Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media

Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media

ABEER ALDAYEL*, School of Informatics. The University of Edinburgh, Edinburgh, UK
WALID MAGDY†, School of Informatics. The University of Edinburgh, Edinburgh, UK

To what extent user's stance towards a given topic could be inferred? Most of the studies on stance detection have focused on analysing user's posts on a given topic to predict the stance. However, the stance in social media can be inferred from a mixture of signals that might reflect user's beliefs including posts and online interactions. This paper examines various online features of users to detect their stance towards different topics. We compare multiple set of features, including on-topic content, network interactions, user's preferences, and online network connections. Our objective is to understand the online signals that can reveal the users' stance. Experimentation is applied on tweets dataset from the SemEval stance detection task, which covers five topics. Results show that stance of a user can be detected with multiple signals of user's online activity, including their posts on the topic, the network they interact with or follow, the websites they visit, and the content they like. The performance of the stance modelling using different network features are comparable with the state-of-the-art reported model that used textual content only. In addition, combining network and content features leads to the highest reported performance to date on the SemEval dataset with F-measure of 72.49%.

We further present an extensive analysis to show how these different set of features can reveal stance. Our findings have distinct privacy implications, where they highlight that stance is strongly embedded in user's online social network that, in principle, individuals can be profiled from their interactions and connections even when they do not post about the topic.

Additional Key Words and Phrases: Social Media, Opinion mining, Stance detection

ACM Reference Format:
Abeer ALDayel and Walid Magdy. 2019. Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media. 1, 1 (August 2019), 20 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

This is a preprint of an article accepted for publication by CSCW 2019.

1 INTRODUCTION

Stance towards a given topic is the position towards this topic either being in-favour or against it [10]. Recently, large attention has been directed to automatic stance classification (detection) because of its wide range of applications, especially in the field of social media analysis. Earlier work focused on stance detection on argumentative debates in Online-forums [30, 43, 47]. With the wide spread of social media platforms, such as Twitter, which have become a common place for users to share their opinions towards various topics, research has been directed towards stance detection on these platforms. Detecting stance has widespread applications in social media analysis, opinion evolution, polarization detection, and rumours detection [24, 58]. Many studies used stance

Authors' addresses: Abeer ALDayel, a.aldayel@ed.ac.uk, School of Informatics. The University of Edinburgh, Edinburgh, UK; Walid Magdy, wmagdy@inf.ed.ac.uk, School of Informatics. The University of Edinburgh, Edinburgh, UK.

XXXX-XXXX/2019/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

detection to analyze social media as main component of investigating the users aligns toward a given topic or entity [3, 7, 33, 34, 36, 40].

Most research on stance detection modeled stance as a text classification task, where text of on-topic posts are used as the features [21, 40, 46, 55]. Some other work showed the effectiveness of using user's network as the features [14, 33, 34, 36]. However most of these studies were focused on one topic with no real examination to its generality on other topics or domains. Another limitation of the existing approaches for stance detection is the reliance on signals from active users only who frequently post on social media, where user's stance is modelled either by user's posts or interaction with other users (retweet in case of Twitter). There has been a growing interest on characterizing "silent user" in social media platforms [8, 27]. This group of users known as "lurkers" or "invisible participants" tends to contribute with a little or no content. Some users prefer to interact quietly on social media using other means of interactions instead of directly posting or sharing contents, such as following others and liking posts [27]. Most of stance detection studies used the network representation of the active users only and overlooked the silent users [14, 34, 36].

Du Bois [20] argues that stance taking is a subjective and inter-subjective phenomenon in which stance-taking process is affected by personal opinion and non-personal factors such as cultural norms. Stance taking is a sophisticated process relates to different personal, cultural and social aspects. For instance, a political stance taking depends on experiential behavior as stated by [37]. Thus users in social media might express their opinion directly by posting about the topic or their stance could be inferred indirectly through their interactions and preferences. Our hypothesis is that user's embedded viewpoint in a post is related to the user's identity which could be better modeled by their interactions and connections in the social network. This idea is related to the concept of homophily in which users with same believes tend to have common interests and group together [2, 14, 23].

In this paper, we apply an extensive analysis to the possible online signals that can reveal the user's stance. To that end, we examine four groups of signals that might indicate the stance, namely: 1) on-topic posts by the user, which models users who explicitly express their stance on a topic; 2) user's interactions on social media with other users or websites, which models users interactions online regardless having them expressing their stance or not (IN); 3) user's preferences the posts they like, which enable modeling silent users who do not post or share content only (PN); and finally 4) the network of users they are connected, which enable modeling passive users who might have no content or interaction on social media, but just follow other accounts online (CN). We compare the effectiveness of each of these groups of features on detecting stance individually and when combined. Our main research question is to understand "What are the factors that can reveal the stance of user online towards a given topic". We further analyse "how" and "why" these factors might be effective for detecting stance. Our list of research questions in this paper are:

- What are the different signals in user's online activity that can reveal their stance, including textual content, networks of interaction (IN), preference (PN), and connection (CN)?
- Does the performance of detection differ by different types of topics?
- What makes any of these signals effective (or ineffective) for detecting stance?

Our experiments are applied on the SemEval stance detection benchmark dataset [40], which contains a set of over 4,000 tweets labeled by stance towards five different topics. The five topics covers multiple domains not just politics, which makes the dataset ideal to examine the generalisability of the stance detection models, unlike most of work in literature that typically focus on studying one political topic at a time [14, 33, 34, 36]. Our results show that training a classification model on pure user network features outperforms the state-of-the-art baseline system [40] which is trained on multiple features extracted from the tweets text content. This includes when using the

preference network features from only the tweets the user likes and also the connection network of the accounts the user follow, where both can model silent users. When different groups of features are combined, including content and network, a significant improvement is observed. Our findings suggest that for the task of stance detection, even when applied on the level of tweet, user's network information are more effective features than the content of the tweet itself. This aligns to the sociolinguistic theory in [6], where it defines stance as the link between linguistic forms and social identities which has the capability to establish the alignment between stance-takers.

We further applied an extensive analysis to the most influential features for each group of network signals to understand how they outperform textual text. It was interesting to find that the overlap between IN, PN, and CN was not large, where the common nodes among them are around 10% only, however, each of those networks still can model user's stance towards a given topic. Our analysis to the most influential features from each network on each of the five topics shows that there is usually some common signals in user online activity that can reveal their stance towards a given topic regardless of the type of the topic. We believe that our findings in this study raises a large concern about protecting the privacy of social media users, where their beliefs and leanings could be easily predicted using any of the footprint signals they leave online. This should motivate social media networks owners and designers to develop methods for protecting the privacy of their users [51].

The collected network information for the SemEval dataset would be made publicly available to allow replication to our experimentation[1].

2 RELATED WORK

There is a considerable amount of work on viewpoint or stance detection; yet, less work compared the role of content and social actor interactions in stance detection [14, 31]. Studying stance needs to cover the intersection dimensions of stance taking process, which are mainly influenced by linguistic forms and social interactions frames [37]. Most of the previous studies define stance as a textual entailment task where the main processing depends on the raw text only [5, 16, 40, 42, 45]. In this form of stance detection, a given text entails a stance towards a premise (target).

It has been shown that constructing a knowledge based dataset about the topic is beneficial in stance detection task [40]. This constitutes a visible hurdle which limits the stance detection task to set of predefined topics. Furthermore, many times the topic is not mentioned in the tweet. One way that was suggested to handle the unmentioned target entity in text is to analyze the opinion to the opponent of the entity or supporter of the entity. For example, [17] constructed a list of keywords that identifies Trump using a dataset labeled with stances toward Hillary. Using this list of keywords help in detecting the unexpressed stand towards Trump. Another study [45] follows the same line by constructing corpus that contains words that are *against* and *in-favor* each target to enrich the models. Similarly, [52] used a domain corpus related to Trump along with lexicon to construct a labeled dataset to detect stance towards Trump. Furthermore, [7] used context of the users tweets to construct author embedding and predict the stance.

There has been some work on studying the integration of network and content with a limited focus on the ideological political views [15, 31, 33, 36]. For instance the study of [31] focused on the liberal and conservative on twitter. Unlike previous work, rather than studying the stance on single topic and using a domain specific data, we study the stance in various domains. This study explores the stance modeling in the social media to know to what extent do network interactions and content interactions reveal an individual's viewpoint. Examining the implications of those

[1]https://github.com/AbeerAldayel/Stance_detection

interactions in detecting users' stances provides a better understanding of stance modeling on social media. In the following we summarise the literature work on stance detection.

2.1 Stance Detection on Twitter

The task of detecting stances takes a way back, focusing on online debates in online forums [35, 43, 50]. With the widespread of social media, they soon become a rich source of argumentative data, which has attracted many researchers to study stance detection on these platforms. As these platforms foster the real-time engagements with the new events, many studies used data collected from social media to predict people stances towards different topics [28, 36, 56]. For instance, the study done by [56] designed a stance detection model using YouTube's comments data.

Over the last decade, Twitter has become the most commonly used platform to study the expressed stance towards various events/topic [18, 22, 33, 36]. This platform featured to be open and capable to reach a significant proportion of audience. As a social media platform, the network structure in Twitter has a profound existence through various features provided within this platform. These features have made Twittersphere an attractive source of data to study stance and detect opinions toward a broad range of topics in real-time. In this platform users can connect and interact with each other directly through reply, retweet or mention. The retweet interactions considered a asymmetric interaction. In this kind of communication the user can retweet a tweet without the author acknowledgment. In contrast, the reply takes a form of symmetric communication where both users are involved in the process of interaction [39]. Basically, each user has a home timeline shows a stream of Tweets from accounts the user have followed on Twitter. Within this home timeline, user can reply, retweet, or like a Tweet from within the timeline. The collection of liked tweets for each user is shown in "Likes" (sometimes referred to "Favourite") timeline. The Likes timeline include only public liked tweets. If a user liked a "private" tweet of protected account they follow, it will not show up in their Likes timeline. Beside the endorsement and interactions, Twitter users can have a set of "Followers" and "Friends". The "Friends" collection contains accounts that a specific Twitter user follows. The Friends and Followers networks have been used effectively in previous studies to capture the social ties [12, 54].

2.2 SemEval Stance Detection Task

One of the well known stance dataset derived from Twitter is the SemEval stance dataset. This dataset is designed for supervised stance detection (task A) [40]. The dataset contains a (topic,tweet) pair for five topics covering political, social, and religious domains. Over 4000 tweets are released in this dataset, each labeled with stance as favor, against, or none to one of the five topics.

19 teams have participated in the task and submitted different models for stance classification. Most of the participants developed models that learn linguistic cues from the given tweet text to identify the stance for the target [21, 40]. Others used text representation methodologies such as LSTM conditional encoding to represent tweet-target pairs [4]. One team, MITRE [55], obtained a result with overall F-score of 67.82% by using two recurrent neural networks (RNN) classifiers. Another team, Pkudblab [53], achieved 67.33% F-score by utilizing convolutional neural network (CNN). While most of these approaches used various methods for creating an effective stance classifier, at the time of the competition, the best reported system used a simple character and word n-grams representation for the tweet text to train a linear SVM model, which achieved an average F-score of 68.98% [40].

Several studies have been published later on the same SemEval dataset reporting similar or marginal improvements to the performance while continuing to use various representations of the text to train different machine learning models. [19] proposed attention-based neural network by using a target-specific information which produced an overall F-score of 68.79%. Another marginal

improvement introduced by [16] by using attention based LSTM model, which achieved 68.84% F-score. Whereas in [57] the usage of bi-directional GRU-CNN yielded an F-score of 69.42%.

To the best of our knowledge, the current highest reported performance on this dataset is by [46] who trained multiple SVM models on only two classes (against and favor), while neglecting the "none" class. Forcing the classifier to predict a polarised stance led to the highest reported result on this benchmark dataset with F-score of 70.03.

2.3 Network Features to Detect Unexpressed View

Another stream of research on social media analysis has shown the large split in the networks of online social media [24, 25]. This has been analyzed as a reason to the social phenomenon of homophily [2, 9] that states the fact of users with similar believes tend to interact with each other, which creates what is so-called echo-chambers in online social networks [13, 29]. Few studies have utilized this phenomenon as a feature to predict unseen views of social media users on topics that they never discussed [14, 34, 36]. This assumption is powered by [32] in which she describes stance as non-transparent act in the text and must be inferred from the empirical study of interactions. Different sets of user features have been introduced in the previous studies with the focus of defining similar users for specific events. For instance [36] utilized user-interaction to predict users who would share hate-speech against Muslims after Paris attacks in 2015 using other accounts that a user mention, retweet, and reply to. They reported that using user's network interactions can predict the user's future attitude towards Muslims with an accuracy of 88% even when the user never discussed any topic related to Muslims before. Another work by [14], proposed a user similarity measure that is based on the distance between users in a social network to predict the users' stance towards two different topics. Similarly Lai et al. integrated network features to study users political leaning in the US elections 2016 and the Italian political debates [34].

Further work studied the users interactions as a factor to model the stance in the social media. The work of [49] used graph partitioning method with social interactions to cluster the users based on their viewpoint. Furthermore the study of [48] used the interactions between users with focus on the retweet and reply network as way to cluster the users with the same views. Similarly, [22] used the users interactions and the textual features to model the users stance by proximity graphs.

These studies highlights the importance of social network interaction of users to detect their position towards specific events or entities. Nevertheless, they are limited to focusing on one specific topic from the political domain, which lacks examining the generalisability of these approaches on multiple topics from different domains. In addition, it focuses on network interactions that can only model active users who retweet, mention, and reply other accounts.

In this paper, we utilise the SemEval benchmark dataset to apply an extensive comparison on stance detection using multiple sets of features and compare it to the state-of-the-art. In addition, we introduce the use of the preference network as a new way to model the stance and examine the possibility of detecting the stance of the silent users. We compare the performance of this new set of feature with content-based and other networks based features for stance detection.

3 STANCE DETECTION METHODOLOGY

In the following, we discuss our proposed methodology including the set of features used and the machine learning method applied. But initially, we discuss the implications of our approach from the conceptual point of view.

3.1 User vs Tweet Level Stance Detection

The SemEval dataset is labeled for stance on the tweet level, while we are examining user-level features. To enable comparison to state-of-the-art methods on the same dataset, we apply our

detection on the tweet level. This would not be an issue if each tweet in the dataset is coming from a different users. However, we noticed that 167 users (out of 3,528) in the dataset produced multiple tweets. This means that our classifiers trained on network features would always give the same classification to any tweets posted by the same user. We argue that this is acceptable based on the assumption that user stance for a given topic is not expected to change within a short period of time [11]. Moreover, we hypothesize that even if the user stance gets changed, it would be accompanied by a change in the network interactions of the user [51].

To further validate our assumption, We examined the 167 users who produced multiple tweets on the same topic. Out of those, 104 users have fixed stance in their multiple tweets and 42 have fixed polarized stance with some tweets with no stance (labeled *none*). Only 19 users have a mix of *favour* and *against* stance in the same topic, but with clear dominance for one of them (e.g. 16 vs 1 tweets). This quick analysis, shows that the majority of tweets from the same user are expected to have a fixed stance on a single topic. Thus, we believe that having a fixed set of features, based on user's network, for all tweets of the same user can be seen as an acceptable approach for stance detection on the tweet level.

3.2 Feature Extraction

We define four features sets to model the stance in social media. These sets are: on-topic content, user's network interactions, preferences and connections. Those are defined as follow:

- **On-Topic Content (TXT)**, models the text of the tweet, including features combining both word and character n-grams as presented in the best performing system in SemEVal 2016 [40]. This set of features models stance of users who explicitly express it in text.
- **Interaction Network (IN)**, models the network the user interacts with in their posts. It includes the mentioned accounts ($IN_{@}$) and website domains (IN_{DM}) the user interacts with directly either by retweeting, replying, mentioning, or linking.
- **Preference Network (PN)**, models the network the user prefers from the tweets they like. It includes the mentioned accounts ($PN_{@}$) and linked website domains (PN_{DM}) in the tweets the user likes.
- **Connection Network (CN)**, models the online social ties between the users, which includes the accounts who follow the users (followers CN_{FL}), and those the user follows (friends CN_{FR}).

Table 1 shows a detailed explanation of the feature sets. It's worth noting that *IN* features are independent of having users expressing their stance towards the target topic, since it depends on the social and web networks the user interact directly with regardless to the content in tweets. Both *PN* and *CN* features enable modeling silent or passive users who do not post or share content rather than just following or liking tweets from others. Our objective is to understand how each of these feature sets would compare to each other and to the textual features which have been studied heavily in literature.

3.3 Stance Detection Model

Since our main contribution is on stance representation to analyse the effectiveness of different social signals in detecting stance, we used our proposed set of features to train an SVM model with linear kernel for two main reasons: 1) It achieved the best performing model over 19 participating groups at SemEval 2016 [40] while outperforming more sophisticated model that used deep learning [4, 53, 55]. 2) SVM models built with linear kernel are easily to interpret, which would enable us to apply feature analysis for a better understanding to the influential features and their role in stance

Feature Set	Description
TXT	word and character n-grams of the tweet text.
IN:	user's interaction network. Extracted from user's *Home* timeline.
- $IN_{@}$	the list of accounts the user retweet for, reply to, or mention in their timeline.
- IN_{DM}	the list of web domains the user link in their tweets.
PN:	user's preference network. Extracted from user's *Likes* timeline.
- $PN_{@}$	the list of accounts mentioned in the tweets the user likes.
- PN_{DM}	the list of web domains in the tweets the user likes.
CN:	user's connection network. Accounts user connected to.
- CN_{FL}	the list of followers of the user, i.e. accounts that follow the user.
- CN_{FR}	the list of followees/friends, i.e. accounts that the user follows.

Table 1. List of feature sets examined in our experiments with their description.

	Full dataset		Existing Users	
Topic	Train	Test	Train	Test
Atheism (A)	513 (434)	220 (196)	380 (302)	170 (146)
Climate change is a real concern (CC)	395 (347)	169 (145)	317 (269)	144 (120)
Hillary Clinton (HC)	639 (556)	295 (250)	447 (364)	223 (178)
Feminist movement (FM)	664 (620)	285 (256)	354 (312)	170 (141)
Legalization of abortion (LA)	603 (496)	280 (228)	471 (365)	199 (147)
Total	2814 (2453)	1249 (1075)	1969 (1612)	906 (732)

Table 2. Number of tweets used for training and testing with respect to Semeval 2016 topic. The number of unique users authored the tweets are shown in brackets.

detection. We used Scikit-learn[2] implementation of SVM, which use cross-validation with k=5. For all the features, we use vector representation with Boolean value to indicate the presence or absence of the feature's values. We have examined other feature values, such as frequency and tf-idf, but Boolean values showed the best performance.

4 EXPERIMENTAL SETUP

4.1 Data Collection

Our experimentation has been applied to the benchmark dataset of the SemEval 2016 stance detection task [40]. The dataset contains a set of 2814 and 1249 tweets for train and test respectively covering five topics. These topics are: Atheism (A), Climate Change (CC), Feminist Movement (FM), Hillary Clinton (HC), and Legalisation of Abortion (LA). As could be noticed, these topics are not just political, but actually covers topics of social (e.g. 'FM', 'LA') and religious (e,.g. 'A') natures.

We further used the Twitter REST API to collect the network information of the users in SemEval stance dataset. Basically, we collected two timelines for each of the users posted the tweets in our dataset, namely *Home* timeline[3], which we use to contruct the user's IN; and the *Likes* timeline[4], which we use to construct the user's PN. In addition, we collected the user's list of followers and friends to construct the user's CN[5]. Unfortunately, we found that around 25% of these users have been deleted or suspended. Therefore, we end up with smaller number of tweets in the collection

[2]Scikit-learnhttp://sciki-learn.org/

[3]https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-home_timeline.html

[4]https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-favorites-list.html

[5]https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/overview

that we can apply our approach to them, exactly 1969 training and 906 testing data[6]. Table 2 shows the distribution of tweets (and users authored them) that we could retrieve in our dataset compared to the original SemEval dataset.

For each of the users in our collection, we managed to collect an average of 2,552 and 1,801 tweets from the *Home* and *Likes* timelines respectively. For each user, the set of mentions in those timelines were extracted and saved separately. In addition, we collected the set of friends and followers of each of the users in our collection. $IN_{@}$, $PN_{@}$, CN_{FR}, CN_{FL} represent the set of unique accounts appeared in the user's tweets, likes timelines, list of friends, list of followers of the user respectively. In addition, all the links appeared in the timelines were extracted and expanded (in case they were shortened). The domain of each link was then extracted and saved. IN_{DM} and PN_{DM} represent the set of unique web domains appeared in the user's tweets and likes timelines respectively.

4.2 Baselines and Evaluation

We created two baseline systems that achieve the highest reported performance on the SemEval dataset based on the best performing participating system in the SemEval task [40] that is trained on the *TXT* features. For the first baseline, an SVM with linear kernel trained on the three stance classes using a combination of both word and character n-grams was used to represent the textual content of the tweet to be classified. Word n-grams was used with $n = \{1, 2, 3\}$, and character n-grams was used with $n = \{2, 3, 4, 5\}$. These features were used to train the SVM classifier with linear kernel. We only used the subset of training data that we managed to retrieve its users network information to allow a direct comparison to our models. The outcome of this model achieved an average F-score of 68.48 on our subset of the test data, which should be comparable to the reported best model in [40] that achieved an average 68.98 F-score but on the whole dataset.

From a sociolinguistic perspective, it has been argued that there is no complete neutral stance as people use to position themselves with favor or against the object of evaluation [32]. To comply with this argument, we created our second baseline by retraining the same SVM classifier with the same set of features, but with considering only the two polarised classes {favor, against} and neglecting the 'none' class. In this way, we force classifier to have a decision on the polarised stance of the user. While this approach will misclassify the samples in the test set with ground-truth 'none' stance, it was shown in the current state-of-the-art system [46] that this approach actually outperform the three-class classifier, where they achieved F-Score of 70% when trained a binary SVM classifier with tree kernel after neglecting the 'none' class. When we applied this approach, the overall F-score of the system got an actual improvement to reach 69.8%, which is comparable to [46].

After building the linear SVM baselines (both with the three and binary classes models), we trained the same models with the different set of suggested network features. We test each feature set separately and compare their performance to the models that depends on the tweet textual content; then we apply different combination of the features to observe any potential improvement in the performance.

To evaluate the performance of our method, we used the official SemEval-2016 macro-average of the F1 score for the 'Against' and 'Favour', where the F-score on the 'None' class is discarded from calculating the average [40]. The same evaluation script provided by SemEval stance detection task was used to report the results. In addition, we show the performance over each of the five topics separately for a deeper analysis of the performance.

	Topic					Overall		
Model	A	CC	HC	FM	LA	F_{favour}	$F_{against}$	F_{avg}
TXT (Baseline)	61.38	42.86	58.91	52.01	60.96	63.09	73.87	68.48
$IN_{@}$	68.94	40.09	62.15	54.80	56.25	60.77	75.57	68.17
IN_{DM}	56.86	38.46	34.20	38.67	53.31	49.19	61.76	55.47
$IN_{@}$+IN_{DM}	70.16	39.81	61.59	57.63	64.16	64.04	76.18	**70.11**
$PN_{@}$	73.30	36.36	56.82	48.43	56.41	55.81	73.39	**64.60**
PN_{DM}	62.99	35.18	58.01	46.71	48.49	50.85	70.26	60.56
$PN_{@}$+PN_{DM}	64.55	37.13	54.27	49.00	56.44	55.73	70.14	62.94
CN_{FR}	66.71	30.11	63.87	51.51	53.10	51.15	72.76	**61.96**
CN_{FL}	40.78	20.29	54.11	46.80	56.38	39.55	65.82	52.68
CN_{FR}+CN_{FL}	49.66	28.14	66.95	48.76	49.72	44.85	67.98	56.42

Table 3. Stance detection performance using different set of features using SVM classifier trained on three classes. F-Score (%) is reported on the SemEval stance detection task for each topic and overall.

	Topic					Overall		
Model	A	CC	HC	FM	LA	F_{favour}	$F_{against}$	F_{avg}
TXT (Baseline)	61.91	42.86	59.53	52.21	62.40	63.53	76.07	69.80
$IN_{@}$	68.30	54.14	59.05	50.40	60.82	61.89	77.90	69.89
IN_{DM}	63.24	42.86	53.91	61.24	60.17	61.51	76.82	69.17
$IN_{@}$+IN_{DM}	67.65	42.86	62.64	55.87	63.93	64.04	79.07	**71.56**
$PN_{@}$	73.49	42.86	59.26	49.63	63.87	63.70	77.70	70.7
PN_{DM}	67.14	42.17	58.33	51.62	61.79	60.18	77.28	68.73
$PN_{@}$+PN_{DM}	68.03	42.86	59.00	52.57	65.50	63.91	78.60	**71.25**
CN_{FR}	63.83	42.86	64.01	60.93	59.58	64.53	78.25	**71.39**
CN_{FL}	35.97	42.86	58.51	52.70	62.68	56.08	69.73	62.91
CN_{FR}+CN_{FL}	50.00	42.86	68.21	57.38	54.13	58.07	73.41	65.74

Table 4. Stance detection performance using different set of features using *binary* SVM classifier. F-Score (%) is reported on the SemEval stance detection task for each topic and overall.

5 RESULTS

5.1 Stance Detection Results

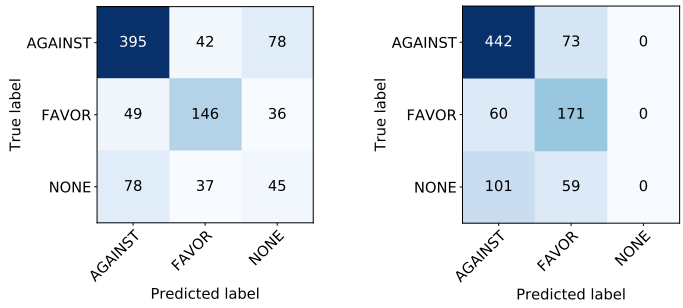
Tables 3 and 4 report the performance of the three-class classifier and binary classifier for stance detection, respectively. The general observation from the tables is that the binary classifier outperforms the classifier that is trained on three classes. While the binary classifier misclassifies tweets with no stance, it is more effective in detecting the polarised stance. This initial observation shows that forcing automatic classifiers to decide on a given stance might be a more effective approach than allowing them to have the 'none' option about stance, which makes it more confusing following Jaffe's argument that there is no complete neutral stance [32]. We analyse this further in the following subsection.

The second observation, for the binary classifier (Table 4), is that all the three set of network features - that are totally independent of the tweets contents - have better overall performance than the state of the art systems that depend on tweets textual content. In fact, the activity network (*IN*) and the preference network (*PN*) features that combine the accounts and domains features achieve better results than the baseline on all the five topics. This confirms the consistent performance of network features over text on topics of different domains. In the connection network (*CN*),

[6]List of ids of tweets and users network information would be made available

Model	F_{favour}	$F_{against}$	F_{avg}
(A) TXT+$IN_{@}$+IN_{DM}	67.21	76.49	71.85
(B) TXT+$IN_{@}$+IN_{DM}	66.67	78.31	**72.49**

Table 5. The result of baseline linear SVM model when combining both text and network features. Model (A) and (B) shows the result when trained on three and two classes, respectively.



(a) Three Classes $IN_{@}$ + IN_{DM} (b) Two classes $IN_{@}$ + IN_{DM}

Fig. 1. Confusion matrices for the best three vs two classes prediction models.

the friends network (CN_{FR}, the accounts the user follows) outperformed the baseline, while the follower network (CN_{FL}) achieved the lowest average F-score among all classifiers, even when combined with the friends network. This is potentially because of the sparsity of this network, where finding common followers among different users is less likely compared to finding common accounts they might follow, where it is expected to have people of similar stance following common accounts as a part of the homophily phenomena in social media [2, 23].

While user's interaction network showed the best overall performance among all feature sets, Table 4, it was interesting to see preference network outperformed all models in two of the five topics when using the binary classifier. These results support the hypothesis about stance detection, which is the online social network activity of a user posting a tweet contains enough signals to detect the stance of tweet regardless of its content. Furthermore, we show that the preference network of user's likes on Twitter still can achieve decent detection of stance, which enables detecting stance for silent users.

We further tested combining the best performing network features from the two networks ($IN_{@}$+IN_{DM}) with (TXT) to see if this can further improve the performance. Table 5 shows the best achieved average F-score when we combined the network with content features, where the best performance achieved when we combined the interaction network with text for both the three-class and the binary classifiers[7]. This result was found to be statistically significantly better than the state-of-the-art baseline model using two-tailed t-test with $p - value < 0.05$ (we also tested significance using Mann-Whitney U test [38], but it did not show significance).

5.2 Performance Discussion

As shown earlier, forcing the stance model to predict in-favor and against stances and ignore the 'none' stance consistently leads to better performance using all feature sets. This is an interesting

[7]we also tested other combinations of feature sets, but TXT+$IN_{@}$+IN_{DM} achieved the highest results

result, since a binary classifier will always misclassify the 'none' class leading to larger number of false positives to the other two main polarised classes, which should reduce the performance. To better understand this, we plot the confusion matrices for the best performing model for both three/two class classifiers in Figure 1. As it is shown, the binary classifier led to larger number of false positives for both the polarised classes; however at the same time, it led to larger number of true positives for both classes. This led to improvement in recall with some reduction in precision, with an overall improvement in the average F-score.

Another observation from Tables 3 and 4, is the low performance of classifying stance on the climate change (CC) topic, where it has the lowest F-score among all topics. We conducted a further analysis and we noticed a large difference in the class distribution between the 'in-favor' and 'against' classes, where 176 samples in the training set are labeled as 'in-favor', while only 8 samples are labeled as 'against'. This led the classification models to predict the majority class in most of the cases, which led to random-like performance for this topic.

Our obtained results for stance classification are the highest to be reported to date on the SemEval dataset, which confirms the large impact of utilising user's network activity as features in boosting the performance of stance detection, especially when combined with textual features. Our results highlight that user's stance towards given topics could be inferred with various types of features from their activity online. In the following section, we apply an extensive analysis to these features to understand its role and influence in revealing the user's stance.

6 FEATURE ANALYSIS

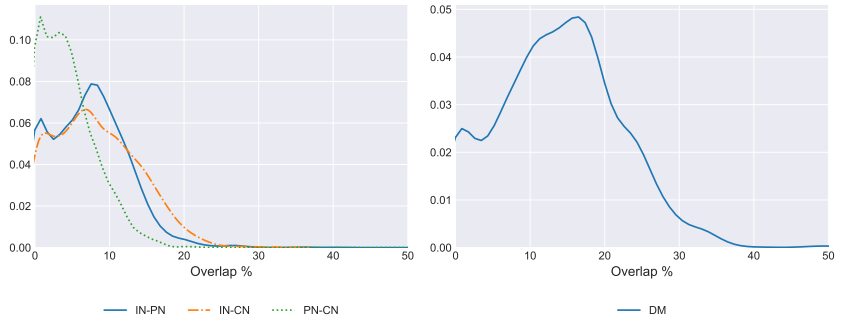
In this section, we analyse each of the network features that showed to be effective in detecting stance. We apply our analysis to the binary classifier, which achieved the highest results. Our analysis includes studying differences between our three networks, analysing most influential features per network and per topic, and giving examples of how these features might be effective.

6.1 Similarity between Networks

From the results obtained in Table 4, it is noticed that the scores achieved by the three groups of networks (IN), (PN) and (CN) are relatively similar. The average F-scores obtained by ($IN_{@}$ + IN_{DM}), ($PN_{@}$ + PN_{DM}) and (CN_{FR}) are around 71% and their results were found to be statistically indistinguishable from each other using both t-test and Mann-Whitney U test. This motivates to further examine the overlap among these networks, since it is highly possible that users interact with and like content of the same set of accounts they follow. Hence, we measure the overlap between the features of (IN), (PN) and (CN) to gauge the similarity among them.

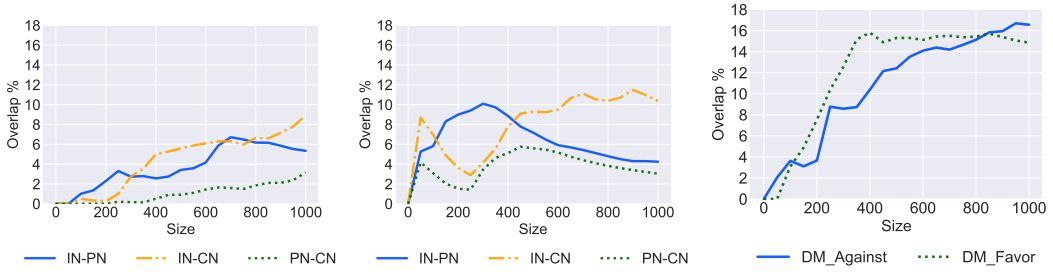
For each user, we compute the similarity between their $IN_{@}$, $PN_{@}$, and CN_{FR} features using Jaccard similarity, then we plot the distribution of the similarity score across all users. We repeat this process for the domains features by computing the similarity between IN_{DM}, PN_{DM}. Figure 2 shows the similarity distribution between the network's sets, where zero indicates no overlap and 100% means identical sets. We observe that there is a noticeable difference in each network for the same feature component. The overall similarity between accounts in each of the three networks ranges between zero and 20%, and it ranges between 0 and 35% for domains. This result means that users tend to interact and like contents from users out side their connection network, and like tweets with links generally different from the domains they link in their tweets. This is actually an interesting finding, which actually raises further research question about the reason of having the performance of the three networks in stance detection similar when they are mostly different.

There is a hypothesis behind the similar performance, that actually the small percentage of similar accounts (domains) between the three networks are those which create the most influential features for the classification, and thus the three classifiers achieved comparable performance.



(a) $IN_{@}$-$PN_{@}$-CN_{FR}. (b) IN_{DM} and PN_{DM}.

Fig. 2. Similarity between CN, IN and DM in users dataset.



(a) $IN_{@}$, $PN_{@}$ and CN_{FR} - Favor. (b) $IN_{@}$, $PN_{@}$ and CN_{FR} - Against. (c) IN_{DM} and PN_{DM}.

Fig. 3. Similarity between CN, IN and DM for (In-favor and Against) stances with respect to the top features.

Therefore, we further analyse the similarity between the most influential features of the three networks sets, where influential features are identified as those having the highest weights for each of the classes for each topic. We use Jaccard similarity to compute the similarity between the top N influential features of IN, PN and CN and plot the similarity for N={1 $\rightarrow$ 1000}. Figure 3 presents the similarity for each network features influencing favor and against stance. Again, it is observed that similarity between the most influential features is not high for any of the networks for both 'favor' and 'against' classes, where the similarity does not exceed 10% for the accounts, and 17% for domains.

These findings confirm the differences between the three networks, and show that each network represents a different set of accounts and domains for the same user. Even with the most influential features for models trained on each classifier the set of features is different than the other. This means that a more in-depth analysis to these features is required to understand the high performance of the classifiers trained on the three fairly independent networks.

6.2 Which Network Features Reveal the Stance?

To get meaningful insights about the contribution of the features to infer the stance, we identify the most influential feature of the best model from (CN), (IN), and (PN) network with regards each topic.

T	NW	Favor	Against
A	IN	@atheism_tweets, @atheistrepublic, @god_stupid	@ChristianInst, @godlesstheory,@godbiblechurch
	PN	@thetweetofgod,@foxnews,@nytimes	@prayerbullets, @reuters, @cnn
	CN	@Stephenfry, @RichardDawkins, @MarilynManson	@baptism_saves, @srisri, @artofliving
CC	IN	@telegraph, @independent, @climatereality	@skynewsbreak, @nytopinion, @reuters
	PN	@nytstyles, @news4anthros, @fox2now	@cnn, @foxnews, @nythealth
	CN	@barackobama, @potus, @mashable	@foxnews, @sentdcruz, @cnn
HC	IN	@washtimes,@hillaryclinton,@realdonaldtrump	@trumpstudents, @foxnewssunday,@brianschatz
	PN	@cbsnews,@nbcnews, @hillaryfornh	@govchristie,@drbiden, @sentedcruz
	CN	@hillaryclinton, @billclinton,@shehasmyvote	@foxnews, @realdonaldtrump, @madam_presiden
FM	IN	@mtv,@goodreads,@feministculture	@feministfailure, @goodmenproject, @womenwriters
	PN	@feministajones,@ppfa,@foxnews	@nytopinion, @mtvnews,@weneedfeminlsm
	CN	@vday,@Schofe,@twitterfashion	@Truth_seeeeker,@femalefedupwith,@thepowrhouse
LA	IN	@humanesociety,skynews ,@ppactionca	@ppfa, @nbcnews, @bible_time
	PN	@savewomenslives,@dallasnews,@citynews	@onejesusloves, @younglife, @yahoonews
	CN	@thedemocrats,@barackobama,@hillaryclinton	@prolifeyouth,@march_for_life, @lifeteen

Table 6. Top features extracted from the best model in each case and trained on two classes, CN_{FR}, $IN_{@}$,$PN_{@}$.

T	DM	Favor	Against
A	IN	sciencealert, thinkprogress, washingtonpost	nationalpost, washingtonpost, newsweek
	PN	reuters, newhumanist, telegraph	faithreel, bible,prayerbullets
CC	IN	thetimes, nytimesarts, nbcnews	bbc, naturalnews, washingtontimes
	PN	abc, newswire, nypost	cbc, telegraph, washingtontimes
HC	IN	nytimes, thedailybeast, cnbc	opposingviews, washingtontimes, foxnews
	PN	nytimes, theguardian, nbc	cnn, foxnews, newsfoxes
FM	IN	cnn, buzzfeed, nytimes	dailymail, bbc, theguardian
	PN	apnews, washingtontimes, feministing	independent, dailymail,activistpost
LA	IN	newstatesman, nytimes, cnn	nypost, dailymail, cbsnews
	PN	bpas, ahealthblog, thenation	lifenews, gotquestions, cnsnews

Table 7. Top features extracted from the best model in each case and trained on two classes, IN_{DM},PN_{DM}.

We hope this would give some explanation to the good performance of these models, especially after we found that these networks do not highly overlap.

Table 6 and 7 show the top features that have a noticeable influence on the stance classification for each topic with respect to the weights of features in the linear SVM model for the best features from each network group: ($IN_{@}$+IN_{DM}), ($PN_{@}$+PN_{DM}) and (CN_{FR}).

In the (CN_{FR}) network, the social influence manifest through the users' friends (following network). Users tend to follow the accounts that support their stance. For instance, users with against stance toward legalisation of abortion (LA) tend to follow accounts that oppose the abortions such: '@prolifeyouth', '@march_for_life'. The same for the users with favor stance to Hillary Clinton where the top followers are '@Hillaryclinton', '@billclinton', '@shemyvote'. Users who have a favor stance towards Atheism tend to follow social actors with the same believes such:'@Stephenfry', '@RichardDawkins', '@MarilynManson'. Similarly, users with favor stance toward feminist movement follow the accounts that support feminism. One of the top features that identifies the in-favor stance toward feminism is '@vday', which is an activist movement account that supports the feminist movement as this account description indicates: "to End Violence Against Women & Girls.". For the climate change and legislation of abortion, the politicians and news outlets are the most influential accounts in predicting the stance. We can not specify whether these users follow such account because they support their opinion towards each topic.

Unlike CN, influential accounts for IN and PN include news accounts. For instance, the news accounts '@washtimes' and '@cbsnews' are one of the distinguishing features to detect the favor stance to Hillary Clinton in $IN_{@}$ and $PN_{@}$. In addition, '@telegraph' in $IN_{@}$ has a positive correlation with favor stance to climate change. Users with favor stance to the legalization of abortion interact with '@skynews' account. In contrast, news accounts have a minimal effect in detecting stance toward feminist movement and atheism, where the top mentions features that capture a favor stance are accounts that support the topic: '@atheism_tweets' and '@feministcultur'.

Also, another difference between IN and PN, is that IN usually contains accounts of opposing view since in this case the interaction can be through replying or quoted retweets with opposing comments. This case can be seen in Hilary Clinton topic, where '@realDonalTrump' is one of the top features for the 'favor' stance in IN. It can be imagined that the interaction here is not for support as shown in table 8 (Example 3). In addition, interacting with accounts that have a related meaning to the topic seems to have a visible correlation with detecting the against stance of users. For instance, the interaction with '@godlesstheory' and '@godbiblechurch' has an influence in detecting the against viewpoint to atheism. Similarly, '@bible_time' captures the against stance toward abortion. Furthermore, famous accounts with clear support to a related social issue have a clear influence in detecting the stance. For instance, users with against stance to feminism interact with '@feministfailure'. In addition, users who oppose the legalisation of abortion interact with '@ppfa', Planned Parenthood account.

For the web domain features, it can be noticed that the top domains features IN_{DM} and PN_{DM} are mostly news websites. News websites and media outlets such as 'washingtonpost' and 'sciencealert' are one of the distinguishing features to detect favor stance toward Atheism. In contrast to mentions, the news websites have a noticeable effect in detecting users view points toward feminist movements. We can see that users with against stance to feminist movement tend to share contents from 'dailymail', 'bbc' and 'theguardian' websites. Users with support stance to feminist movement tend to share contents from 'cnn'. Users with against stance to Hillary Clinton share contents from news websites such as 'opposingviews', 'washintontimes' and 'foxnews'. The website 'nytimes' has a positive effect in identifying the favor stance to Hillary Clinton. We can notice some overlap between IN_{DM} and PN_{DM}, where it seems users like and interact with the same news and media outlets in the PN and IN networks. For instance, users with against stance to Hilary Clinton tweet interact and like news contents from 'foxnews'. The same for users with against stance to the feminist movement, the users like and interact with 'daily-mail'. In general, there is a tendency for the users to like and share content from the same media as described in the next section.

6.3 The Context of the Features

We carried a further qualitative analysis to identify the context in which the IN and PN features correlate with the topic of the target. Table 8 shows a sample of tweets from the users' timelines (IN) and Favorite timeline (PN) with respect to topic-stance pair and highlights the interactive nature of the user with the top features. As explained in the previous section, what sets apart users with support/against stance to climate change are those pertaining to news portals. For instance, the most dominant mentioned accounts that influence the supporting and opposing position toward climate change is '@telegraph' and '@SkyNewsBreak'. Users interaction with these news accounts in the sense of re-tweeting and liking the news that has no relation to climate change (Example 1 and 6).

Tweets from '@NBCNews' with no relevance to Hilary Clinton or the presidential candidates tend to be liked by users with a stance supporting Hillary, (Example 2). The same with users who support feminist movement, they interact with account '@goodreads' with no topical relation to stance topic, (Example 4). The user mentioned @goodreads to promote to the novel "Beginnings"

#	T	Feat	Example tweets (favor)
1	CC	$IN_{@}$	RT @Telegraph: Prince Charles reveals his gardening inspiration: a hidden Buckingham Palace veg plot https://t.co/tBZB5DSKt5
2	HC	$PN_{@}$	@NBCNews Kill the bear for BEING A BEAR! What's wrong with this?
3	HC	$IN_{@}$	You are an idiot on so many levels, @realDonaldTrump https://t.co/keptgYgTed
4	FM	$IN_{@}$	I'm your nightmare come true," said Angela. #YAlit #vampire #paranormal #Action #humor https://t.co/MCvYEvdz8Z @goodreads
5	A	$IN_{@}$	@god_stupid @userid just the ignorant, racist, sexist, child abusing fanboys that roll play #christianity.#Atheist and proud
#	T	Feat	Example tweets (against)
6	CC	$IN_{@}$	RT @SkyNewsBreak: Former Labour Prime Minister Tony Blair has told Sky News Theresa May will win the General Election #GE2017
7	A	$IN_{@}$	RT @ChristianInst: Romans 8:28 And we know that for those who love God all things work together for good, for those who are called accordi
8	A	$PN_{@}$	@prayerbullets: Turn every curse sent my way into a blessing -Neh. 13:2 #Prayer

Table 8. Sample of tweets and the context of IN and PN in relation with stance and topic.

which is a teen romance, sci-fi and fantasy story. Furthermore, example 6 demonstrates how the interaction with @SkyNews helps in predicting the against stance towards Climate Change (CC) even with news that does not concern with climate change. In contrast, users opposing atheism tends to mention religious accounts to support their stance against atheism. For instance, users with against stance toward atheism interacted with '@ChristianInst' by retweeting verses from scripture (Example 7). Furthermore, users who have an against stance toward atheism tend to like religious's content from accounts such as '@prayerbullets' (Example 8). Users supporting atheism interact with accounts that are sarcastic toward religions such as '@god_stupid' account, in a sense of hashtag as a way of expressing the against viewpoint towards the religious people. The account '@god_stupid' is a sarcastic account, yet the interaction with it tends to take a kind of attacking the religious means as shown in (Example 5). Similarly, Users supporting Hilary Clinton defending their viewpoint by attacking '@realdonaldtrump' (Example 3).

7 DISCUSSION

In this work, we studied the possible signals that can reveal the user's stance from their publicly available online data. Unlike most of the literature in this area, which mostly focuses on achieving a high accuracy without in-depth analysis, our main focus is to understand how stance could be revealed throughout different sets of signals. This led us to explore multiple sets of features including some that have not been examined before (such as the preference network), and test it on a stance benchmark dataset of multiple topics of different genres.

7.1 What factors reveal stance and how?

Our study investigates three main research questions that have not been sufficiently explored in earlier studies on stance detection.

Our first research question is concerned with exploring the different signals from user's public social media profiles that can reveal their stance. We have defined three sets of network features, including interaction (IN), preference (PN), and connection (CN) networks, and compared their performance to textual features that represent the state-of-the-art models on the SemEval dataset. Our findings showed that user's stance can be detected with many signals, including textual content and different sets of network features. We found that using network features leads to a more accurate stance detection than using content-based features solely, and the performance becomes statistically significantly better when both sets of features are combined together. We also

noticed that when building a stance classifier, a binary classifier is more superior than a classifier that allows neutral stance, which could be linked to the argument that there is no "neutral" stance and everyone should have some leanings [32].

Our second research question focused on how the performance of the stance detection using these features would differ across different topics. Our analysis of the five topics in our dataset showed that network features consistently achieve better performance on average compared to textual features. We only found that the performance for one topic (CC) has always the lowest F score. Our investigation to the distribution of the stances on this topic suggests that the problem stems from the large imbalance in the training samples, which leads the prediction model to predict only the majority stance class, which is independent of the set of features used.

As for our third research question, which concerns with investigating what makes the introduced features effective for stance detection; we initially analyzed the overlap between the accounts and web domains for each of the users in our dataset in the three networks: IN, PN and CN to ensure that their similar performance is not the reason for their high similarity in their nodes. It was surprising to find them mostly dissimilar with low overlap among them with <20% similarity between them. This was interesting to see that each of them captures one side of the user's activity, and each can reveal their stance. We further investigated the top features in each network model. We noticed that the top features can sometimes be topically unrelated to the target and yet have a high impact on deciding the stance of the topic. For instance, the interactions with accounts as @goodreads and @SkyNews help in detecting the stance towards feminist movement (FM) and climate change (CC) respectively, as shown in section 6.3. Since these features have no direct relation to the topic of the stance, this indicates that the user's stance can be detected with many signals regardless of the topic. We showed that using content-less features help in detecting the stance for the users with an implicit point of view toward a topic where the users may not directly express their point of view by using keywords related to the target. As the top features extracted from the two networks ($PN_{@}$) and ($IN_{@}$) have no direct relation to the stance's topic. For instance, the '@Telegraph' was one of the top features that predicts the in-Favor stance towards Climate Change topic.

Furthermore, one of the key findings from our study is the high performance of PN and CN for stance detection, which outperforms the state-of-the-art baseline TXT model. This shows that detecting stance for silent/passive users (who never tweet or share any content) is doable, given the condition that they have enough common signals in their preferences and connection networks. This raises a real concern about the privacy of social media users in general, and motivates future research in the direction of protecting those users from having their leanings and beliefs revealed unconsciously [51].

Our experiments and analysis was applied to a set of five topics of political, social, and religious natures. Our findings show that regardless to the topic, there are usually signals in the users' online activity and connections that can reveal the stance of those users towards this topic.

7.2 Ethics and Privacy Considerations

Using Twitter as a central platform for this study is supported by a robust literature[14, 18, 41, 44]. Previous studies have shown that the stance detection in social media provides useful information to understand better the way in which people communicate and express their viewpoint towards a topic. Stance detection has been used as the first step toward solving fake-news, polarization, and rumours [24, 26, 58]. Hence, the famous Semeval stance detection benchmark dataset has been constructed using public tweets [40]. Furthermore, Twitter does not force demographic information of the user upon registration. Consequently, the accounts are not linked to the physical identity of the users. In the process of collecting the additional tweets to extend the Semeval stance dataset we are using authorized developers accounts approved by Twitter application developer portal[1].

This study does not store non-public Twitter content, such as direct messages or other private or confidential information. The collected tweets are the publicly available data on Twitter as further indicated in the Twitter Developer Policy [8].

8 CONCLUSION AND FUTURE WORK

In this paper, we present a thorough analysis of the four main scenarios to predict the stance on social media. We investigate with a stance detection approach that can be text-independent, where the stances are predicted from users online activity. We introduce three sets of networks to represent users, which are the interaction, preference and connections network. The interaction network includes the accounts the user interacted with and the website domains the user shared; and the preference network represents the accounts and website domains in the tweets the user liked. Finally, the connection network is the set of friends and followers of the user. We conducted the experiments on SemEval 2016 stance benchmark dataset, and showed the superiority of network features over textual features when compared with the baseline model. All three network-based models outperformed the state-of-the-art methods that depend on textual features only. We also presented an analysis of the top features to identify the correlation between stance and topic with respect to the features groups. We explored with the key important features that have a positive effect on detecting stance for each target. The results denote accurate learning of the stances at the user-level representation that improves the content-related features model.

For future work, more analysis of the network features could be applied, since retweeted accounts might denote different preferences than replied or mentioned ones. In addition, it is essential to create new sets of data covering more topics to validate the generalisability of our findings. Finally, since our work raises some concerns about the vulnerability of users by having their stances easily detectable even when not directly discussing the topic, it becomes highly essential to develop methods to counter those automatic methods for detecting user's leanings as a step to protect user's privacy.

REFERENCES

[1] Wasim Ahmed, Peter A Bath, and Gianluca Demartini. 2017. Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. In *The Ethics of Online Research*. Emerald Publishing Limited, 79–107.

[2] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *ICWSM* 270 (2012), 2012.

[3] Abeer Aldayel and Walid Magdy. 2019. Assessing Sentiment of the Expressed Stance on Social Media. In *Proceedings of the 11th International Conference on Social Informatics (SocInfo 2019)*.

[4] Isabelle Augenstein, Tim Rocktaschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *EMNLP 2016* (2016). https://doi.org/10.18653/v1/d16-1084

[5] Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 2016. Usfd at semeval-2016 task 6: Any-target stance detection on twitter with autoencoders. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 389–393. https://doi.org/10.18653/v1/s16-1063

[6] Reem Bassiouney. 2015. Stance-Taking. *The International Encyclopedia of Language and Social Interaction* (2015). https://doi.org/10.1002/9781118611463.wbielsi139

[7] Adrian Benton and Mark Dredze. 2018. Using Author Embeddings to Improve Tweet Stance Classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*. 184–194.

[8] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 21–30. https://doi.org/10.1145/2470654.2470658

[9] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics* 225, 10 (2016), 2047–2059. https://doi.org/10.1140/epjst/e2015-50319-0

[8] https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

[10] Douglas Biber and Edward Finegan. 1988. Adverbial stance types in English. *Discourse processes* 11, 1 (1988), 1–34. https://doi.org/10.1080/01638538809544689

[11] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind Egyptian political polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 700–711. https://doi.org/10.1080/01638538809544689

[12] Carter T Butts. 2009. Revisiting the foundations of network analysis. *science* 325, 5939 (2009), 414–416. https://doi.org/10.1126/science.1171022

[13] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication* 64, 2 (2014), 317–332. https://doi.org/10.1111/jcom.12084

[14] Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved Stance Prediction in a User Similarity Feature Space. In *ASONAM'17*. https://doi.org/10.1145/3110025.3110112

[15] Kareem Darwish, Peter Stefanov, Michaël J Aupetit, and Preslav Nakov. 2020. Unsupervised User Stance Detection on Twitter. *ICWSM 2020* (2020).

[16] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention. In *European Conference on Information Retrieval*. Springer, 529–536. https://doi.org/10.1007/978-3-319-76941-7_40

[17] Marcelo Dias and Karin Becker. 2016. INF-UFRGS-OPINION-MINING at SemEval-2016 Task 6: Automatic Generation of a Training Corpus for Unsupervised Identification of Stance in Tweets. *Proceedings of SemEval* (2016), 378–383. https://doi.org/10.18653/v1/s16-1061

[18] Sarah Djemili, Julien Longhi, Claudia Marinica, Dimitris Kotzinos, and Georges-Elia Sarfati. 2014. What does Twitter have to say about ideology?. In *NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication/Social Media-Pre-conference workshop at Konvens 2014*, Vol. 1. Universitätsverlag Hildesheim, http–www.

[19] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence. https://doi.org/10.24963/ijcai.2017/557

[20] John W Du Bois. 2007. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction* 164 (2007), 139–182. https://doi.org/10.1075/pbns.164

[21] Heba Elfardy and Mona Diab. 2016. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 434–439. https://doi.org/10.18653/v1/s16-1070

[22] Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. 2018. Stance Classification through Proximity-based Community Detection. (2018). https://doi.org/10.1145/3209542.3209549

[23] Kiran Garimella et al. 2018. Polarization on Social Media. (2018).

[24] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 81–90. https://doi.org/10.1145/3018661.3018703

[25] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 3. https://doi.org/10.1145/3140565

[26] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2018. Stance Detection in Fake News A Combined Feature Representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 66–71.

[27] Wei Gong, Ee-Peng Lim, and Feida Zhu. 2015. Characterizing Silent Users in Social Media Communities.. In *ICWSM*. 140–149.

[28] Yupeng Gu, Ting Chen, Yizhou Sun, and Bingyu Wang. 2017. Ideology Detection for Twitter Users via Link Analysis. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 262–268. https://doi.org/10.1007/978-3-319-60240-0_32

[29] Lei Guo, Jacob A Rohde, and H Denis Wu. 2018. Who is responsible for Twitter's echo chamber problem? Evidence from 2016 US election networks. *Information, Communication & Society* (2018), 1–18. https://doi.org/10.1080/1369118x.2018.1499793

[30] Kazi Saidul Hasan and Vincent Ng. 2013. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 816–821.

[31] Itai Himelboim, Stephen McCreery, and Marc Smith. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication* 18, 2 (2013), 154–174. https://doi.org/10.1111/jcc4.12001

[32] Alexandra Jaffe. 2009. *Stance: sociolinguistic perspectives*. OUP USA. https://doi.org/10.1111/j.1548-1395.2012.01144.x

[33] Mirko Lai, Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Friends and enemies of clinton and trump: using context for detecting stance in political tweets. In *Mexican International Conference on Artificial Intelligence*.

Springer, 155–168. https://doi.org/10.1007/978-3-319-62434-1_13

[34] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance Evolution and Twitter Interactions in an Italian Political Debate. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 15–27. https://doi.org/10.1007/978-3-319-91947-8_2

[35] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the tenth conference on computational natural language learning*. Association for Computational Linguistics, 109–116. https://doi.org/10.3115/1596276.1596297

[36] Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016. # isisisnotislam or# deportallmuslims?: Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*. ACM, 95–106.

[37] David McKendrick and Stephen A Webb. 2014. Taking a political stance in social work. *Critical and Radical Social Work* 2, 3 (2014), 357–369. https://doi.org/10.1332/204986014x14096553584619

[38] Patrick E McKnight and Julius Najab. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology* (2010), 1–1.

[39] Yelena Mejova, Ingmar Weber, and Michael W Macy. 2015. *Twitter: a digital socioscope*. Cambridge University Press.

[40] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets.. In *SemEval@ NAACL-HLT*. 31–41.

[41] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)* 17, 3 (2017), 26.

[42] Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic Stance Detection Using End-to-End Memory Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Vol. 1. 767–776. https://doi.org/10.18653/v1/n18-1070

[43] Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 869–875.

[44] Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 153–160. https://doi.org/10.1007/978-3-319-05579-4_19

[45] Dong Rui, Sun Yizhou, Wang Lu, Gu Yupeng, and Zhong Yuan. 2017. Weakly-Guided User Stance Prediction via Joint Modeling of Content and Social Interaction. In *CIKM17*. Singapore. https://doi.org/10.1145/3132847.3133020

[46] Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2018. Stance Detection on Microblog Focusing on Syntactic Tree Representation. In *International Conference on Data Mining and Big Data*. Springer, 478–490.

[47] Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 226–234. https://doi.org/10.3115/1687878.1687912

[48] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. 2017. Users are known by the company they keep: Topic models for viewpoint discovery in social networks. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 87–96. https://doi.org/10.1145/3132847.3132897

[49] Amine Trabelsi and Osmar R Zaiane. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Twelfth International AAAI Conference on Web and Social Media*.

[50] Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate.. In *LREC*. 812–817.

[51] Marcin Waniek, Tomasz P Michalak, Michael J Wooldridge, and Talal Rahwan. 2018. Hiding individuals and communities in a social network. *Nature Human Behaviour* 2, 2 (2018), 139. https://doi.org/10.1038/s41562-017-0290-3

[52] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 384–388. https://doi.org/10.18653/v1/s16-1062

[53] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 Task 6: A Specific Convolutional Neural Network System for Effective Stance Detection.. In *SemEval@ NAACL-HLT*. 384–388.

[54] Wei Xie, Cheng Li, Feida Zhu, Ee-Peng Lim, and Xueqing Gong. 2012. When a friend in twitter is a friend in life. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 344–347.

[55] Guido Zarrella and Amy Marsh. 2016. MITRE at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784* (2016).

[56] Michael Wojatzki Torsten Zesch. 2018. Comparing Target Sets for Stance Detection: A Case Study on YouTube Comments on Death Penalty. (2018), 19–21.

[57] Yiwei Zhou, Alexandra I. Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*. Springer, 18–32.

https://doi.org/10.1007/978-3-319-68783-4_2

[58] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management* 54, 2 (2018), 273–290. https://doi.org/10.1016/j.ipm.2017.11.009