Twitter Sentiment Analysis, 3-Way Classification: Positive, Negative or Neutral?

Mestan Fırat Çeliktuğ
Computer Engineering
Bilkent University, Gazi University
Ankara, Turkey
firat.celiktug@bilkent.edu.tr, firat.celiktug@gazi.edu.tr

***Abstract*—People face with the huge amount of information on each day with the advent of big data era. The data amount stored and processed by Facebook, Twitter and other big social networks store (e.g. Instagram) is massive in those days. Online social networks provide great opportunity for propagation of almost any type of information. It's actually much much easier to disseminate an idea/knowledge than previous times. Naturally, this creates information validity and immediate curiosity about mass evaluation problem in general. In this regard, sentimental polarity detection in social media (e.g.**

Classification of a tweet as negative or positive or neutral) is highly valuable for certain institutions, organizations. The study's main focus is to classify negative, positive and neutral approaches of three (3) annotated twitter datasets. Effect of oversampling, unigram features and other features on overall and class-based accuracy ratios is worked on the datasets. Baseline is reached in dataset-2 experiments. 88% overall accuracy was observed in dataset-1 experiments which outperforms the prior art.Unigram features has shown significant effect on overall accuracy, class-based accuracy balance.

***Index Terms*—3-Way Sentiment Analysis, Polarity detection, Social Media Analysis, Twitter Analysis, Twitter Sentiment Analysis.**

I. Introduction

Online real-time information sharing is relatively new concept due to recentness of online social network (OSN) paradigm in civilization history. Before the OSNs, there were television channels, journals, gazettes and many other information sharing devices. Each of them generally gave information with significant latencies in general. Social media community is generating pseudo real-time data frequently and enormously.

Nevertheless, main objective of evaluating the information is to decide whether to trust or untrust the information. Since, people are psychologically curious about whether the information is credible or/not and/or right or/not and/or fake or/not. Besides personal trust, many institutions and organizations try to capture what is happening in peoples minds.

Tweets contain rich information about peoples preferences, users usually discuss between each other and declare their opinions in twitter. Besides, tweets are small in length, and hence they are relatively unambiguous. In those regards, data scientists especially try to make the analysis on maybe the most popular micro-blogging site Twitter, also it's more comprehensive and public compared to Facebook where social interactions are often private.

Twitter community reactions usually differ in face of information and of misinformation. Misinformation is generally more questionable than the valid information. There could be more negative sentimental reaction against misinformation. So, sentimental polarity detection in twitter could be very beneficial in scope of misinformation analysis and in many other credibility analyses.

In this study, the 3-way sentimental polarity classification (i.e. Twitter sentiment analysis) of three (3) annotated tweet dataset is presented.

The rest of article is organized as follows. Section II describes motivation of the study. Section III outlines and describes the 3-way sentimental polarity classification methodology adopted. Section IV presents and discusses the empirical results. Section V summarizes and concludes.

II. Motivation

Twitter sentiment analysis has numerous benefits and wide application area. Not restrictively but selectively, its exemplary applications and benefit would be described.

Managing a brand or a political campaign may require to keep track of your institutional popularity, the sentiment analysis provides a convenient way to take the pulse of the tweeting public.

Numerous business decision-making process demands automated sentimental polarity detection. For instance, sale decision of a candidate/actual product could be finalized as a result of the user satisfaction survey(s) and the sentiment analysis.

The evaluation of OSN information shared in a pseudo-real time has a place in contemporary society. People are not usually capable to assess automatically whether the information is fake or credible.Automated credibility/fakeness analysis would be so useful for increasing the benefit from the OSN information (e.g. Tweets shared on Twitter). In this regard, the sentiment analysis is highly valuable as a prior work of the credibility/fakeness analysis.

All in all, twitter sentiment analysis has practical and research value in numerous applications.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

III. Methodology

The sentiment analysis is a classification job. Positive, negative, neutral are three classes in the sentiment analysis. The analysis would be made via using machine learning (ML) and information retrieval (IR) methods, additionally human operators could be utilized for validating the results. In narrow scope of our study, we would use ML and IR methods for the analysis.

The methodology consists of two main steps in general. First step is feature extraction from the annotated tweets for the analysis. Second step is to learn how to classify the sentimental polarity using chosen ML algorithms (Table-V).

In this study, three (3) labelled datasets (Table-I) which contain three main sentimental polarity classes (i.e., Positive, negative and neutral.) is utilized. Features are extracted from tweet text. Subsequently, the sentimental classification experiments with/out oversampling is done for each ML algorithm in case of any unbalanced distribution of sentiments on the dataset (Table-II, III).

A. Dataset Description

We have three (3) annotated datasets having the main sentimental polarity classes. The datasets (Table-I) contain topic and context information. Sentimental polarity distributions of the datasets is presented in Table-II.

TABLE I
Dataset Description

No	Dataset
1	Twitter Sentiment Corpus of Sanders Analytics ([5])
2	Expression of feelings about US Airline Travelling in February 2015 on Twitter ([6])
3	First 2016 GOP Presidential Debate Twitter Sentiment Dataset ([7])

TABLE II
Dataset Sentimental Polarity Distribution

Dataset No	Neutral	Negative	Positive	Irrelevant
1	2333	572	519	0
2	3099	9178	2363	0
3	3109	2304	8460	0

The sentiment analyses are made separately on each dataset. Training and test sets are derived from the given corpuses. Shuffling for randomization is utilized during acquisition of training and test sets. Dataset is always divided into two parts which are equal or almost equal as training set and test set. Numerically speaking, 50% of each corpus is used as a training set, other 50% of each corpus is used as a test set in a randomized way by shuffling.

Ground truth is the annotated datasets (Table-I). Accuracy calculations is done based on those datasets.

When there is a dominance of any class, this kind of unbalanced distribution of tweets yielded to make oversampling in order to see the effect of the oversampling to the accuracy rates. Since, there is a danger of not learning well small sentimental polarity classes having less amount of tweet instances. Oversampling the tweets is made with shuffling for randomization and omitting dominance of examination order of each tweet. Oversampling statistics of each dataset is presented in Table-III. As seen from Table-III, all of minority class tweet amounts is made equal to majority class.

TABLE III
Dataset Sentimental Polarity Distribution in Oversampling

Dataset No	Neutral	Negative	Positive	Irrelevant
1	2333	2333	2333	0
2	9178	9178	9178	0
3	8460	8460	8460	0

B. Feature Engineering: Background and Application

Utilized feature extraction technique is based on two prior works ([9], [10]) and one prior application. Mainly, there are five (5) types of features(i.e., n-gram features, lexicon features, part-of-speech features (POS), micro-blogging features, 100-senti features). There is a critical importance of feature selection and analysis in this abundance of features. It could yield to success or failure in general.

There is not a clear view of success of each feature type (Table-IV). Features in one prior application and unigram features are utilized. Two different approaches exist in terms of n-gram features extraction. In first approach, there is use of huge set of features ([10]), in second approach, top-1000 features have been used for experiments ([9]). Only unigram features is used via former approach. Effect of unigram features on accuracy is observed by making experiments with/out unigram features.

Best accuracy ratios seen so far are changing between 70%-80%. They are not exact baselines (are approximate baselines) due to fact that there are different approaches in experiments. Especially, major differences are sizes of test sets and of training sets, training size/test size proportion, method of collecting training datasets and also method of achieving ground truth. Without any unigram features, accuracy ratios achieved in the study are 71% (dataset-1), 68% (dataset-2) as overall accuracy. It was reached somehow to baselines. With unigram features use and oversampling, 88% accuracy ratio is obtained in dataset-1. Its clearly outperforming the baselines.

Feature selection could ease the classification job by decreasing storage complexity and increasing the efficiency in general sense. Feature selection made on dataset-1, in order to observe relative importance of each feature and to reduce the dimensionality by looking at many subset of features, has shown that a found subset of features yielded decreased, negative results. So, the selection is not preferred in general in this study.

TABLE IV
FEATURE TYPE PREFERENCE

No	Feature Type
1	Presence of certain sentimental polarity words
2	Presence of certain sentimental polarity emoticons
3	Presence of certain grammatical structures Especially negations
4	Unigram features

C. Machine Learning Approach

Table-V contains ML method preference. The methods are selected based on the prior art and on ease of achieving experimental results.

Especially, ensemble learning methods such as Random Forest and Multi Class Classifier are used for seeing effect of each ensembled classifers learning different classes. The main idea behind ensemble learning is that one algorithm could learn better one class, the other could learn better the other class, so ensembling each algorithm with proper theoretical base could help to increase class-based accuracy ratios and conclusively overall ratio.

Artificial Neural Network (or Multilayer Perceptron) took too much time to train in case of only unigram features (for 5560 features). Therefore, even if it's a valid system in the literature, it wasn't used.

TABLE V
MACHINE LEARNING METHODS

No	Machine Learning Method
1	Support Vector Machine (SVM)
2	Naive Bayes Classification
3	J48 Decision Tree
4	Random Forest
5	MultiClassClassifier (also, its updatable version)
6	IterativeClassifierOptimizer

First following experimental result part (Section 3.4-6) would contain only results in absence of unigram features, in other words just features 1, 2, 3 (Table-4) have been used. For each dataset, there would be two types of results with/out oversampling due to imbalanced class size for each dataset (Table-1, 2).

IV. RESULTS

First, experimental results without unigram features use is presented with/out oversampling separately for each dataset (Section-IV-A,IV-B,IV-C). Second, experimental results with unigram features on dataset-1 is presented by describing effect of the feature set on accuracy.

A. Dataset-1 Results

In this dataset, all irrelevant labelled tweets are removed from the dataset. In one previous work, examination of accepting irrelevant tweets as neutral tweets was done. It increased accuracy ratios up 90%. However, it's not preferred. Since, there could be bias in terms of neutrality, a possibility of accepting wrong classification as right.

As seen from Table-VI, there is a lack of learning and prediction of negative and positive classes. It comes from fact that there is a dominance of neutral tweets in dataset-1 (Table-II). The fact is reflected on TP and FP rates (Table-VI). In each method, the best TP rate belongs to neutral class, also, FP rate of neutral class dominates all other classes for each method. It means mostly neutral class is wrongly estimated in place of either negative class or positive class.

Overlearning of neutral class is eliminated to some extent by oversampling as its seen from Table-VII, it's in accordance with oversampling intuition. Nevertheless, we lose the generality and variety in oversampling due to duplication of given data, even if it's made in randomized manner by shuffling.

Best method in terms of overall accuracy ratio is MultiClassClassifier (0.711). Close overall accuracy ratios comes from Random Forest(0.707), SVM (0.706).

As seen from Table-VII, oversampling has positive effect on class-based accuracy ratios. But, it decreased overall accuracy, didn't increase sufficiently class-based accuracy ratios.

TABLE VI
EXPERIMENTAL RESULTS OF DATASET-1
IN ABSENCE OF UNIGRAM FEATURES AND OVERSAMPLING

	Positive	Negative	Neutral
SVM			
TP Rate	0.17	0.15	0.94
FP Rate	0.05	0.19	0.80
Naive Bayes Classifier			
TP Rate	0.12	0.12	0.93
FP Rate	0.05	0.03	0.85
J48 Decision Tree			
TP Rate	0.10	0.00	0.96
FP Rate	0.04	0.00	0.93
Random Forest			
TP Rate	0.17	0.23	0.92
FP Rate	0.05	0.03	0.76
MultiClass Classifier			
TP Rate	0.16	0.21	0.93
FP Rate	0.04	0.02	0.78

B. Dataset-2 Results

Best-performing method in terms of overall accuracy is again MultiClass Classifier with 0.681. It is closely followed by Iterative Classifier Optimizer (0.677), Random Forest (0.674) and J48 Decision Tree(0.672), SVM. MultiClass

TABLE VII
EXPERIMENTAL RESULTS OF DATASET-1
IN ABSENCE OF UNIGRAM FEATURES AND IN CASE OF OVERSAMPLING

	Positive	Negative	Neutral
SVM			
TP Rate	0.47	0.32	0.76
FP Rate	0.17	0.06	0.50
Naive Bayes Classifier			
TP Rate	0.43	0.25	0.78
FP Rate	0.18	0.04	0.55
J48 Decision Tree			
TP Rate	0.48	0.20	0.80
FP Rate	0.17	0.03	0.57
Random Forest			
TP Rate	0.51	0.34	0.76
FP Rate	0.16	0.05	0.49
MultiClass Classifier			
TP Rate	0.48	0.35	0.75
FP Rate	0.16	0.07	0.48
Iterative Classifier Optimizer			
TP Rate	0.45	0.27	0.78
FP Rate	0.17	0.05	0.74

Classifier and Random Forest are two top methods in accordance to the experiments without oversampling. It shows us value of ensemble learning to certain extent.

Negative class dominance is observed. It has approximately same effect as neutral class has in dataset-1 on TP and FP rates. Oversampling effect is naturally examined due to imbalanced class frequency distributions (Table-II). Class-based accuracy ratios are ameliorated thanks to oversampling in general. However, overall accuracy ratio has dropped in some extent similar to dataset-1 case.

C. Dataset-3 Results

Worst overall accuracy ratios came from dataset-3. Best accuracy ratio achieved is 0.612 and it is achieved by J48 Decision Tree method. Accuracy ratios of all others are above 60% and very close to J48 Decision Tree.

Interesting, contradictory observation is that negative class has dominance over other classes even if positive class has more tweets in this dataset (Table-II). Its main intuitive interpretation is oversampling should affect class-based accuracies, since there is a clear difference between class-based accuracies (Table-VIII, Table-IX) and distributions (Table-II, Table-III). The intuition is in accordance with results to certain extent (Table-VIII, Table-IX).

However, neutral class accuracy and overall accuracy is very low with respect to previous results in case of oversampling (Table-VII,IX,XI). This dataset is the worst-case of the study from certain perspectives.

TABLE VIII
EXPERIMENTAL RESULTS OF DATASET-2
IN ABSENCE OF UNIGRAM FEATURES AND OVERSAMPLING

	Positive	Negative	Neutral
SVM			
TP Rate	0.41	0.75	0.00
FP Rate	0.06	0.77	0.00
Naive Bayes Classifier			
TP Rate	0.38	0.93	0.47
FP Rate	0.05	0.77	0.02
J48 Decision Tree			
TP Rate	0.38	0.97	0.01
FP Rate	0.04	0.79	0.00
Random Forest			
TP Rate	0.36	0.96	0.02
FP Rate	0.04	0.79	0.01
MultiClass Classifier			
TP Rate	0.38	0.97	0.01
FP Rate	0.04	0.79	0.00
Iterative Classifier Optimizer			
TP Rate	0.35	0.97	0.01
FP Rate	0.03	0.81	0.01

TABLE IX
EXPERIMENTAL RESULTS OF DATASET-2
IN ABSENCE OF UNIGRAM FEATURES AND IN CASE OF OVERSAMPLING

	Positive	Negative	Neutral
SVM			
TP Rate	0.68	0.84	0.05
FP Rate	0.14	0.50	0.02
Naive Bayes Classification			
TP Rate	0.67	0.84	0.03
FP Rate	0.14	0.51	0.02
J48 Decision Tree			
TP Rate	0.68	0.83	0.05
FP Rate	0.14	0.51	0.02
Random Forest			
TP Rate	0.68	0.86	0.08
FP Rate	0.12	0.49	0.02
MultiClass Classifier			
TP Rate	0.70	0.86	0.05
FP Rate	0.14	0.49	0.01
Iterative Classifier Optimizer			
TP Rate	0.69	0.82	0.02
FP Rate	0.15	0.51	0.01

TABLE X
EXPERIMENTAL RESULTS OF DATASET-3
IN ABSENCE OF UNIGRAM FEATURES AND OVERSAMPLING

SVM			
	Positive	Negative	Neutral
TP Rate	0.00	0.99	0.00
FP Rate	0.00	0.99	0.05
Naive Bayes Classification			
TP Rate	0.10	0.96	0.00
FP Rate	0.04	0.93	0.00
J48 Decision Tree			
TP Rate	0.02	0.99	0.00
FP Rate	0.00	0.99	0.00
Random Forest			
TP Rate	0.03	0.99	0.01
FP Rate	0.01	0.97	0.01
MultiClass Classifier			
TP Rate	0.04	0.99	0.01
FP Rate	0.01	0.97	0.01
Iterative Classifier Optimizer			
TP Rate	0.01	0.99	0.00
FP Rate	0.00	0.99	0.00

TABLE XI
EXPERIMENTAL RESULTS OF DATASET-3
IN ABSENCE OF UNIGRAM FEATURES AND IN CASE OF OVERSAMPLING

SVM			
	Positive	Negative	Neutral
TP Rate	0.30	0.84	0.02
FP Rate	0.16	0.76	0.01
Naive Bayes Classification			
TP Rate	0.26	0.87	0.00
FP Rate	0.14	0.80	0.00
J48 Decision Tree			
TP Rate	0.28	0.85	0.02
FP Rate	0.14	0.78	0.01
Random Forest			
TP Rate	0.30	0.85	0.03
FP Rate	0.14	0.77	0.02
MultiClass Classifier			
TP Rate	0.01	0.99	0.00
FP Rate	0.03	0.94	0.00
Iterative Classifier Optimizer			
TP Rate	0.28	0.87	0.00
FP Rate	0.14	0.79	0.00

D. Discussion: Effect of Unigram Features

Effect of unigram features is examined on only dataset-1. We removed stopwords during extraction of unigram features based on term frequencies. Since, other datasets are of big sizes, there has been computational complexity problems. For instance, number of unigram features extracted from a single tweet is 15638 in dataset-2. There are 14327 tweets without oversampling in dataset-2. Therefore, it took too much time and didnt finish in a reasonable amount of time. It shows us need of parallelism in order to have a reasonable speed-up and more extensive experiment.

Dataset-1 is tested twofold, i.e., tested with only unigram features and tested with unigram features plus other features (Table-IV).

General observation is existence of unigram features balanced class-based accuracies dramatically and increased overall accuracy visibly.

When only unigram features are used, overall accuracy ratio of 0.876 is reached with Random Forest in case of oversampling (Positive Accuracy: 0.930 Negative Accuracy : 0.949 Neutral Accuracy: 0.749), SVM was also remarkable with overall accuracy ratio of 0.863 and balanced class-based accuracy ratios.

Its interesting that both methods reached little bit better accuracy ratio in presence of previously used features with unigram features (Table-IV). Overall accuracy ratio of 0.883 is reached with Random Forest in case of oversampling (Positive Accuracy: 0.948 Negative Accuracy :0.933 Neutral Accuracy : 0.765), SVM was again remarkable with overall accuracy ratio of 0.867 and balanced class-based accuracy ratios.

Interpretation that unigram features plus already used features (Table-4) and oversampling would yield better results, that at least just unigram features use with oversampling would have positive effect is natural projection.

Baselines, whose their accuracy ratios are are changing between 70%-80%, are reached without any unigram features in dataset-1 (71%, Section-IV-A), dataset-2 (68%, Section-IV-B) in overall accuracy. Dataset-1 accuracy ratios with unigram features use and oversampling outperformed the baselines.

V. CONCLUSION

Numerous institutions, organizations, society try to capture whats going on others minds with certain objectives. In this regard, data scientists especially try to make the sentimental polarity detection on maybe the most popular micro-blogging site Twitter.

In this study, twitter sentiment analysis is done on three annotated datasets collected from twitter and annotated manually.

Mainly, Machine Learning systems from the prior art (e.g. Support Vector Machine) and Information retrieval techniques (e.g. Stopword removal based on term frequency).

Best overall accuracy ratio (88%) came from dataset-1 with use of all features listed in Table-IV, of oversampling and finally of an ensemble learning method (i.e. Random Forest). This approach yielded to balanced class-based accuracies. We outperformed the baselines(70%-80%) seen on the literature.

As a future work, experiments on unigram features effect (Section-IV-D) should to be extended dataset-2, dataset-3. Finally, very large experiment could be done by solving computational complexity problems to observe behaviour of the analysis on large scale. Baseline is caught in dataset-2 experiments by 68%. However, this is not the case in dataset-3 experiments. We are 10% percent back baseline work accuracy ratio range (i.e. 70% - 80%). Especially, the dataset-3 would be good to be trained with unigram features and with/out already used features.

It is observed that critical stage of twitter sentiment analysis is feature extraction, analysis. Our experiments with unigram features yielded dramatic increase in overall accuracy (appx. 18%) and also dramatic balanced class-based accuracy results. In this respect, SVM exhibited remarkable performance by using only unigram features (86.3%) and by using all features listed in Table-IV (86.7%) in accordance with prior work. Besides, ML methods and their general effectivenesses are highly important. In other words, not all ML methods demonstrated the dramatic increase. Only some of them have given increased results. Nevertheless, there were a balanced class-based accuracy ratios in presence of unigram features for most of them.

Oversampling showed great benefit in terms of solving unbalanced class frequency distributions effect on class-based accuracy ratios and of solving unbalanced class-based accuracy ratios. However, it decreased overall accuracy while balancing the class-based ratios without unigram features case. At the same time, it increased dramatically overall accuracy ratio while highly balancing the unbalanced class-based ratios in presence of unigram features.

Unigram features seems to have significant importance in the twitter sentiment analysis. There are similar reportings regarding its positive effect on twitter sentiment analysis prediction ratios in the literature.

References

[1] Benevenuto, F., Magno, G., Rodrigues, TBenevenuto, Fabricio, et al. ”Detecting spammers on twitter.” Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). Vol. 6. 2010.

[2] Rajdev, Meet, and Kyumin Lee. ”Fake and Spam Messages: Detecting Misinformation during Natural Disasters on Social Media.” 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Vol. 1. IEEE, 2015.

[3] Mohanraj, V. ”A Survey on Spam Detection in Twitter.”International Journal of Computer Science and Business Informatics 14.1 (2014).

[4] Pang, Bo, and Lillian Lee. ”Opinion mining and sentiment analysis.” Foundations and trends in information retrieval 2.1-2 (2008): 1-135.

[5] http://www.sananalytics.com/lab/twitter-sentiment/

[6] https://www.kaggle.com/crowdflower/twitter-airline-sentiment

[7] https://www.kaggle.com/crowdflower/first-gop-debate-twitter-sentiment

[8] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. ”Twitter sentiment analysis: The good the bad and the omg!.” Icwsm 11 (2011): 538-541.

[9] Agarwal, Apoorv, et al. ”Sentiment analysis of twitter data.” Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011.

[10] http://hughchristensen.co.uk/papers/socialNetworking/Twitter%20Sentiment%20Analysis.pdf

[11] http://www.mit.edu/%7E6.863/spring2009/readings/ngrampages.pdf

[12] https://en.wikipedia.org/wiki/N-gram

[13] https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf

[14] http://www.cs.columbia.edu/%7Ekathy/NLP/ClassSlides/Class3-ngrams09/ngrams.pdf

[15] http://www.statmt.org/book/slides/07-language-models.pdf

[16] Go, Alec, Richa Bhayani, and Lei Huang. ”Twitter sentiment classification using distant supervision.” CS224N Project Report, Stanford 1 (2009):12.