#Brexit: Leave or Remain? The Role of User's Community and Diachronic Evolution on Stance Detection

Mirko Lai [a,*] and Viviana Patti [a] and Giancarlo Ruffo [a] and Paolo Rosso [b]
[a] *Dipartimento di Informatica, Università degli Studi di Torino, C.so Svizzera 185, 10149, Turin, Italy*
[b] *PRHLT Research Center, Universitat Politècnica de València, Camino de Vera s/n. 46022 Valencia, Spain*

Abstract. Interest has grown around the classification of stance that users assume within online debates in recent years. Stance has been usually addressed by considering users posts in isolation, while social studies highlight that social communities may contribute to influence users' opinion. Furthermore, stance should be studied in a diachronic perspective, since it could help to shed light on users' opinion shift dynamics that can be recorded during the debate. We analyzed the political discussion in UK about the BREXIT referendum on Twitter, proposing a novel approach and annotation schema for stance detection, with the main aim of investigating the role of features related to social network community and diachronic stance evolution. Classification experiments show that such features provide very useful clues for detecting stance.

Keywords: Stance Detection, Twitter, Brexit, NLP, Community Detection

1. Introduction

Online debates are a large source of opinion-sharing dialogue on current socio-political issues, and several works rely on finer-grained sentiment analysis techniques to analyze politics. In the last decade, social media gained a very significant role in the public debate, specially in political activism. Indeed, several politicians - or their staff - actually use Twitter or other social media directly to spread their opinions or to reinforce their political campaign. On the other hand, Twitter users take part in the discussion, in particular during elections and events of public interest. Social media provide a powerful experimental tool to investigate how individuals are exposed to diverse viewpoints. Although there is an on-going scientific discussion on the existence and the real extent of the so called "echo chambers" and "filter bubbles", it is however possible to observe that online and offline forms of political participation can have both positive and negative effects [10,42]. The ever growing number of messages posted on social media platforms has progressively motivated the increasing need of intelligent systems able to assess the contents that users generate. In particular, a growing interest has been expressed for the classification of users' stance, i.e. the detection of positions pro or con a particular target entity that users assume within polarized debates, applied to data from microblogging platforms such as Twitter [30].

Social studies highlight that users' social communities can play a crucial role in determining stance within polarized debates since social relations may contribute to influence users' opinion. It has been observed that individuals that share the same stance toward a specific target are likely to belong to the same community, thus it would be an interesting cue to exploit in stance detection. Moreover, stance should be studied in a diachronic perspective, since people may change their opinions or their communication style after some particular events that happened when the debate is still active. However, stance has been mostly addressed by considering users posts in isolation, focusing on the

*Corresponding author. E-mail: lai@di.unito.it

content analysis of the single posts, without considering the surrounding context. In this work, we aim at investigating this issue, by proposing a new approach to stance detection (SD henceforth) where the role of three orthogonal contextual facets, which may influence user's stance, is explored: (i) social network community, (ii) diachronic evolution and (iii) common knowledge on the entities involved in the debate.

Our approach relies on the use of content analysis techniques within a computational social science scenario [26] for extracting from the social media data information on social relations between users sharing contents and opinions towards a given target entity.

In order to evaluate our approach, we analyzed the political discussion in United Kingdom (UK) about the European Union membership referendum, held on June 23rd 2016, commonly known as BREXIT, an abbreviation for "British exit", collecting about 5M of English tweets containing the hashtag #brexit. We propose a novel stance detection annotation schema to apply to our social media data that takes into account diachronic evolution, allowing to study stance from a diachronic perspective. As a side product of our work, we released the Twitter diachronic Brexit corpus (TW-BREXIT, herceforth), where diachronic triplets of tweets posted by 600 users active in the debate have been annotated for stance. The core idea is to consider the evolution of the user's stance over the time monitoring her posts in different time windows corresponding to three 24-hour time steps forthcoming elections.

Overall, the analysis of the corpus and the classification experiments show that:

- Analysis of the community network structure helps to improve the performance of the SD task. Therefore, the intuition that user's stance is strongly related to the social media network community, in accordance with the homophily principle [28,5,7], seems to be confirmed in our setting.
- User's stance changes after relevant events, in accordance with theoretical studies from political sciences [14]. This confirms the importance of taking a diachronic perspective in the SD task.
- Context-based features include community, diachronic evolution, and common-knowledge. All three at once obtain significant results on ablation test.

The corpus, the annotation guidelines and the source code of the experiments are available for the community to foster reproducibility of the experiments[1].

The paper is structured as follows. Related work is discussed in Section 2. Section 3 describes the features proposed in our model. Section 4 presents the development of the TW-BREXIT corpus. Classification experiments are reported and discussed in Section 5. Section 6 concludes the paper.

2. Related Work

Political sentiment and stance detection

Recent trends in monitoring political sentiment consist in considering texts from social media and in exploiting computational techniques for extracting information about the political landscape in the offline world. In particular, techniques such as sentiment analysis (SA) and opinion mining [36], developed within the context of computational linguistics for extracting several kinds of information about humans' behavior, can be specifically declined for monitoring political contents and are gradually considered as especially useful for this purpose, with possible different focuses: detecting users stance, detecting the polarity of messages expressing opinions about candidates in political elections [4], detecting deception in text [17]. Among such new areas derived from SA, stance detection is one of the most interesting [1,27]. In 2016 a shared task on Stance Detection on Twitter has been proposed at SemEval-2016 (SemEval2016-Task6). Mohammad et al. [30] describe the task as follows: "Given a tweet text and a target entity (person, organization, movement, policy, etc.), automatic natural language systems must determine whether the tweeter is in favor of the target, against the given target, or whether neither inference is likely". The SemEval2016-Task6 dataset included six commonly known targets in the United States, such as: "Atheism", "Climate Change is a Real Concern", "Feminism Movement", "Hillary Clinton", "Legalization of Abortion", and "Donald Trump".

Most participating teams exploited standard text classification features such as n-grams and word embedding vectors. Some SA features relying on well-known lexical resources, such as EmoLex [33], MPQA [46], Hu&Liu [19], and NRC Hashtag [31], were also exploited. Baselines were established based on n-

[1] https://github.com/mirkolai/leave-or-remain

grams, char-grams and Majority Class (MC), but no team outperformed such baselines, confirming the difficulty of the task. An enhanced version of the annotated corpus for SD used in SemEval2016-Task6 was subsequently released, which includes two additional labels: the opinion target, and the sentiment polarity [39]. Mohammad et al. [32] and Lai et al. [21] exploited the new labels, showing that information on sentiment polarity and additional knowledge about the target of the opinion help in detecting stance. In particular, Lai et al. [21] proposed an approach for detecting stance in Twitter that relies on domain knowledge by considering the context surrounding a target of interest. The proposed approach was evaluated on two targets of the SemEval2016-Task6: Hillary Clinton and Donald Trump. Three groups of features were considered: *structural* (hashtags, mentions, punctuation marks, etc.), *sentiment* (a set of four lexica to cover different facets of affect ranging from prior polarity of words such as AFINN [34] and Hu&Liu Lexicon, to fine-grained emotional information such as LIWC [37] and the Dictionary of Affect in Language was exploited), and *context-based* features. Attempting to capture the contextual information surrounding a given target, the authors use the concepts "friends" and "enemies" for defining the relationships between the target and the politicians and the parties mentioned in the text. The proposed approach outperform the state-of-the-art results, showing that information about enemies and friends of politicians help in detecting stance towards them. In Lai et al. [23] the role of social relations was been analyzed together with the users' stance towards the BREXIT referendum. The study shows two main results that may be of particular interest for addressing SD: users sharing the same stance towards a particular issue tend to belong to the same social network community, and users' stance diachronically evolves. A similar experiment has been performed by Lai et al. [22,24] analyzing the political debate on Twitter about the Italian Constitutional referendum held in 2016. Notice that the importance of using contextual features has been highlighted by many scholars in relation also with different NLP tasks aimed at detecting subjective information in user-generated contents. For instance, Wallace et al. [43] used communities belonging to different Reddit's subreddits in order to improve irony detection. Furthermore, Bamman and Smith [2] used information such as historical terms, topics, and sentiment from both author and audience in order to improve sarcasm detection. The current proposal is settled on this line of research: we aim at exploring the importance of information about interpersonal relations, time evolution, and common knowledge about the target in SD, which can be all seen as different facets of context.

Recently, stance detection was addressed also in the context of political debates featured by languages different than English. The very controversial issue *Independence of Catalonia* was chosen as target in order to perform Stance Detection in Tweets (StanceCat[2]) in the framework of the evaluation campaign IberEval 2017 [41]. The aim of the shared task was to detect the author's stance towards the Independence of Catalonia in tweets written in Spanish or Catalan, collected during the regional elections in Catalonia, which took place on September 27, 2015. A dataset in both languages was released, containing a set of tweets manually annotated as in favor, against or neutral towards the target of interest. Well-known classifiers such as SVM and novel techniques such as deep learning approaches were used by the ten different teams participating in the shared task. As features n-grams and word embeddings were the most used. The best performed approach in both languages on the shared task, called iTACOS [20], consisted in a supervised approach that considers three groups of features: *Stylistic* (bag of: n-grams, char-grams, part-of-speech labels, and lemmas), *Structural* (hashtags, mentions, uppercase characters, punctuation marks, and the length of the tweet), and *Context* (the language of each tweet and information coming from the URL in each tweet). The obtained results validate the importance of considering contextual information in stance detection tasks.

The political debate around the BREXIT referendum has been at the center of a growing interest, before and after the outcome of the vote. The kind of shock related to the outcome of the Brexit referendum is leading many scholars in different disciplines to focus on the stance dynamics underlying this political debate, with the aim to understand the processes that led to the result. Social media data could contribute also to shed some light on such dynamics related to political sentiment [29]. An annotated corpus of tweets and news blogs about the BREXIT Referendum, collected in the month preceding the referendum date (June 23rd 2016), has been discussed in [6]. It includes sentiment and agreement/disagreement labels on isolated posts. Nevertheless, the goal of that study is different, mainly

[2]Detecting the gender of the author of a given tweet was also a sub-task to be addressed in the shared task.

targeted at forecasting the referendum outcome. Our main aim here is, instead, to investigate the the predictive power of community detection features in SD and the importance of taking a diachronic perspective on stance, that, at the best of our knowledge, has been never studied before within this research community.

Community detection and stance

Social media foster novel possibilities to investigate social networks embedding new forms of interpersonal relations. Research in *Network Science* have, thus, focused on exploring social networks, with the twofold aim to corroborating sociological theories and providing evidences to develop new ones [15,26,16,44].

In this work we exploited a classical method to analyze the network structure, which is commonly used in social media networks, Community Detection (CD), that consists in identifying groups or communities in a given large scale network [13]. Some recent works attempted to investigate the social media network structure in relation with sentiment information extracted from posted contents. Xu et al. [47] introduce the concept of *sentiment community*, trying to maximize both the intra-connections of nodes and the sentiment polarities using ratings of movies information collected from Flixster[3]. Deitrick et al. [8] combined SA and CD techniques on Twitter by using replies, mentions and retweets, hashtags, and sentiment classification of tweets to iterative increment edge weights in a social networks based on follower and friend relations. Those works showed that CD and SA can be mutually supportive. However, to the best of our knowledge, community detection techniques have been never used for detecting stance in Twitter. Some preliminary results in this direction have been reported in [23] that investigate the BREXIT debate for inspecting users' social network based on followings relations. This work, using the TW-BREXIT corpus for automatically predicting the stance of the increased number of unannotated users that took place in the debate on Twitter, showed that users having the same stance towards this particular issue tend to belong to the same social network community and that the neighbours are more likely to have similar opinions. In this paper, we present an enhanced stance model, enriched with the novel set of contextual features described in details in the next section, and present an in-depth evaluation of the model in a set of stance classification experiments on the TW-BREXIT corpus.

[3] http://www.flixster.com

3. Methodology

Our methodology comprehends two steps. First, we developed a novel annotated corpus for studying stance in a diachronic perspective, which takes into account evolution of users' stance over time. Then, a novel set of features related to community and diachronic evolution has been introduced and evaluated in a set of stance classification experiments. The development of the Twitter annotated corpus TW-BREXIT, with focus on the debate on the BREXIT referendum which took place in UK, is described in the next section. Here, we will quickly describe our features, with a focus on the rationale behind the choice of the set of novel features introduced to capture information related to community, diachronic and contextual facets of stance, which were also influencing the design of a novel annotation scheme.

3.1. Context-based features

3.1.1. Community features

According to the homophily principle observed in on line social media as well as traditional social network [28] people tend to bond with similar persons. Therefore, an interesting hypothesis to test in polarized political debate contexts is if people with the same stance belong to the same social media network community [5,7]. In particular, we assumed that the value homophily is involved [25], considering that Twitter users tend to bond with others who think in a similar ways, regardless of any differences in their status characteristics (i.e. gender, age, social status). Intuitively, the feature is based on the creation of a social network in witch a relation between two Twitter's users exits if one *follows* the other (in Twitter the term "to follow" refers to the specific relationship a user can create to be a follower of another user). After extracting the social communities from the network structure, the process results into the definition of the *community-context-based* feature. It consists in a binary feature vector including one element for each detected community; value is 1 when the community corresponds to the one the tweeter belongs to, 0 otherwise.

3.1.2. Diachronic evolution features

People's opinion is influenced not only by pre-existent ideology and party identification, but also by information about events happened during the political discussion [14]. Therefore, we hypothesize that the evolution of the political debate affects the stance

of each voter. Dividing the dataset of Tweets in discrete temporal phases delimited by significant events occurred around the consultation period, could allow to track the evolution of stance of users involved in the online debate. This assumption does not necessarily imply that users effectively change opinion, but that something changes in the way they write about the topic. The feature *diachronic-evolution-context-based* consists in a binary feature vector composed of an element for each considered time window. The value of the element corresponding to the time window in which the user posts the tweet is set to 1.

3.1.3. Common knowledge features

There is a general agreement on the idea that language cannot be investigated in isolation from culture and social organization. Such elements define an important context surrounding the event to be examined, and the use of resources embedding external knowledge can be important for a better interpretation [9]. The external knowledge could be extracted from resources nowadays available such as Wikipedia and DBpedia. To address this issue, we introduced two binary features considering the relations of friendship and enmity among the target of interest and the related entities such as politicians and parties mentioned in the text of the tweet as a signal of stance that should be taken into account: *party-stance-context-based*, a binary sub-feature vector of three elements (party_against, party_favour, party_neutral) considering the presence of a mentioned party in the text and its corresponding stance; *politician-stance-context-based*: a binary feature vector of three elements (politician_against, politician_favour, politician_neutral) considering the presence of a mentioned politician in the text and her corresponding stance. Moreover, we introduce another binary feature of two elements, *explicit-stance-context-based*, considering the words used for expressing the stance (in the context of BREXIT the words are "remain" and "leave"). Aggregating the three sub-features, the *Common knowledge features* feature consists is in a binary feature vector of seven elements.

3.2. Sentiment-based features

Stance detection is strongly related to sentiment [21,30,32,39,48]. We are not aware of sentiment analysis lexica retrieved specifically in the political domain; thus, we exploited a wide range of resources available for English [35]. We used a set of four lexica to cover different facets of affect, ranging from sentiment polarity of words to fine-grained emotional information [21]: AFINN [12], Hu&Liu [19], LIWC [38], and DAL [45]. AFINN was selected since contains several slang and profanity; Hu&Liu and LIWC, since they are widely used in tasks related to analysis of subjective information, and DAL in order to explore different emotional dimensions. Therefore, the *Sentiment-based* feature consists in a continuous feature vector of six element. Three elements respectively store the sum of the polarity calculated by AFINN, Hu&Liu, and LIWC (considering as 1 the presence of a positive word in the text and as -1 the presence of a negative one) and other three elements for storing the the sum of the values of each word for each of the three emotional dimensions explored by DAL: pleasantness, activation, and imagery.

3.3. Structural features

We also experimented structural characteristics of tweets taking into account the use of metadata and punctuation marks [11]. Therefore, the *structural-based* features consists in a continuous feature vector of nine elements containing the number of elements of the following structural characteristics extracted from the text: number of hashtags, number of mentions, and number of punctuation marks (i.e., frequency of exclamation marks, question marks, periods, commas, semicolons, and, finally, the sum of all the punctuation marks mentioned before). In addition, binary unigrams of bag of hashtags and of bag of mentions are also included.

4. The TW-BREXIT corpus

We analyzed the political discussion in United Kingdom about the European Union membership referendum commonly known as BREXIT, an abbreviation for "British exit". We gather about 5M English tweets containing the hashtag #brexit using the Twitter Stream API in the time-frame between 2016-06-22 and 2016-06-30 and we used the dataset in order to create a novel linguistic resource annotated for stance. A novel SD annotation schema has been applied to the data, focusing on users' stance in a diachronic perspective. First, we grouped tweets according to three 24-hour short and highly focused time steps forthcoming elections. For the ease of the reader, we defined each time windows by a name that corresponds to a relevant event that happened during the 24 hours:

– "Referendum Day" (RD): includes the 24 hours preceding the polling stations closing.
– "Outcome Day" (OD): includes the 24 hours following the formalization of referendum outcome.
– "After Pound Falls" (APF): includes the 24 hours after the financial markets' turbulence occurred three day after the formalization of the referendum outcome.

Then, we randomly selected a sample of 600 users over 5,148 that wrote at least 3 tweets in each temporal interval. We decided to require three tweets for each user in each time windows due to our assumption is that human annotators may guess more easily the user's stance considering a context of several tweets instead of only one. Therefore, we defined a *triplet* as a collection of three random tweets written by the same user in a given time interval. Finally, we created the TW-BREXIT corpus, which consists of 1,800 triplets. Overall, for each of the selected 600 users, we have three triplets in the corpus, one for each time interval.

4.1. Annotation Process

We employed CrowdFlower[4], a popular crowdsourcing resource, to annotate the corpus. Eligible annotators (i.e. "contributors") were required to live in the UK or in Gibraltar, since we wanted to be sure to they were directly involved in the political debate at issue, and aware about the local political situation. The proposed HIT (Human Intelligence Tasks[5]) request to the human contributors to annotate the user's stance on the target BREXIT (i.e. UK exit from EU). In particular, given a triplet posted by a user, they had to infer the user's stance. The instructions given to contributors for determining stance are shown below.

"What is the stance of the user that wrote those three messages?"

The available answers are:

– **Leave**: you think that the user is *in favor* of the UK exit from EU.
– **Remain**: you think that the user supports staying within the EU, being *against* the UK exit from EU.
– **None**: you could not infer user's stance on BREXIT due to:
  * all the messages are unintelligible
  * the user does not express express any opinion about the target
  * the user expresses opinion about the target, but the stance is unclear.

It is worth to be noticed that the classification task is very similar to the stance classification task formulated in Semeval-2016 [30]. Indeed, the labels "favor", "against", and "neither" can be considered equivalent to "leave", "remain", and "none", respectively, if used to define stance towards the target "UK exit from EU" (BREXIT). Similarly to Mohammad et al. [30], we use only a label "none" for identifying both cases when the annotator can infer from the tweet that the tweeter has a neutral stance towards and when there is no clue in the tweet to reveal the stance of the tweeter towards the target.

We required two judgments for each triplet. When the two contributors disagree on the stance evaluation of a triplet, we requested an additional judgment by a new contributor. Crowdflower provides a quality control mechanism to evaluate contributor reliability based on answers given to a set of test questions. Two human domain experts created 70 balanced test questions w.r.t. the three stance labels. We set the reliability threshold for our job, requiring that contributors should have correctly answered to at least 80% of the test questions proposed during the task.

4.1.1. Annotation Results

The trusted contributors that perform the annotation process were twenty-nine[6]. Crowdflower provides a measure of the agreement between contributors calculated as the average confidence obtained by each HIT, which results to be 91.04%.

In order to select the true label we used majority voting. Overall, the final gold TW-BREXIT resulted in 1,760 labeled triplets, after removing triplets in disagreement (where all three contributors provided different annotations). The corpus and the annotation guidelines are available to the community[7]. The unbalanced distribution for stance is not unexpected. Indeed, we used the hashtag #brexit for quickly collecting data, due to its wide use in the debate for marking posts which express different stance on the target.

[4] http://www.crowdflower.com/

[5] This is the expression for denoting questionnaires posted on a Crowdflower's job.

[6] Four contributors did not pass the quality control during the annotation task and they were considered untrusted.

[7] https://github.com/mirkolai/leave-or-remain

Leave	Remain	None
961 (51%)	236 (14%)	563 (35%)

Table 1
Label distribution

However, also apparently neutral hashtags are often biased, and a recent study [18] shows that most of the tweets containing #brexit were posted from people that express stance in favor of BREXIT. The bias is not critical here, since our focus is on stance detection and evolution, and not in predicting the referendum outcome.

Most interestingly, the label distribution changes over the time as shown in Figure 1.

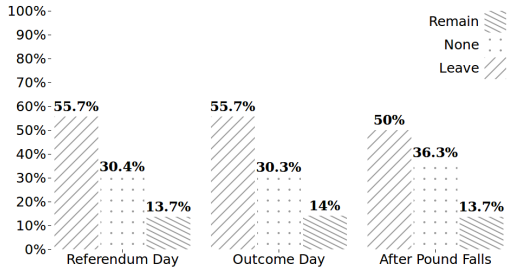

Fig. 1. Label distribution over the time

The label distribution changes consistently in the "After Pound Falls" temporal interval, when we observe a drop in the percentage of label "leave", and an increase for what concerns the label "none".

We also explored if users' stance changes over time. We find that 57.66% of the users was labeled with the same stance in all three intervals (37.16% leave, 15.5% none, 5% remain). Very interestingly, 42.33% of users' stance changes across different time intervals. In particular, 9.5% of users' stance varies from "leave" to "none" (7% leave $\rightarrow$ leave $\rightarrow$ none; 2.5% leave $\rightarrow$ none $\rightarrow$ none). Overall, the analysis of the corpus supports the conjecture that there is a relation between diachronic evolution and stance. Furthermore, we observed a progressive decrease of the agreement (93.72 in the RD interval; 90.61 in the next OD interval; 88.78 in the last APF interval). The analysis of the manual annotated corpus could imply that something changed in users' communication style up to the point that annotators used a different label to annotate the stance of the same user in different time windows. Furthermore, the progressive decreasing of the agreement among annotators could mean that users tend to express their opinion in a less explicit way in the following phases of the debate. Also Messina et al. [29] show how people's views over technical arguments about BREXIT could quickly change due to emotive events (i.e. national humiliation). For this reason we are very encouraged by this preliminary analysis to keep trying to investigate stance in a diachronic perspective.

5. Experiments and Results

In this section, we first describe the methods used for detecting communities in the social network and for gathering common knowledge. Then, we report on three types of classification experiments for SD on the TW-BREXIT corpus: stance at the triplet level, at tweet level, and at tweet level in different temporal intervals. The source code of the experiments are available for the community[8].

5.1. Retrieving common knowledge

For retrieving common knowledge, we created a gazetteer of relations (politician $\rightarrow$ stance and party $\rightarrow$ stance) using sources freely available such as Wikipedia and DBpedia. First, We extracted the declared stance (leave, remain, none) of UK's, Northern Ireland's, and Gibraltar's parties from Wikipedia [9]. Second, we gathered political affiliation of each politician in DBpedia, and we inferred politicians' stance from the stance of the party they are affiliated to. We populated the gazetteer with politicians' and parties' alias provides by DBpedia and the inferred stances. One problem we had to face is that our knowledge sources are not constantly updated. Several politicians changed their political affiliation during their career and we couldn't infer the current political affiliation. To deal with such cases, if a politician was affiliated at party with a different stance in the past, we infer several stances for her. Overall, we automatically obtained 225 alias for 17 parties and 5945 alias for 1838 politicians (48 of them having multiple stances due to the reasons highlighted above). The feature extraction consists in giving the values of the elements of the feature vector *Common knowledge features*. For example, the element party_against is set to one if the text contains

[8] https://github.com/mirkolai/leave-or-remain

[9] https://en.wikipedia.org/wiki/United_Kingdom_European_Union_membership_referendum,_2016

a party identified to have campaigned for a "remain" vote.

5.2. Community detection process

First, we used Twitter API "GET friends/id" in order to gather the list of the *friends* (the users that a user follows) of 5,148 users who posted at least three tweets containing the hashtag "brexit" in all the defined temporal intervals. Then, we created a graph that consisted of about 13M edges and 4M users and friends. We filtered a sub-graph consisting of about 200K nodes after removing the friends have less than 10 relations (we consistently reduced graph dimension). We decided to use the Louvain Modularity algorithm[10] in order to extract social media network communities, since it performs better in terms of computer time and modularity compared to other methods [3]. The algorithm extracts 6 communities; we added a seventh community for 195 users that were isolated from the graph after the filtering. Figure 2 shows the average of the distribution over the communities of the 600 users' stance resulting from the manual annotation process.

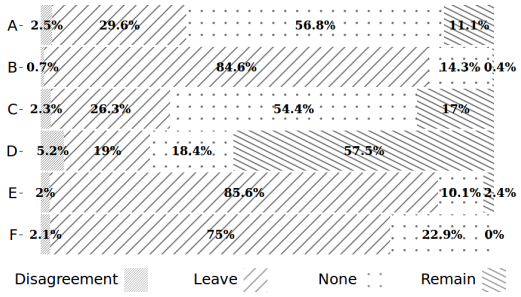

Fig. 2. The figure shows the users' stance distribution of users belonging to each community. Only the 600 users resulting from the manual annotation are considered. The distribution is expressed as the average over the three temporal phases.

We observed that the percentage of users' stance in community D is evidently biased towards the stance "remain" (about 57% users were annotated with the label "remain"); in communities B, E, and F towards the stance "leave"(more then 70% users with the label "leave"); in communities A and C towards the stance "none" (more then 50% users with the label "none"). We also noted that the disagreement among contributors is higher for the users belonging to the community D (the annotators do not reach an agreement for about more 5% of cases), maybe because the hashtag #brexit is biased in favor of BREXIT [18] and might have contributed to create more ambiguity during the annotation process in a community mostly composed by users annotated with the stance "remain".

[10] We used the software package NetworkX.

5.3. Classification experiments

We experimented the use of several supervised learning algorithms - Naïve Bayes (NB), linear support vector machine (SVM), Random Forest (RF), Decision Trees (DT) - on the TW-BREXIT[11]. In addition, we experimented with different feature sets for SD, and evaluated them performing 5-fold cross validation for each run. We used the macro-average of the AGAINST and FAVOR F1-score metrics and the baselines such as Majority Class, SVM-unigrams (word unigram features), SVM-ngrams (1-, 2-, and 3- word grams features and 2-, 3-, 4-, and 5- character grams features) proposed in Semeval-2016 [30]. The macro-average of the F1-score metrics was redefined, replacing labels "favor" and "against" with labels "leave" and "remain", respectively:

$$F_{\text{avg}} = \frac{F_{leave} + F_{remain}}{2}$$

We carried out three kinds of experiments:

- **Stance at the triplet level**: we looked for the better feature combination set in order to predict users' stance extracting features from the textual contents of the three tweets grouped in a triplet.
- **Stance at the tweet level**: we tried to find the best feature combination set in order to predict users' stance extracting features from a tweet isolated from a triplet inheriting triplet label.
- **Stance at the tweet level in different temporal intervals**: we used the best feature combination obtained in the experiment "Stance at the tweet level", by grouping single tweets from each triplet according to the temporal interval.

5.3.1. Stance at the triplet level

First, we experimented SD at the level of triplets. Here, the classifiers, similarly to the human annotators,

[11] We used the scikit-learn (`http://scikit-learn.org`) implementation of the machine learning algorithms with default parameters.

are trained with triplets - three tweets for each user in each temporal phase. In other words, the training and the test instances are triplets instead of single tweets. Thus, we performed features extraction concatenating the textual content of three tweets (instead of relying only on the single tweet). We experimented 63 different feature combination using six groups of features: *BoW*, *structural-based* (structural), *sentiment-based* (sentiment), *common-knowledge-context-based* (comm-know-cxt), *diachronic-evolution-context-based* (de-cxt), *community-context-based* (comm-cxt). Results are showed in Table 2.

Classifier	Feature set	F_{avg}
Baselines		
MC	-	35.25
SVM	unigrams	58.25
SVM	ngrams	60.14
Our Classifiers		
NB	BoW + comm-cxt	53.77
DT	comm-cxt	63.74
RF	comm-cxt	63.76
SVM	structural + sentiment + de-cxt + comm-cxt	**67.01**

Table 2
Best feature set on stance at triplet level

The *BoW* and *structural-based* features are relevant due to the presence of three tweets in a triplet (more words than in a text of a single tweet) respectively in Naïve Bayes and SVM. The *community-context-based* feature is significant especially in Decision Tree and Random Forest. In addition, all the best feature combinations for each classifier contain the *community-context-based* feature. The *diachronic-evolution-context-based* feature shows its relevance only in SVM. Table 3 shows the results obtained in the ablation test using linear SVM, i.e. the machine learning algorithm that achieved the best performance in the above mentioned experiments[12]. F_{avg} decreases of 14.6% and 0.12% removing singularly *community-context-based* and *diachronic-evolution-context-based* features respectively. Removing only the *common-knowledge-context-based* feature F_{avg} improved of 0.49%. Therefore, the *common-knowledge-context-based* feature does not improve F_{avg} and the *diachronic-evolution-context-based* feature is not decisive in results. Using the whole group of *context-based* features improves F_{avg} more than using only the *community-context-based* features (16.78%).

[12]Notice that we will report on ablation test results only for classification experiments at the triplet level, since this is here our main novel focus for what concerns the task of detecting stance.

Features	F_{avg}	Decreasing	Percentage decreasing
All	65.61	0	0%
All - context-based	54.60	-11.01	-16.78%
All - comm-cxt	56.03	-9.58	-14.6%
All - de-cxt	65.53	-0.08	-0.12%
All - comm-know-cxt	65.93	0.32	0.49%
All - sentiment	65.99	0.38	0.58%
All - structural	65.81	0.2	0.3%
All - BoW	65.66	0.05	0.08%

Table 3
Ablation Test (linear support SVM on stance at triplet level)

Wallace et al. [43] already noted how the mention of a specific political entity in a determined community could be a useful feature for irony detection on political debates. *Sentiment-based*, *structural-based*, and *BoW* features could be removed without significant influencing F_{avg}.

5.3.2. Stance at the tweet level

Second, we experimented SD at the level of tweet. The classifiers, differently from the situation faced by the human annotators, deal only with single tweets, isolated from triples, but inheriting the triplet label. We experimented the same 63 different features combinations used in the previous experiment. Table 4 shows the best feature set combination for each classifier.

Classifier	Feature set	F_{avg}
Baselines		
MC	-	35.25
SVM	unigrams	51.98
SVM	ngrams	52.66
Our Classifiers		
NB	sentiment + community-ctx	52.00
DT	community-ctx	63.75
RF	community-ctx	63.74
SVM	BoW + sentiment + de-cxt + comm-cxt	**65.68**

Table 4
Best feature set on stance at tweet level

SVM obtains the best result using *BoW*, *sentiment-based*, *diachronic-evolution-context-based*, *community-context-based* based features. Furthermore, it is important to highlight that all classifiers improve consistently the Majority Class. In addition, all the best feature combinations for each classifier contain the *community-context-based* feature. In particular, the *community-context-based* feature in Decision Tree and Random Forest singularly is the best feature. The feature *diachronic-evolution-context-based* seems to be relevant only in SVM. Moreover, Naïve Bayes did not improve SVM unigram and ngram baselines.

5.3.3. Stance in different temporal intervals

We experimented with the best features sets obtained in the previous experiment over the three different temporal intervals selected. We decided to carry out this experiment since we observed that the agreement between the annotators varies over the temporal intervals, in particular we observed a progressive decrease of the agreement (93.72 in the RD interval; 90.61 in the next OD interval; 88.78 in the last APF interval). We hypothesized that the proposed features could detect cues which influence annotators' agreement over the time windows. Table 5 shows the results obtained by the features set combinations over the three temporal phases.

Classifier	Feature set	RD	OD	APF
Baselines				
MC	-	35.97	36.07	33.6
SVM	unigrams	54.46	49.93	45.77
SVM	ngrams	56.60	48.49	45.35
Our Classifiers				
NB	sentiment + comm-cxt	47.56	54.20	43.25
DT	comm-cxt	67.07	61.95	**62.22**
RF	comm-cxt	67.14	61.96	62.07
SVM	BoW + sentiment + comm-cxt	**69.75**	**62.03**	58.30

Table 5
Results in different temporal intervals

We did not use the *diachronic-evolution-context-based* feature, since the feature would become superfluous after grouping tweets by temporal interval. The F_{avg} changes over the time for each classifier. We observe that F_{avg} decreases for SVM and Naïve Bayes in the time interval "After Pound falls" as with annotators' agreement.

6. Conclusions

In this work we investigated the use of several context-based features related to common knowledge, social network community, and diachronic evolution in the stance detection task. We focus on the political debate on BREXIT in Twitter. A novel annotation scheme, which takes into account the temporal evolution of stance has been proposed and applied to our social media data. Results of classification experiments confirm that the entire group of context-based features is very relevant for the stance detection task, in particular the community based feature. Then, the analysis of the annotated corpus confirms that users' labeled stance may consistently change over temporal phases. We can only speculate that users not only could effectively change opinion, but also they could change their communication style, probably influencing annotators' choices. However, even if a deeper investigations on the possible causes of the opinion shifts are needed, calling also for competencies from other disciplines such as sociology or social psychology, this finding confirms that it is interesting to investigate SD in a diachronic perspective, since opinion fluctuations within the debates occur even in short time spans. It also suggests that people's stance depends not only on their pre-existent ideology and party identification, but also on the information about events happened during the political discussion [14].

We are currently planning to investigate the debate about Brexit during the last three year particularly focusing on European elections in order to explore a wider time windows.

As a future work, a deeper linguistic analysis is also needed in order to clarify what has been observed here. In particular, we would like to investigate not only stance, but also the communications among unlike-minded users focusing on the use of rhetorical devices, such as sarcasm [40], and hostility. Here, we considered a static network structure, whereas on this future line of research it will be interesting to consider the *mutual evolution of stance and network structure* for observing the dynamics of the polarization within the debate. Furthermore, the proposed method could be useful for training intelligent systems capable of gathering and analyzing real time user generated contents from social media for supporting automatic prediction of citizens' stance toward public issues, based on big amount of people's opinions spontaneously expressed. Thus, policy makers and public administrators could better meet population's needs and could have a

data-driven support for developing policies to prevent strong polarization among different groups in society.

Acknowledgments

The work of P. Rosso was partially funded by the Spanish MICINN under the research projects MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech(PGC2018-096212-B-C31) and PROMETEO/2019/121 (DeepPattern) of the Generalitat Valenciana.
The work of V. Patti and G. Ruffo was partially funded by Progetto di Ateneo/CSP 2016 *Immigrants, Hate and Prejudice in Social Media* (S1618_L2_BOSC_01).

References

[1] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 1–9, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] David Bamman and Noah A Smith. Contextualized sarcasm detection on Twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:10008, October 2008.

[4] Cristina Bosco and Viviana Patti. Social media analysis for monitoring political sentiment. In R. Alhajj and J. Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*. Springer, 2017.

[5] Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira, Jr., and Virgílio Almeida. From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 150–158, New York, NY, USA, 2011. ACM.

[6] Fabio Celli, Evgeny Stepanov, Massimo Poesio, and Giuseppe Riccardi. Predicting brexit: Classifying agreement is better than sentiment and pollsters. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 110–118, Osaka, Japan, December 2016. ACL Anthology.

[7] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. Political polarization on Twitter. In *International AAAI Conference on Web and Social Media*, 2011.

[8] William Deitrick and Wei Hu. Mutually enhancing community detection and sentiment analysis on twitter networks. *Journal of Data Analysis and Information Processing*, 1:19–29, 2013.

[9] Alessandro Duranti and Charles Goodwin. *Rethinking context: Language as an interactive phenomenon*. Cambridge University Press, 1992.

[10] Erick Elejalde, Leo Ferres, and Eelco Herder. The nature of real and perceived bias in chilean media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, pages 95–104, New York, NY, USA, 2017. ACM.

[11] Ash Evans. Stance and identity in Twitter hashtags. *Language@Internet*, 13(1), 2016.

[12] Finn Årup Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[13] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[14] Andrew Gelman and Gary King. Why are american presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 23(04):409–451, 1993.

[15] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on Twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, 2011.

[16] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, Jun 2008.

[17] Ángel Hernández-Castañeda, Hiram Calvo, and Omar Juárez Gambino. Impact of polarity in deception detection. *Journal of Intelligent & Fuzzy Systems*, 35(1):549–558, July 2018.

[18] P. N. Howard and B. Kollanyi. Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum. *ArXiv e-prints*, June 2016.

[19] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[20] Mirko Lai, Alessandra Teresa Cignarella, and Delia Irazú Hernandez Farias. iTACOS at IberEval2017:Detecting Stance in Catalan and Spanish Tweets. In *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL),CEUR Workshop Proceedings. CEUR-WS.org, 2017*, Murcia, Spain, september 2017.

[21] Mirko Lai, Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. Friends and enemies of Clinton and Trump: Using context for detecting stance in political tweets. In Grigori Sidorov and Oscar Herrera-Alcántara, editors, *Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23–28, 2016, Proceedings, Part I*, pages 155–168. Springer International Publishing, Cham, 2017.

[22] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance evolution and Twitter interactions in an Italian political debate. In Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems*, NLDB, pages 15–27, Cham, Switzerland, June 2018. Springer International Publishing.

[23] Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Extracting graph topological information and users' opinion. In Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl,

Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings*, pages 112–118. Springer International Publishing, Cham, 2017.

[24] Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter. *Data & Knowledge Engineering*, 2019.

[25] P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, and C. Page, editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, New York, 1954.

[26] David Lazer, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723, Feb 2009. 19197046[pmid].

[27] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 109–116, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[28] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[29] Enza Messina, Elisabetta Fersini, and Joe Zammit-Lucia. All Atwitter About Brexit: Lessons for the Election Campaigns, 2017.

[30] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics.

[31] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *CoRR*, abs/1308.6242, 2013.

[32] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3):26:1–26:23, June 2017.

[33] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[34] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, Heraklion, Crete, Greece, 2011. CEUR-WS.org.

[35] Malvina Nissim and Viviana Patti. Semantic aspects in sentiment analysis. In Fersini Elisabetta Pozzi, Federico, Enza Messina, and Bing Liu, editors, *Sentiment Analysis in Social Networks*, chapter 3, pages 31–48. Elsevier, 2017.

[36] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.

[37] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

[38] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. *Linguistic inquiry and word count: LIWC 2001*. Lawrence Erlbaum Associates, Mahwah,NJ, 2001.

[39] Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 159–169, Berlin, Germany, August 2016. Association for Computational Linguistics.

[40] Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132 – 143, 2016.

[41] Mariona Taulé, Maria Antònia Martí, Francisco Manuel Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti. Overview of the task of Stance and Gender Detection in Tweets on Catalan Independence at IBEREVAL 2017. In *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL 2017)*, volume 1881 of *CEUR Workshop Proceedings*, pages 157–177, Murcia, Spain, September 2017. CEUR-WS.org.

[42] Yannis Theocharis and Will Lowe. Does Facebook increase political participation? Evidence from a field experiment. *Information, Communication & Society*, 19(10):1465–1486, 2016.

[43] Byron C Wallace, Do Kook Choe, and Eugene Charniak. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1035–1044. ACL, 2015.

[44] Lilian Weng, Márton Karsai, Nicola Perra, Filippo Menczer, and Alessandro Flammini. Attention on weak ties in social and communication networks. *arXiv preprint arXiv:1505.02399*, 2015.

[45] Cynthia Whissell. Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language. *Psychological Reports*, 105(2):509–521, 2009.

[46] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[47] Kaiquan Xu, Jiexun Li, and Stephen Shaoyi Liao. Sentiment community detection in social networks. In *Proceedings of the 2011 iConference*, iConference '11, pages 804–805, New York, NY, USA, 2011. ACM.

[48] Zhihua Zhang and Man Lan. Ecnu at semeval 2016 task 6: Relevant or not? supportive or not? a two-step learning system for automatic detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 463–469, 2016.