Modeling Argumentation Framework for Twitter Private Opinion Accounts

2010127 Zhong Jiaqi

Nowadays, it is convenient for users of social medias to express their thoughts and opinions. Among many tools, Twitter is one of the most widely used and has open APIs which allow developers to extract large amounts of data. However, when users want to obtain some information to make decisions, there is no way to distinguish the reliability of what other users post.

In Japanese, there is a special expression - private opinion (本音) and public sound (建前). Private opinion means one's real opinion, public sound means one's public stance. Private opinion is a method that does not make the other person uncomfortable by returning it in bland words without saying something that is unpleasant in the heart, and it is regarded as common sense in Japanese society. It's hard to judge public sound as the lie, but it's still a persistent treatment for Japanese people whose virtue is to read between the lines. In order to really understand the real thoughts of Japanese, we can focus on accounts that only post private opinions. These special accounts do exist among Japanese netizens, which are called private opinion accounts (本音垢). At the same time, there are also accounts posting vicious words which are more straightforward than the private opinion accounts, which called the venom accounts (毒舌垢). Obviously, we can judge that private opinion has a higher degree of reliability than public sound, therefore we choose to crawl and analyze the private opinion accounts and the venom accounts in the research. Our research can be viewed as a development of XAI and reliable AI, which is in very high demand in recent years.

Our objective is to identify private opinion structure in Twitter data for modeling argumentation graphs with attack relations. Our research based on the Argumentation Framework (AF), which is introduced by Dung, are pairs consisting of a set $\mathcal{AR}$ of arguments and a binary relation between arguments, representing attacks. Formally, an AF is any $\langle \mathcal{AR}, attacks \rangle$ where $attacks \subseteq \mathcal{AR} \times \mathcal{AR}$. Bench-Capon defines the valued-based argumentation framework (VAF) by attaching to each argument the social values that it promotes. Formally, a VAF is a 5-tuple $\langle \mathcal{AR}, attacks, R, val, valpref \rangle$, where R is a non-empty set of values, val is a function which maps from elements of $\mathcal{AR}$ to elements of R, and $valpref$ is a preference relation on $R \times R$.

In our research, we extend the notion of VAF to introduce a Weighted Annotated Discussion Graph (WADisG). Let Γ be a non-empty set of tweets and $G = \langle V, E, A, R, W \rangle$ be a WADisG; for every tweet in Γ there is a node in V and if tweet a attacks tweet b, there is a directed edge (a, b) in E, A is an annotation function for edges $A : E \to S$, where S represents

attack relations for any directed edge (a,b), and the value's range of S is {attacks,none}. We defined the valued-based argumentation frameworks for as $F = \langle V, attacks, R, W, Valpref \rangle$, where the weighting function for arguments is $W : V \rightarrow R$ for tweets, and the preference relation $Valpref \subseteq R \times R$ is the ordering relation over R. We give weight to arrows rather than nodes because our data set is special compared with typical twitter conversation analysis. In a typical twitter conversation, a tweet usually has many replies from different users, which is suitable for the traditional VAF that gives weight to different nodes (arguments). In our dataset, users rarely receive a reply to a tweet, but they usually quote the arguments of others in a tweet before giving a counter-argument. Therefore, we divide such tweets into sub tweets and annotate the sub tweets with attack relation. The weight of the same user on a topic is calculated by the weighting function therefore the node weights at both ends of the attack are the same. We give weights to different arrows to represent the overall reliability of the tweet. In the proposed WADisG, the grounded extension $S \subseteq T$ of F is the accepted set of tweets based on the weighting scheme W and we refer to it as the solution of G, i.e. the set of tweets with high reliability.

We crawled the data from Twitter API and have done some processing work such as format conversion, solving the garbled code problem, cleaning our data, and annotate for the tweets. After we finish processing the tweets, we carry out some basic morphological analysis work, such as word segmentation to get word frequency. Then, we combine TF-IDF and Japanese grammar (refer to Japanese papers) to select the most useful features for training our model. We carried out two binary classifications, argument classification and attack classification. Later, we use some commonly used machine learning algorithms on our dataset and evaluate the performance of these algorithms based on the confusion matrix and F Score. We found that Multinomial Naive Bayes performs the best and Passive Aggressive the worst in our classification experiments. In the end, we use answer set programming to calculate the set of reliable tweets and visualized with argumentation graph.

Finally, we analyze the structure of private opinion from morphological analysis and syntactic analysis aspect. We look forward to some future work such as making a more complete information retrieval system, including but not limited to adding a more friendly interactive interface, adding indicators from other users such as Net Promoter Score, using the threshold setting in order to provide users in need with a reliable decision-making reference tool.