This shows a reasonable characterization of the authors' attempts to influence the debate forum and thus defend one's opinions. Tweets tend to contain more claims than forum messages, which corresponds well to the particularly brief nature of tweets.

The distribution of *stimulation* markers (Figure 3) shows a large predominance of *Agreement;* this is a reasonable response to the predominance of *Argumentation* among *stimuli* markers because agreement is an alignment of opinions while argumentation is used to support an opinion. The predominance of *stimuli* markers over the *decision* ones (see Figure 3) shows that it is rare for an influence process to reach its conclusion.

A gold dataset was created only for the *stimuli* markers, on the one hand because we did not have time to develop for *stimulation* detection and on the other hand because we performed *change of mind* (*decision*) detection by using as reference the "delta" system in the "Change My View" forum (Deturck, 2021): when initial authors of discussions change their mind because of messages, they have to cite them with a new message including a "delta" symbol and an explanation of their change of mind, then, an automatic moderation validates or not the delta.

We present in Table 2 the number of annotated messages in the gold dataset per marker, with the percentage of these messages that contain at least one occurrence of the marker.

Marker	Number of annotated messages	% of messages containing the marker
Claim	1126	45%
Pedagogy, Argumentation	716	14% for *pedagogy*, 7% for *argumentation*

Table 2: Message volumes by marker in the gold dataset

Quantity differences among markers are directly related to the session configurations (see Table 1): one more session was dedicated to *claim* annotation, also, tweets are more represented than forum messages, which explains the higher proportion of *pedagogy* compared to *argumentation*.

5.2 Qualitative Synthesis

5.2.1 Inter-annotator Agreement

Since it is argued that an annotation is more reliable if it is reproduced by several annotators (Krippendorff, 2004), we measured inter-annotator agreement. Two measures have been specifically designed for unitizing annotation tasks: the Alpha family (Krippendorff et al., 2016), and the Gamma family (Mathet, 2017). Alpha measures cannot be applied to annotations containing overlapping entities, as may be the case in our annotation task. We will therefore use Gamma measures.

We use two coefficients in the Gamma family: the standard Gamma coefficient, which takes the location and categorisation of annotations into account, and the GammaCat coefficient, which focuses on the categories associated with the selected units. This allows us to distinguish between two forms of disagreement: (1) a confusion between categories or (2) differing boundaries of relevant text.

	Gamma score	GammaCat score
Session 1	0.38	N/A
Session 2	0.35	0.53
Session 3	0.48	0.7
Session 4	0.62	0.88
Session 5	0.71	0.91

Table 3: Average inter-annotator agreement scores

Table 3 shows the Gamma inter-annotator agreement measures for each session. These results were calculated by averaging the scores of all groups in a session. We present sessions 2 to 4 in a different colour because they are fully comparable in terms of the annotated categories (the *stimuli* ones).

We observe an interesting improvement in results between sessions 2 and 4, both for Gamma and GammaCat. These three sessions were specifically designed using the same traits to evaluate the annotation progression. This improvement confirms the relevance of our iterative approach, especially as regards improving the annotation guide.

Overall, we notice that the GammaCat coefficient gives much better results than the Gamma coefficient. We can therefore conclude that the disagreement measured is mainly due to a problem in delimiting the units rather than to a difficulty in identifying the presence of categories in the messages. This is a positive result for the use of annotations since the units found, even not exact in their boundaries, are consistent with the defined categories.

5.2.2 Annotation Mistakes

Error type / Expected	*Claim*	*Pedagogy*	*Argumentation*
Claim confusion	N/A	20%	17%
Pedagogy confusion	25%	N/A	25%
Argumentation confusion	2%	16%	N/A
Delimitation error	49%	36%	32%
Out of the scope	24%	28%	26%

Table 4: Statistics on error types regarding *stimuli*

We manually identified the "mistakes" made by annotators, that is those annotations, among disagreements, that contradict the guidelines. It is a necessary step to determine annotation difficulties and improve the annotation guide.

Besides confusion between markers, we distinguished between two error types that we describe below.

- *Delimitation error*: boundaries incorrect, but semantics are valid, for example, the two claim annotations in "[[Most to all mass shootings in the US are where carrying guns is banned]1 (for the laws abiding)]2,
- *Out of the scope*: semantics are not valid; it is a critical error, for example, *"These types of calculations aren't helpful"* is *out of the scope* because it is a judgement alone, without argumentation

We present the distribution of these error types for *stimuli* markers (see Table 4), which constitute a significant part of the annotations. A large proportion of annotation errors relates only to the delimitation of units. This is a relatively positive observation as regards the quality of the annotations since annotations of this type still contain relevant statements.

Confusion between marker types is important due to similarities: pedagogical discourse may contain claims, pedagogy explains a fact and argumentation explains a point of view. *Out of scope* errors are globally in a minority; they are mainly due to the difficulty of distinguishing factual from viewpoint statements.

6. Conclusion

We have described an annotation campaign organized as part of the development of a system to detect influencers. The annotation schema is composed of linguistic markers corresponding to our influence model.

The annotation task was particularly difficult, on the one hand because the linguistic markers involved the interpretation of statements and on the other hand because it required annotators to precisely identify the text segments that corresponded to each marker. To deal with this difficulty, we chose to design an iterative annotation campaign, involving multiple annotation-revision cycles.

Inter-annotator agreement measures throw different annotation sessions showed that our method allowed to build a relative consensus. It may be a validation of our approach to get reliable annotations, but it may also reflect overtraining due to the reconciliation phases. The resulting *gold* annotations have been used to train models that we applied for influencer detection.

7. Bibliographical References

Ajzen, I. (1996). The social psychology of decision making. Social psychology: Handbook of basic principles, 297-325.

Binali, H., Wu, C., and Potdar, V. (2010). Computational approaches for emotion detection in text. In *4th IEEE international conference on digital ecosystems and technologies* (pp. 172-177). IEEE.

Bonin, F., Finnerty, A., Moore, C., Jochim, C., Norris, E., Hou, Y., ... and Michie, S. (2020). HBCP corpus: A new resource for the analysis of behaviour change intervention reports.

Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology*, *9*(2), e1002854.

Deturck, K. (2021). *Détection des influenceurs dans des médias sociaux* (Doctoral dissertation, Institut National des Langues et Civilisations Orientales-INALCO PARIS-LANGUES O').

Deturck, K. (2021). *Guide d'annotation en discours pour la détection d'influenceurs* (Doctoral dissertation, Institut National des Langues et Civilisations Orientales).

Dillard, J. P. and Wilson, S. R. (2014). Interpersonal influence. *Interpersonal communication*, *6*, 155.

Eckle-Kohler, J., Kluge, R., and Gurevych, I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 2236-2242).

Fernandez, M., Asif, M., & Alani, H. (2018). Understanding the roots of radicalisation on twitter. In Proceedings of the 10th ACM conference on web science (pp. 1-10).

Germesin, S. and Wilson, T. (2009). Agreement detection in multiparty conversation. In *Proceedings of the 2009 international conference on Multimodal interfaces* (pp. 7-14).

Hidi, S. and Baird, W. (1986). Interestingness—A neglected variable in discourse processing. *Cognitive science*, *10*(2), 179-194.

Hovy, E. and Lavid, J. (2010). Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, *22*(1), 13-36.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public opinion quarterly*, *21*(1), 61-78.

Katz, E. and Lazarsfeld, P. F. (2017). *Personal influence: The part played by people in the flow of mass communications*. Routledge.

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology*, 47-76.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and quantity*, *38*, 787-800.

Krippendorff, K., Mathet, Y., Bouvry, S., and Widlöcher, A. (2016). On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, *50*(6), 2347-2364.

Mason, W. A., Conrey, F. R., and Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review*, *11*(3), 279-300.

Mathet, Y. (2017). The Agreement Measure γ cat a Complement to γ Focused on Categorization of a Continuum. *Computational Linguistics*, *43*(3), 661-681.

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

Strötgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3746–3753, Istanbul, Turkey, May. European Language Resource Association (ELRA).

Trusov, M., Bodapati, A. V., and Bucklin, R. E. (2010). Determining influential users in internet social networks. *Journal of marketing research*, *47*(4), 643-658.

Turner, J. C. and Oakes, P. J. (1986). The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, *25*(3), 237-252.

Voormann, H. and Gut, U. (2008). Agile corpus creation.

Wyer Jr, R. S., and Shrum, L. J. (2015). The role of comprehension processes in communication and persuasion. *Media Psychology*, *18*(2), 163-195.