Topic-centric Classification of Twitter User's Political Orientation

Anjie Fang[1], Iadh Ounis[2], Philip Habel[2], Craig Macdonald[2] and Nut Limsopatham[2]
University of Glasgow, UK
[1]a.fang.1@research.gla.ac.uk, [2]{firstname.secondname}@glasgow.ac.uk

ABSTRACT

In the recent Scottish Independence Referendum (hereafter, IndyRef), Twitter offered a broad platform for people to express their opinions, with millions of IndyRef tweets posted over the campaign period. In this paper, we aim to classify people's voting intentions by the content of their tweets—their short messages communicated on Twitter. By observing tweets related to the IndyRef, we find that people not only discussed the vote, but raised topics related to an independent Scotland including oil reserves, currency, nuclear weapons, and national debt. We show that the views communicated on these topics can inform us of the individuals' voting intentions ("Yes"–in favour of Independence vs. "No"–Opposed). In particular, we argue that an accurate classifier can be designed by leveraging the differences in the features' usage across different topics related to voting intentions. We demonstrate improvements upon a Naive Bayesian classifier using the topics enrichment method. Our new classifier identifies the closest topic for each unseen tweet, based on those topics identified in the training data. Our experiments show that our Topics-Based Naive Bayesian classifier improves accuracy by 7.8% over the classical Naive Bayesian baseline.

1. INTRODUCTION

Both citizens and politicians are increasingly embracing social media to disseminate information, particularly during significant political events and campaigns. Twitter emerged as an especially popular platform during the recent IndyRef held in September 2014—a vote that, if successful, would have created an independent Scotland—a secession of the 300 years union with Great Britain. This critical event attracted unprecedented levels of political activity both offline and online. Social media usage was essential to both campaigns, with more than 5.8 million tweets using the *indyref* hashtag within the year leading up to the vote [6]. Therefore, we propose here a technique to analyse the voting intentions of users, based on data mining and machine learning approaches. Indeed, the general approach we advance could be used to understand vote intentions in other major elections.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00
DOI: http://dx.doi.org/10.1145/2766462.2767833 .

To analyse voting intentions, we capture two months (August 1 to September 30, 2014) of Twitter data related to the IndyRef. To form a ground truth, we label users based upon hashtags appearing in their tweets, and we verify the reliability of this approach using the users' followee networks. Those in favour of an independent Scotland used hashtags such as *VoteYes* and *YesScotland*; those opposed used *BetterTogether* and *VoteNo*–which serves as our ground-truth. After removing the hashtags from these tweets, we then focus on the remaining terms, treating each term as a feature. However, the referendum created an evolving discourse, with different topical themes (such as *oil*, *currency*, and *debt*), which make the accurate classification of users' voting intentions more challenging. For instance, the word (feature) "change" is indicative of a "No" voter in the *currency* topic, and of a "Yes" voter in the *nuclear weapons* topic. That is, there was a significant discussion over whether Scotland would need to "change" its currency if it obtained independence, while the "Yes" camp purported that the nuclear arsenal base could "change" in an independent Scotland. The dichotomy of the term "change" in indicating voting intentions across different topics highlights the main benefit of our approach. Indeed, this paper contributes to the use of topical clusters to identify the topic of discussion in a tweet and subsequently to classify users' voting intentions. Our approach, called *Topics-Based Naive Bayesian* (TBNB) demonstrates marked improvements over a classical Naive Bayes (NB) classification baseline.

2. BACKGROUND AND RELATED WORK

Recently Al Zamal et al. [1] focused on the inference of latent attributes, such as age, gender and political orientation, based on textual and retweeting features. They achieved a high accuracy (90%); however, Raviv et al. [4] demonstrated that classification of political orientation was still a difficult problem and that the earlier result was exaggerated since it used easily classifiable political data. Conover et al. [5] showed that users' tweeting behaviours–such as the actions of re-tweeting, mentioning, and replying–can be used to classify the political polarization of users. In contrast, instead of leveraging the users' profile data or tweeting behaviours, we focus on the content of tweets to classify the users' voting intentions. We use as a starting point a classical Naive Bayesian (NB) classifier, which is an application of Bayes theorem within a probabilistic model to capture the conditional class probabilities of each feature. Since the number of features can be very large, it is common to use feature selection approaches to prune the features. For example, the following feature selection approaches are commonly

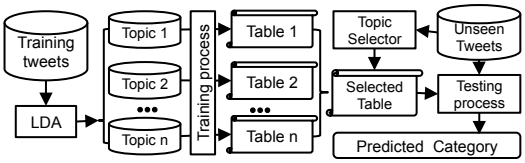

Figure 1: Data flow for TBNB.

used for the NB classifier: $Frequence(word)$ (denoted FR), $LogProbRatio(word)$ (LR), $ExpProbRatio(word)$ (ER), $OddsRatio(word)$ (OR) and $weightedOddsRatio(word)$ (WOR) (see [7]). OR is reported to obtain a high accuracy for the Naive Bayes classifier [7]. Each selection approach ranks and selects the F informative features (in our case we treat each term in a tweet as a feature) based on the training data (e.g., $F = 1000$). Of course, not every selected feature will appear in the unseen test tweets - we denote the number of such "activated" features as F_{test}. For instance, a testing tweet containing "Scotland has remained in the media spotlight throughout 2014" has 9 terms. If only "Scotland", "remained", "media" and "spotlight" were selected as features, the number of activated features is $F_{test} = 4$.

3. TOPICS-BASED NAIVE BAYESIAN

The IndyRef discussions on Twitter revolved around a number of topics, for which people's opinions usually reflected their vote intentions. For example, many "Yes" voters believed that revenues derived from the North Sea *oil* fields belonged to Scotland and could sustain it. On the other hand, many "No" voters argued that these sources were insufficient in the long run. A feature's **dissimilarity** represents the usage difference of this feature across topics, and the difference in usage of "oil" across different topics is therefore high. For a given topic, a feature's **variance** refers to the difference of the conditional probabilities of the occurrence of such a feature in different categories. For example, the conditional probability of "oil" in the "Yes" category is higher than in the "No" category. Typically, the feature selection approaches select features with higher variances between categories. Thus if a feature differs between topics, it will be treated as different features in the TBNB model. Thus TBNB can capture term dependencies between topic and user voting intentions. On the other hand, since the essence of the NB classifier is to learn those features with high variance from the categories, the TBNB classifier is believed to work better by leveraging both the features' dissimilarities across topics and their variances in the categories.

We assume that a single tweet involves a single topic. In the training step, the topics are first detected by Latent Dirichlet Allocation (LDA), a probabilistic graphical model introduced in [3]. For each topic, a corresponding probability table is produced, where each feature has two associated conditional probabilities related to the two possible voting intentions ("Yes"/"No"). Consequently, during the training step, we produce as many feature tables as the number of used topics. In the testing step, we treat a user as a virtual document and this document contains the users' tweets. For each tweet in the user's virtual document, the topic that is closest to the tweet's content is selected. As a topic can be represented by a mean vector of its tweets' vectors, the closest topic can be selected by computing the cosine similarity between the vector representation of the unseen tweet and the vector representations of the topics. We use standard TF vectors to represent both the tweets and topics. Terms in an unseen tweet are then examined using the probability table generated during the training step for the topic with which this tweet is associated. In this way, terms in different tweets are treated differently based on their associated topics, and the TNBN classifier applies, for each unseen tweet, those features that were learned from the corresponding topic. The detailed TBNB algorithm is presented in Algorithm 1. An overview of the whole TBNB classification process is shown in Figure 1. Note that the feature selection approaches can naturally be applied to the TBNB classifier. For example, if F (see Section 2) is set to 1000, the top 1000 features learned from each topic are selected.

Algorithm 1 Topics-Based Naive Bayesian (TBNB).

$topic_n, n = \{1, 2, .., N\} \leftarrow topic\_detection(tweets_{training})$
$n \leftarrow 1,\ c = \{\text{"}Yes\text{"}, \text{"}No\text{"}\}$
for $n \leq N, n++$ **do**
  if w **in** $topic_n$ **then**
    $p(w|c_i) = \frac{number\ of\ the\ word\ w\ in\ topic_n\ in\ c_i}{total\ number\ of\ word\ w\ in\ topic_n}, \forall c$
    $probability\_table_n.add(p(w|c_i))$
  end if
end for
$ProbProduct_{c_i} \leftarrow 1$
for $tweet$ **in** $user_{testing}$ **do**
  $n \leftarrow topic\_selector(tweet_{testing})$
  for w **in** $tweet$ **do**
    if w **in** $probability\_table_n$ **then**
      $ProbProduct_{c_i} \times = p(w|c_i)$
    end if
  end for
end for
$p(c_i) = \frac{number\ of\ users\ belonging\ to\ c_i}{total\ number\ of\ users}$
$class(user) = argmax_{c_i}(p(c_i) \times ProbProduct_{c_i})$

Table 1: Topics and associated terms in the IndyRef.

Topic	Tweets%	Associated Terms
currency	20.25%	currency,money,change,pay,future
salmond	15.88%	salmond,alex,debate,audience,answer
glasgow	10.95%	glasgow,team,games,great,gold
women	9.82%	patronisingbtlady,women,undecided
oil	7.91%	oil,sea,privatisation,billion,gas,cuts
fear	7.87%	country,future,voting,fear,change
lastnight	7.32%	tonight,undecided,time,wearenational
debt	7.03%	scottish,debt,government,share,pay
weapon	6.84%	nuclear,weapon,clyde,year,glasgow
edinburgh	6.13%	edinburgh,johnjappy,minister,time

We use the LDA implemented in Mallet[1]. We investigate various topic numbers ($T = \{5, 10, 20, 30\}$). Table 1 shows the topic terms extracted using LDA for 10 topics. For readability purposes, the first column of Table 1 provides the general theme of the extracted topic[2]. For example, we can see that tweets related to *currency* and *oil* were common. Other oft-used topics and features included references to Alex Salmond, who was both the leader of the Scottish National Party (SNP) and the "Yes" campaign.

4. REFERENDUM DATA

The corpus pertaining to the Scottish Referendum event was collected from the Twitter network by searching for a number of referendum-specific hashtags (e.g. #IndyRef) and associated terms (e.g. 'vote', 'referendum') using the Twitter Streaming API[3]. The (uncompressed) 33GB dataset

[1] http://mallet.cs.umass.edu/ [2] These themes are manually annotated. [3] https://dev.twitter.com/

contains 6 million tweets from over 1 million users collected from August 1, 2014 to September 30, 2014.

In our dataset, 79.7% of users posting tweets with more than one. The most commonly used hashtags indicating the users' voting intentions are listed in Sets 1 and 2 below. As can be seen, certain hashtags were associated with a "Yes" vote, and others with a "No" vote. To reduce sparsity, we retain only users with more than 30 tweets posted during the timeframe of the collection. To generate our ground truth, we assume that if a user's tweets are only tagged by hashtags in Set 1, then this user is labeled as a "No" voter. Similarly, if a user's tweets contain only hashtags in Set 2, then the user is labeled as a "Yes" supporter, favoring independence.
Set 1: #NoBecause, #BetterTogther, #VoteNo, #NoThanks
Set 2: #YesBecause, #YesScotland, #YesScot, #VoteYes

Using this method, we find 5326 "Yes" users and 2011 "No" users. Together these 7337 users account for more than 420K tweets. After labelling, all hashtags in Set 1 and Set 2 are removed from their original tweet text. The resulting tweets constitute our classification dataset. We use this dataset to examine the usefulness of enriching the NB classifier with the extracted topics. Without the hashtags, the classification task is naturally more challenging, but importantly, the resulting generalisable classifier does not require the presence of hashtags.

Next, we verify our ground-truth's reliability using the users' followee networks. In particular, members of the Conservative Party (CONV) were staunchly opposed to independence, with post-election surveys showing that 95% of Conservatives voted "No"[4]. Thus, we argue that if a user mainly follows Conservative politicians, this person is likely to be a "No" voter. In contrast, 86% of SNP party voters favoured independence[4], and hence if a user follows SNP politicians, their vote intention is more likely to be "Yes"–in like manner to a previously used method to classify users' political orientation [2]. We then examined the networks of the 7337 users in our dataset, and used the Twitter REST API[3] to identify who these users follow among the 536 public Twitter accounts corresponding to Members of the British (MPs) or Scottish (MSPs) Parliaments. We use two verification approaches, denoted c_{V1} and c_{V2} for verifying the reliability of our ground truth: c_{V1} assumes an exclusive followee membership, while c_{V2} assumes a marked tendency to follow politicians of a given political party, namely:

$$c_{V1}(u) = \begin{cases} \text{“}Yes\text{”} & \text{if } n_{CONV}(u) = 0 \ \wedge \ n_{SNP}(u) > 0 \\ \text{“}No\text{”} & \text{if } n_{CONV}(u) > 0 \ \wedge \ n_{SNP}(u) = 0 \end{cases}$$

$$c_{V2}(u) = \begin{cases} \text{“}Yes\text{”} & \text{if } n_{SNP}(u) - n_{CONV}(u) > 20 \\ \text{“}No\text{”} & \text{if } n_{CONV}(u) - n_{SNP}(u) > 20 \end{cases}$$

where $n_p(u)$ is the number of times user u follows a politician (MPs/MSPs) of party p. We test our ground truth by comparing a user's label allocated using the hashtags versus that allocated using the two verification methods. If the two labels are concordant, then the user voting intention is said to be verified, i.e. it is likely to be correct. Table 2 reports the agreement statistics between our hashtag labelling method and the two verification methods. Comparing the hashtag labelling method with the two verifications, we find that c_{V1} verifies more users than c_{V2}, but shows lower agreement (c.f. Cohen's Kappa) than c_{V2}. Overall, we find, of the 6332 users verified by c_{V1} or c_{V2}, 87% can be verified into "Yes" or "No" voters, demonstrating that our ground-truth produced by the hashtags labeling method is reasonable and reliable.

[4] http://lordashcroftpolls.com/ 2014/09/scotland-voted/

Table 2: Ground-Truth agreement.

	Verified Users	Agreement Number	Agreement Precision	Kappa
c_{V1}	6339	5424	85.6%	66.2%
c_{V2}	684	619	90.5%	80.0%
$c_{V2} \cup c_{V2}$	6632	5770	87.0%	71.8%

5. EXPERIMENTS

We use the dataset described in Section 4 to compare the performances of the NB and TBNB classifiers. To assess the impact of parameters used in both classifiers, we vary the number of selected features F and the deployed feature selection approach for both NB and TBNB. We also vary the number of topics T in the TBNB classifier. Since the number of unique terms in our collection is 200K, we vary $F = \{5K, 10K, 20K, 50K, 100K, 120K, 150K, 180K\}$ for NB, while, for TBNB, as F depicts the number of features selected for each topic (i.e. the total number of features would be $F \times T$), we do not experiment with $F > 100K$[5].

We use a 10-fold cross validation process over the 7337 users of our dataset to evaluate the performances of the NB and TBNB classifiers. In particular, we use the following performance indicators:
Indicator 1: Accuracy, the standard classification accuracy measure.
Indicator 2: Average Number of Activated Features F_{test}. For an unseen Twitter user, we concatenate their posted tweets into a virtual document and count the number of selected features activated in the virtual document. We average these numbers across the 10 folds to obtain F_{test}. Intuitively, the higher F_{test}, the greater the confidence in the predicted category.
Indicator 3: Average Rank of the Activated Feature R_{test}. Each feature has a rank position ranked by the applied feature selection approach. This indicator represents the average rank position of all testing features of all users in the 10 folds. Intuitively, it reflects the average effectiveness level of the activated features.

We use the three indicators to investigate and explain the performances of the NB and TBNB classifiers, as well as to validate our hypothesis that the TBNB classifier will outperform NB on our IndyRef dataset. In particular, we answer the following related research questions: (i) how effective are the feature selection approaches on the used dataset?; (ii) what is the effect of F_{test} and R_{test} on the Accuracy performances of the classifiers?

Figure 2(a)-(g) shows the performances of the NB and TBNB classifier when varying the parameters of the classifiers. As a baseline, we use NB without feature selection (NB_NO). This baseline has at least a comparable performance to both Support Vector Machine (SVM) and Decision Tree-based (DT) classifiers. Both NB and TBNB perform poorly when F is low. However, all TBNB classifiers markedly outperform the NB_NO baseline when F ranges from 10K to 50K. The highest accuracy of TBNB (90.4%) is achieved when applying the WOR feature selection approach (TBNB_WOR) with T=10 and when the FR feature selection approach is deployed (TBNB_FR) with T= 5. This is a 7.8% improvement over the NB baseline (82.6%). When varying the number of used topics (T), we note that the performance of the TBNB classifier generally increases as T increases. However, once T reaches 30 topics (see Figure

[5] In our dataset, no topic has more than 100K features.

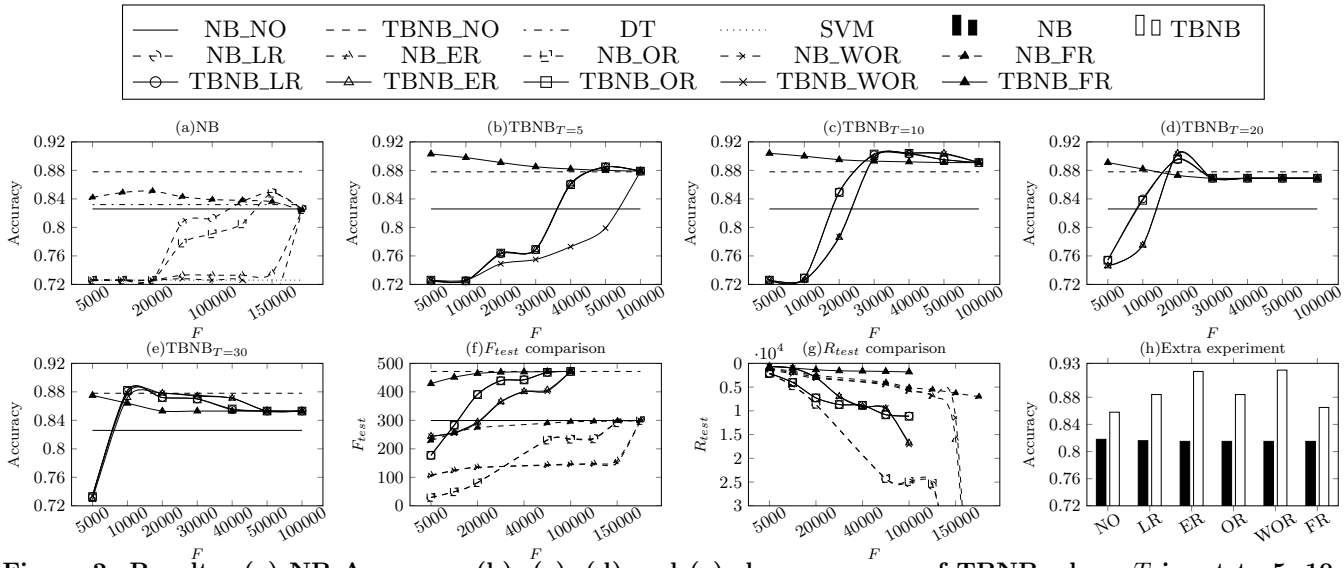

Figure 2: Results: (a) NB Accuracy; (b), (c), (d) and (e) show accuracy of TBNB where T is set to 5, 10, 20 and 30 separately; (f) and (g) show F_{test} & R_{test} for both NB & TBNB classifiers (T=10) while varying F; (h) shows an experiment on a 2nd dataset.

2(e)), the accuracy of TBNB starts decreasing while still outperforming the NB_NO baseline around the IndyRef. This suggests that the tweets corpus reflects 10-20 main discussion themes. On the other hand, each NB or TBNB classifier with feature selection approaches has an optimal F. For instance, the optimal F of NB_OR is 150K while that of TBNB_FR is 5K.

We first contrast the feature selection approaches for the NB and TBNB classifiers. Figure 2(f) shows that the average number of activated features (F_{test}) is lower for the NB classifier across all feature selection approaches than for TBNB with the same feature selection. This demonstrates that the TBNB classifier activates more features in the virtual document of the user, thereby improving its confidence in the voting intention classification. Unlike in previous work where the OR feature selection approach performs best (see Section 2), we found that the WOR and FR feature selection approaches are the most effective on our dataset.

Next, we consider the features selected and activated by each of the classifiers. Firstly, for NB, Figure 2(a) shows that increasing the number of features (F) increases the accuracy, until F reaches an optimal value, and decreases thereafter. The same conclusion is true for TBNB, e.g. for 10 topics (Figure 2(c)). However, contrasting Figures 2(a) & (c), we see that TBNB exhibits higher accuracy than NB, despite using less features (F). Indeed, we observe from Figure 2 (f) that the number of features activated in the unseen tweets (F_{test}) for a given F value is higher for TBNB than for NB - i.e. the classifier has more feature evidence to work with. Moreover, the average rank of those features selected (R_{test}, Figure 2(g)) increases as F increases. Hence, the relatively higher and stable F_{test} and R_{test} values observed for TBNB, in comparison to NB, are indicative of its higher accuracy. In summary, the advantage of TBNB over NB is that the topic-based features are more useful, leading to higher accuracies.

Finally, to show the generalisation of TBNB, we use a second dataset with 6234 labelled users, collected from an earlier period (i.e. July 25 to August 25 2014). For both NB and TBNB, we learn the parameters F and T from the first dataset with the 7337 users. We then use the learned parameters in a 10-fold cross validation on the second dataset. From Figure 2(h), we observe that the TBNB classifier outperforms NB in terms of accuracy, with and without the feature selection approaches, by up to 10.3%. Overall, our experiments validate our hypothesis in Section 5, namely that TBNB will outperform NB on the IndyRef dataset.

6. CONCLUSIONS

We classified the users' voting intentions on Twitter during the IndyRef. We noted that the users tended to focus their discussions on a specific set of topics, reflecting their voting intentions. As a consequence, we proposed to enrich the Naive Bayes classifier by leveraging the underlying topics covered in the tweets. Our proposed approach leverages the dissimilarity of the features across the topics, and their variance across the voting categories to increase the classification confidence. Our results demonstrate the effectiveness of our resulting TBNB classifier on two datasets with and without the use of feature selection approaches. In the future, we plan to analyse the effect of the evolving discussions on the users' voting intentions over time.

7. REFERENCES

[1] F. Al Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *ICWSM*, 2012.

[2] P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1), 2015.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] R. Cohen and D. Ruths. Classifying political orientation on Twitter: It's not easy! In *ICWSM*, 2013.

[5] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on Twitter. In *ICWSM*, 2011.

[6] L. Crossley. From Fife to Fiji: Amazing Twitter heatmap shows how the scottish independence referendum has been followed around the world in the past 30 days. *Daily Mail*, 18 Sep 2014.

[7] D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes. In *ICML*, 1999.