Examining the Automated Inference of Tweet Topics

Saed Rezayi
Department of Computer Science
University of Oregon
saed@cs.uoregon.edu

Abstract

The increasing volume of information exchange over online social networks (e.g. Twitter, Facebook) has led to the growing interest in technique for automated inference of the topic of individual posts/tweets in recent years. Short length, lack of a well defined set of topics, and use of acronyms in tweets are some of the reasons that make topic inference of tweets challenging.

In this study, we examine the feasibility and accuracy of using supervised learning techniques for inferring tweet topics. To efficiently produce a training dataset for a classifier, we explore whether the category of a professional Twitter account can offer a reliable label/topic for generated tweets by that account, e.g. whether the Twitter account of a professional soccer team most generates tweets related to the topic of soccer. We examine this hypothesis by focusing on generated tweets by more than 170 sample Twitter accounts related to 16 specific categories. First, to investigate the clarity of perceived topics for tweets by humans, we recruit human subjects to label tweets of sample accounts. Using these labeled tweets, we study the fraction of tweets for each account whose labels are aligned (and misaligned) with the category of their accounts. We show that these basic characteristics of tweets per account can be viewed as a set of "topic alignment features" that can often specify the category of an account in an automated fashion. Indeed, these features illustrate how the corresponding account owners use Twitter and also reveal the pairwise relationship between some of the selected topics.

We also evaluate the accuracy of classification techniques in three cases with a different level of reliability for training and testing datasets. Our results show how the selection of training sets affects the accuracy of classifications. We also demonstrate that the accuracy of the classification for each account is correlated with its topic alignment features. This suggests that the features can be used to identify accounts whose tweets are more appropriate for training. Finally, we illustrate that the primary selected keywords by classifiers properly represent each topic.

1 Introduction

Growing levels of interactions between individuals and organizations through online social networks such as Twitter or Facebook has turned them into online information societies where users generate, propagate, exchange, receive information and act on it. Thus, there is a growing interest in mining this information for various purposes such as marketing, health, security, economics, etc. [15], [5], and [20].

Extracting information from this online source is challenging because length of a post is often short (for tweets it is 140 characters), and a post could be inherently ambiguous. Besides, use of unconventional language and unclear words and abbreviations adds to the complexity of analysis. One basic issue for information mining is to provide some basic context for a post, such as its topic. More specifically, given a post, can we infer whether it is about soccer, politics, etc. However, There is no widely accepted set of a topics with a clear granularity (e.g. what is a proper granularity for a topic, should we consider sport or soccer as a topic). This issue and the fact that posts could be too simple (no topic) or too complicated (multiple topics) makes the problem more challenging.

Machine learning techniques are promising approaches for such inferences. Prior studies have used Topic Modeling to find a topic of a document. However these algorithms are highly dependent on the number of topics. It might be impossible to figure out the right number of latent topics in LDA Algorithm [4] and such number may not even exist. We address this issue in Section 7. As a result, our goal is to infer a topic of a post using supervised classification.

However, before pursuing our goal we would like to investigate topics of tweets as they are perceived by humans. To make this manageable consider a case with

N specific topics of interest. Toward this end we use categories used by an online marketing website namely socialbakers.com and we collect tweets of well known accounts per category. To tackle the challenge in supervised learning we label the tweets by humans and our hypothesis is that "professional accounts generate tweets related to their category." For example consider the following tweet: "LIVE: President Obama is speaking at the White House" put out by the account Barack Obama. We can intuitively say that Barack Obama falls into the politics category and also its tweet has the topic of politics. That is why we first study whether individual tweets have clear and unique topics as they are perceived by humans rather than simply using a supervised LM technique.

We would like to gain insight about following fundamental questions:

- *How are topics of tweets perceived by humans? Do tweets have one or multiple or no clear topics?* The answer to these questions is important because a tweet is our only source of information that we use to train our model and if we do not train the system precisely how could we except that machine assigns a topic to a short text that has no information in it, "Enjoy the sunshine" for instance!

- *To what extent is topic of a tweet aligned with the category of the account that generated the tweet?* The answer to this question could vary across different categories and even among accounts in a single category. In fact, the alignment of tweet topics with category of an account shows how that entity associated with the account is using Twitter, e.g. announcement, advertisement, voting media, etc.

- *How do professional Twitter accounts use Twitter?* As a result of the above question we are also interested in answering this question.

The rest of the paper is organized as follows: Section 2 reviews the related works in this area. Section 3 presents data collection and data labeling and a summary of our dataset. Sections 4 characterizes the dataset and investigates the alignment of account category and tweet topic. Also we present our feature set that is used for rule based classification in this section. Section 5 then leverages classification technique to infer a topic from a tweet. Section 6 investigates if there are certain keywords that are related to different categories. Finally, Section 9 presents our conclusions.

2 Related Work

Assigning a topic to a document is not a new problem and there have been many efforts in analyzing text. In general, there are two approaches for natural text processing: unsupervised and supervised analysis. Unsupervised analysis is generally called clustering that divides a set of objects into clusters so that objects in the same cluster are similar to each other. These algorithms, e.g. K-means [8], are unsupervised, meaning no human input is necessary. Topic inference has plenty of application from recommender systems[21] to ad placement [1] and interest mining[7].

All studies in this domain are categorized under Machine Learning (ML) techniques. To analyze text and retrieve information from it, classification have been widely used and studied where a model is trained by a set of prelabeled documents (training set) and is asked to classify a new set of unseen documents (test set). [13], [11], and [22], have leveraged popular classifiers on text.

There are other studies that use classification to infer other properties of tweets like sentiment analysis in [6] and [14] or measuring question quality in [23] or link prediction [2]; however the limited information in Twitter text (each tweet is limited to 140 characters) has caused difficulties in the task of topic inference.

There is another emerging technique called topic modeling that can be supervised [3] or unsupervised [17]. These algorithms discover semantic structure of documents, by examining word statistical co-occurrence patterns within a corpus of training documents. Authors in [10] address the problem of using standard topic models in micro-blogging environments (such as Twitter) by studying how the models can be trained on the dataset. L-LDA (Labelled LDA) that is proposed in [18] is based on LDA [4] and is a supervised topic model for assigning topics to a collection of documents.

3 Data Collection & Data Labeling

This section describes our dataset and the way we label tweets. All general statistics are provided here including number of categories, number of accounts per category and number of labeled and unlabeled tweets per account.

3.1 Tweet Collection

To build an effective training set, we select a group of Twitter accounts that are related to a specific *category* [1] and collect all available tweets from these accounts. This approach to data collection not only increases the likelihood of collecting tweets that are related to the selected categories but also enables us to examine to what extent the topic of generated tweets by individual accounts are related to the category of the account. Toward this end, we use web sites, namely socialbakers.com, that publish list of popular Twitter accounts (including their Twitter IDs and the number of followers) that are classified into more than 80 categories. We identify 16 categories and hand pick a set of accounts that represent well known entities (*i.e.*, major teams, companies, brands with a large number of followers) for that category.

While focusing on well-recognized accounts may limit the number of selected accounts in some categories, it intuitively increases the likelihood that their tweets are related to their category as their accounts are likely to be professionally managed. The selected categories essentially define the scope of our study. The list of selected categories along with the number of related accounts and collected tweets in each category is summarized in Table 1. The complete list of all selected accounts for each category and their associated tweets is available in the Appendix 2.

While our goal is to ensure that selected categories are clearly separated, achieving this goal is not trivial. Intuitively, there is some overlap between pairs of selected accounts (*e.g.*, fashion and beauty, or beverage and alcohol), and a category such as news has inherent overlap with a few other categories (politics, finance, or sport). Considering these overlapping categories enables us to explore the potential effect of category overlap on our analysis.

3.2 Tweet Labeling

We recruited a group of UO students to specify the topic (*i.e.*, label) of a subset of tweets in our dataset. Toward this end, each student is provided with a spreadsheet that includes the text of a random selection of tweets and prompts them to assign a topic to each tweet from a drop-down menu. This menu of topics contains all sixteen categories along with two more sensible categories: "no topic" and "other". Students are instructed to assign the label "other" to a tweet if it has a pronounced topic that is not listed in the menu (*e.g.*, music), and assign the label "no topic" if they can not associate any clear topic to a tweet (*e.g.*, "2010 has been an exception year").

The assigned tweets to students are organized into two mutually exclusive groups:

- Three label tweets: Tweets that were labeled by three different students
- Single label tweets: Tweets that were labeled only once.

The multi-label tweets enable us to examine the consistency of label assignment by individuals. Such an inconsistency could be due to genuine disagreement among students on the topic of the tweet or caused by mistakes. The last two columns of Table 1 specifies the fraction of tweets (for each category) that has been labeled once or three times. As this table shows, the recruited students have assigned more than 121.6K labels (including 3 separate labels for 10K tweets).

For each tweet with three labels, we define the notion of Level of Agreement (LoA) that shows the maximum number of similar labels. More specifically, we use the term LoA3, LoA2 and LoA1 for a tweet with three labels to indicate that its number of similar labels are 3, 2, or 1, respectively. We also use the notation of LoA2+ to refer to the collection of tweets that have LoA2 or LoA3 (*i.e.*, LoA2+ = LoA2 $\cup$ LoA3).

4 Characterizing Assigned Topics by Human Labels

We leverage the tweets with three labels to examine the characteristics of assigned topics to tweets by human. These characteristics provide the basic understanding of the clarity of topic for individual tweets and the alignment between the topic of tweets and the category of their associated account. The obtained insights from these characterization effort will inform the evaluation of classification techniques in the second half of the paper.

The task of assigning a label to a tweet may not be trivial when the associated keywords offer diverse clues. For example, a tweet with keywords "Clare Choir, tour, Australia" provides clue about traveling, music and singing, as well as education (since Clare is a college at Oxford University). However, a person who does not know about

[1] Throughout this paper, we use the term "category" to refer to the context of individual Twitter account, and use the term "topic" to indicate the context of individual "tweets". Using different terms should further clarify the focus of each discussion.

topic	No of accounts	No of tweets	No of tweets with one label	No of tweets with three labels
airline	10	32,229	5,393 (%16.7)	600 (%1.8)
alcohol	10	28,339	5,398 (%19.0)	599 (%2.1)
auto	12	38,589	6,472 (%16.7)	720 (%1.8)
basket	9	28,850	4,848 (%16.8)	540 (%1.8)
beauty	10	32,211	5,362 (%17.6)	596 (%1.8)
beverage	10	32,969	5,362 (%16.2)	599 (%1.8)
education	11	33,773	5,923 (%17.5)	655 (%1.9)
electronics	12	37,522	6,494 (%17.3)	720 (%1.9)
fashion	14	34,837	7,109 (%20.0)	702 (%2.0)
finance	11	31,776	5,391 (%16.9)	598 (%1.8)
gaming	6	19,383	3,209 (%16.5)	357 (%1.8)
health	10	27,726	5,395 (%19.4)	599 (%2.1)
news	14	45,044	7,575 (%16.8)	840 (%1.8)
politics	15	36,923	7,722 (%20.9)	781 (%2.1)
soccer	12	38,522	6,175 (%16.0)	677 (%1.7)
telecom	7	22,583	3,775 (%16.7)	420 (%1.8)
total	173	521,276	91,603 (%17.5)	10,003 (%1.9)

Table 1: List of selected topics and fraction of single/multiple-label tweets

the educational context, will not assign the label of education to this tweet. In essence, the available information and context to individuals could affect the way they perceive and thus label tweets with diverse clues.

Despite this challenge, having three labels for each tweet enables us to determine the topic of a tweet with relatively high confidence. In particular, we assume that if at least two assigned labels for a tweet are similar (*i.e.*, any LoA2+ tweet), the common label determines the topic of the tweet since it is unlikely that two individuals make a similar mistake in assigning a label. Note that the common label of a tweet might be *aligned* or *misaligned* with the category of the corresponding account. For example a tweet that has these keywords "Reuters, US Econ, collapse, benefits, $29B, GM" which are associated with a Twitter account with the category of auto and has three similar labels of finance is a LoA3/misaligned.

Hence for each tweet we measure LoAi/x metric where i shows the level of agreement between labels ($i \in \{1, 2, 3\}$) and x indicates the alignment ($x \in \{aligned, misaligned\}$)

We have manually inspected hundreds of LoA2+ tweets to verify the use of common labels as the topic of tweets for LoA2+ tweets that are both aligned and misaligned with their corresponding accounts' category. We observed that for an absolute majority of LoA2+ tweets ($> 95\%$) the common label is the most reasonable topic. The most common exceptions are tweets whose common misaligned label is "no topic" or "other" due to the lack of a dominant context for the tweet. For example, a tweet with keywords "disaster, texting, Redcross" is associated with an account of health category but was labeled twice as "other". Our inspections confirm that the common label for LoA2+ tweets can reliably be used as the topic of the tweet despite stated challenge for human to assign a consistent topic to tweets with conflicting clues. In the rest of this section, we characterize the topic of LoA2+ tweets in order to answer the following key questions:

- Does (and to what extent) the topic of the generated tweets by (professional) Twitter accounts is aligned with their category across different categories?
- Does the level of alignment between the category of a Twitter account and the topic of its tweets vary across different categories?
- What does the alignment between the category of an account and its tweets reveal?

4.1 Alignment of Account Category and its Tweet Topic

To explore the relation between the category of an account and the topic of its tweets, we divide all tweets of each selected account into the following three groups:

- LoA2+/aligned
- LoA2+/misaligned

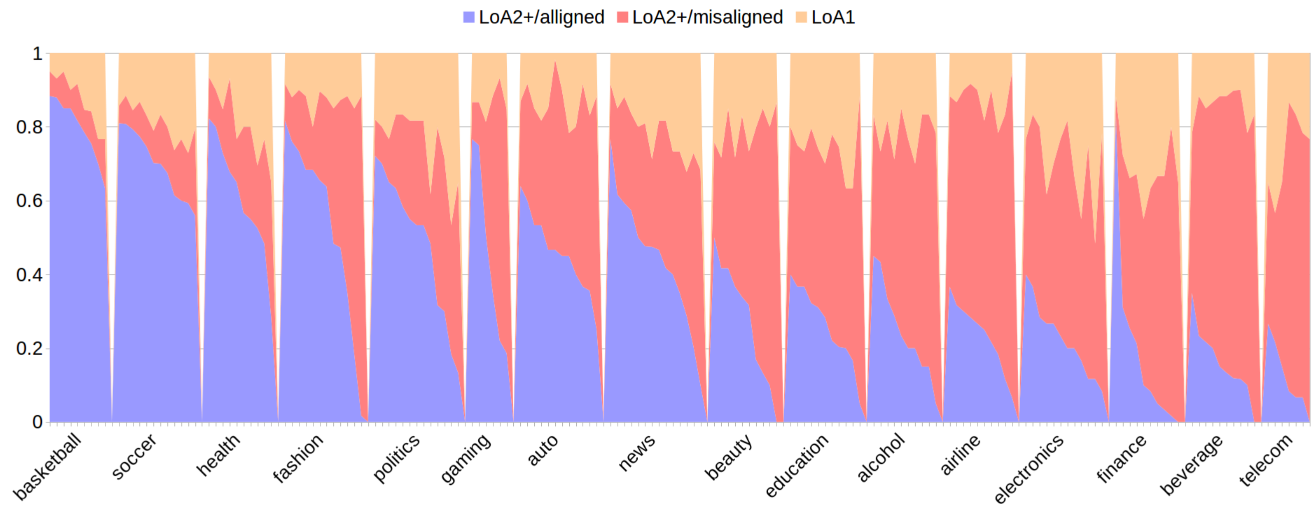


Figure 1: Agreement between tweet labels and account category for three label tweets per account

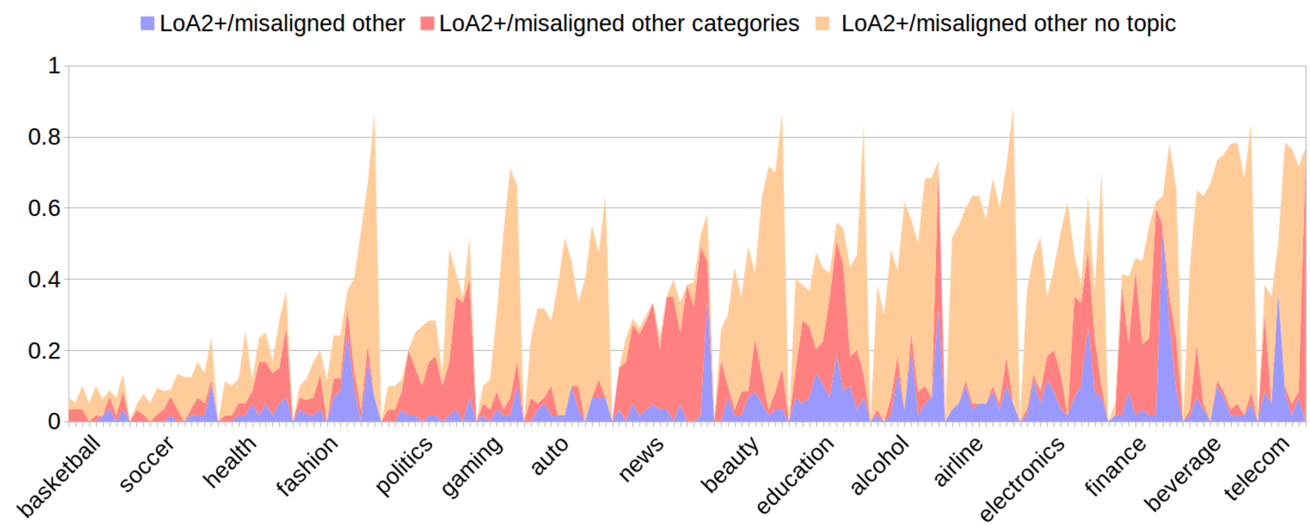


Figure 2: Breakdown of LoA2+/misaligned tweets among "other", "no topic", and "other categories" per account

- LoA1

We refer to these three groups as aligned, misaligned and ambiguous tweets, respectively. Intuitively, these three groups of tweets respectively indicate the extent that generated tweets by an account is related or unrelated to its category or is ambiguous. In essence, the specific division of tweets across these three groups can provide a valuable insight on how these Twitter accounts are used by their owners.

Figure 1 presents the percentage of tweets across these three groups for each account. Furthermore, accounts within the same category are bundled together, categories are ordered (from left to right) based on their average percentage of LoA2+/aligned and within each category accounts are ordered (from left to right) based on their percentage of LoA2+/aligned. This figure illustrates following interesting points:

First, there are some variations in the division of tweets among aligned, misaligned and ambiguous groups within each category. We observe that in some categories (soccer, basketball, health, politics) most accounts clearly exhibit a much larger percentage of aligned tweets than other categories. We refer to these categories as *purposeful* as a significant fraction of their tweets are related to their mission. In contrast, in some other categories (telecom, beverage, finance, electronics, airlines, alcohol, education) a significant percentage of published tweets are misaligned. We refer to these categories as *aimless*. In essence, the relative percentage of aligned and misaligned tweets appears to be largely related to the category of the accounts. Second, the percentage of ambiguous tweets is

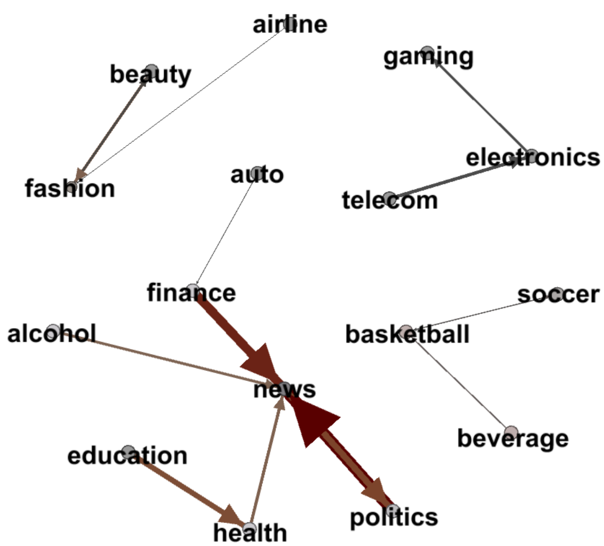


Figure 3: Other major related categories for multi purpose accounts.

around 10% to 30% in most cases and is relatively stable across different categories.

4.1.1 Misaligned Tweets

To gain more insight into the LoA2+/misaligned tweets, we take a closer look at this group by dividing them into the following three subgroups based on their inferred topic (that is misaligned with its category):

- *Other*: tweets whose label is "other"
- *No Topic*: tweets whose label is "no topic"
- *Other Topics*: tweets whose label is the same as one of the other 15 categories.

Note that the characterization of these misaligned tweets are more relevant to aimless categories as most of their tweets are misaligned.

Figure 2 plots the percentage of all LoA2+/misaligned tweets among the above three types for each account, *i.e.*, essentially providing the breakdown of the LoA2+/misaligned in Figure1. This figure clearly illustrates that a significant fraction of misaligned tweets in some "aimless" categories, namely telecom, beverage, airline, alcohol, beauty, auto and gaming, have no topic at all. This reconfirms our earlier assertion that these categories generally appear to be aimless.

In contrast, a majority of misaligned tweets in some other categories, namely finance, education, news, politics, and health are mapped to one of our other categories. We refer to these categories as *multi purpose* categories. In Figure 3 we try to visualize this metric as a graph. In this graph nodes are categories and edges are number of mislabeled tweets between to categories. As can be seen, edges are weighted and directed. Weight represents the number of mislabeled categories and is proportional to thickness. Direction shows in which way we have mislabeling. For example a large number of finance tweets are labeled as news but for news politics is the second major category. Accordingly we draw a conclusion that the edges between two categories shows the overlap between those two categories. This figure also clearly illustrates that news is a multi purpose category and it mainly has overlap with politics and finance. Another pair category is basketball and soccer because they fall into super category of sport. For some sample tweets that shows the multi purpose nature of tweets see Table 2.

4.1.2 Ambiguous Tweets

We now turn our attention to the LoA1 subset of tweets that have very diverse labels. To learn more about these tweets, we divide them into two more groups:

- *LoA1/aligned*: the tweets for which one of their labels is aligned with their category.
- *LoA1/misaligned*: the tweets that none of their labels is aligned with their category.

Figure 4 depicts the break down of the total percentage of LoA1 tweets for each account into LoA1/aligned and misaligned.

We can clearly observe that for many categories, an absolute majority of LoA1 tweets are LoA1/aligned with their category. This implies that tweet's context has some connection with its category but it may not very obvious/strong. Our closer inspection of these tweets revealed that most of these tweets can indeed be reasonably associated with two different topics, the third label is in some cases a very reasonable one and in other cases appear to be a mistake. To demonstrate this point consider the following LoA1/aligned tweets: *"Tories, Labour and Lib Dems to declare opposition to a currency union with Scotland"* with the account category of news that received three reasonable labels of news, politics and finance, or *"Download the new Fox News app for Android. Watch Fox News Channel live"* that has the category of electronics and was properly labeled as telecom, news, and electronics. However, this tweet *"Monica Lewinsky speaks out, says she was made scapegoat"* received two appropriate labels of politics, news and one seemingly in appropriate label of fashion while its category is news.

tweet	label1	label2	label3	category
Pro-Obama nonprofit will no longer divert gifts to allied groups	politics	politics	news	news
Wall Street is sharply divided on 2015 outlook [CNBC Fed Survey]	finance	finance	news	news
Follow the fragrance trail of Jadore from Grasse	beauty	beauty	beauty	fashion
@PlayStation: 12GB PS3 system will be $199 in North America.	gaming	gaming	gaming	electronics
Spurs Connect: Free App for Spurs fans Now on Android	soccer	basketball	basketball	soccer

Table 2: Sample tweets for LoA2+/misaligned with other categories that shows multi purpose nature of some categories.

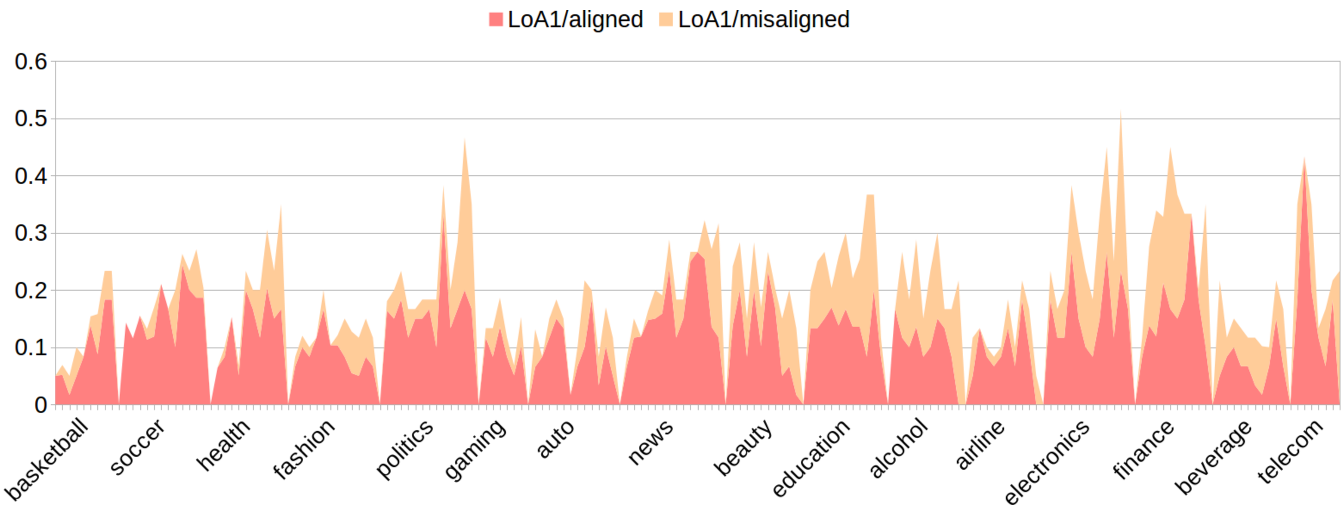


Figure 4: Breakdown of LoA1 tweets for each account into aligned and misaligned

4.2 Automated Classification of Accounts

So far we have broadly classified Twitter accounts based on their LoAi/x characteristics in a hand crafted manner. Each account has a few LoAi/x numbers that can be viewed as its *features*. We can use a classifier to identify the rules for accounts in each category. Obviously, the rules may not be perfect and some accounts are grouped with other categories. We use decision tree classifier to generate these rules and examine whether they are aligned with our earlier hand crafted classifications. This exercise also shows the relative distance between categories.

The list of features that are fed into decision tree classifier are as follows:

feature name	abbreviation
LoA2+/aligned	LoA2+/a
LoA2+/misaligned with other	LoA2+/mo
LoA+/misaligned with no topic	LoA2+/mnt
LoA2+/misaligned with other topics	loA2+/mot
LoA1/aligned	LoA1/a
LoA1/misaligned	LoA1/m

Based on the generated tree, LoA2+/a has the highest information gain and becomes the root for the tree and it splits all accounts into two imbalanced subgroups. The tree is generated graphically and is available in Appendix 1. Here we list some sample rules that show these features lead us to the correct point. Also Figure 5 is a part of this tree that reveals the following rules.

(LoA2+/a $>$ 45.8%) $\wedge$ (LoA2+/mot $>$ 9.16%) $\wedge$ (LoA2+/mo $>$ 1.68%) $\Rightarrow$ 60% politics

(LoA2+/a $>$ 45.8%) $\wedge$ (LoA2+/mot $>$ 9.16%) $\wedge$ (LoA2+/mo $<=$ 1.68%) $\Rightarrow$ 60% news

These rules confirm our previous observation in Figures 1 and 2. For example in Figure 2, we observed that LoA2+/misaligned with "other" categories has a great share of all LoA2+/misaligned tweets for news and politics, and classification place them in a same branch.

In another branch we see that finance and news has the same number of accounts in one leaf. In other words we can extract following rule:

(LoA2+/a $<=$ 45.8%) $\wedge$ (LoA2+/mnt $<=$ 34.1%) $\wedge$ (LoA2+/mot $>$ 6.7%) $\wedge$ (LoA2+/mo $<=$ 5.8%) $\Rightarrow$ 30% news and 30% finance which is consistent with Figure 3 that shows news and finance have the closest distance after news and politics.

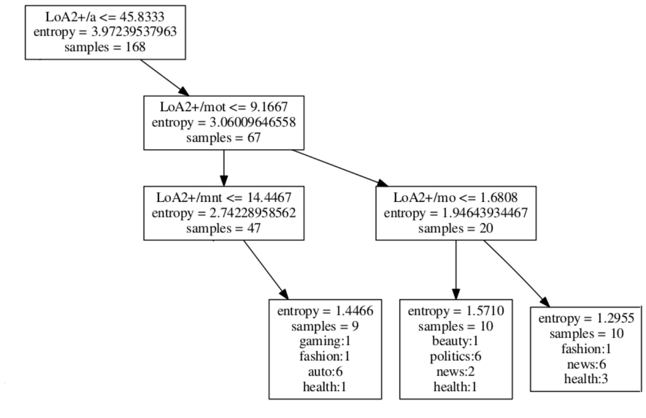


Figure 5: Partial decision tree for politics and news

4.3 Inferring Used Strategy by Accounts/Categories

As a result of above exercise we can elaborate on how certain accounts use twitter, (*e.g.*, informing followers about deals, providing info, asking them to vote) and how this type of use is aligned with classification result (in Section 5), and whether the accounts are managed professionally or casually.

According to the decision tree model, we see none of the leaves is clearly associated with category telecom as telecom accounts are scattered in four different leaves. This suggests that telecom accounts do not use Twitter for telecommunication reasons. We can verify this claim by manually checking the tweets of these accounts.

For example 65% of tweets of account Sprint is the following text!

Please visit *some url* to complete your contest entry!

where *some url* is a url that will be redirected to the sprint website when it is clicked.

Another telecom account Skype uses Twitter very casually and mostly to thank their costumers and ask about their feedbacks. We list some of its tweets in table **??**.

As it is seen nothing informative could be found in these tweets and we can not expect that machine or human could infer an appropriate topic for this account. Such accounts can be found in other categories as well. Redbull is an example of beverage category that uses Twitter

Awesome! We're glad we can be there for you. :)
Wow, you must really love the emotions. Who do we help you stay in touch with? :)
glad we could bring a few extra laughs to your day. Do you and your brother catch up often?
We are here to help. :)
Sounds like someone was a little bit tired ;)
We're glad we can be a part of your daily ritual!

Table 3: Sample tweets for telecommunication account Skype

the exact same way as Skype does and no beverage related keyword could be found in its tweets.

In summary our characterization of labels reveals the clarity and complexity of topics of tweets as they are perceived by humans. We also examined alignment of tweet topic with category of each account. The insight of this section helps our automated topic inference in the next section.

5 Text-based Topic Inference of Tweets

We now turn our attention into the automated classification of tweets from the target account into one of the specified topics.

Dataset: To expand our dataset for this analysis, we use the larger set of single label tweets that are presented in Table 1. Figure 6 shows the division of tweets for each account across four groups based on their labels:

- Aligned: tweets whose category and label agree.
- No topic: tweets that are labeled as "no topic".
- Other: tweets that are labeled as "other".
- Other labels: tweets that are labeled as one of the other categories.

Accounts of each category are grouped together. Categories are ordered from left to right based on their average percentage of aligned tweets and within each category accounts are ordered based on the same criteria. Therefore, Figure 6 is comparable to Figure 1. We observe that the order of categories and accounts in each category in Figure 1 and Figure 6 are exactly the same. Comparing these two figures reveals that three- and single-label tweets for each account exhibit generally similar characteristics.

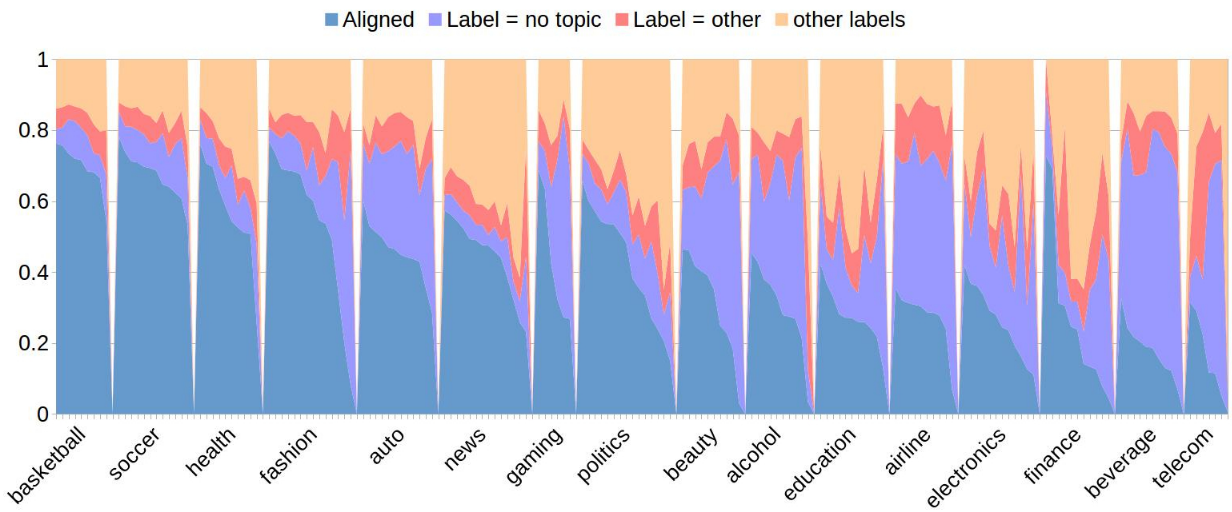


Figure 6: labeling information for single label tweets per account

	case 1		case 2		case 3	
category	NB	SVM	NB	SVM	NB	SVM
soccer	**0.97**	0.95	0.75	**0.87**	**0.93**	0.92
airline	0.64	**0.87**	0.16	**0.71**	0.65	**0.68**
basketball	0.8	**0.84**	0.68	**0.77**	**0.7**	0.69
health	0.76	**0.83**	0.37	**0.68**	0.47	**0.60**
news	0.67	**0.76**	**0.88**	0.6	**0.75**	0.7
politics	**0.78**	0.77	0.28	**0.53**	**0.54**	0.53
fashion	**0.80**	0.7	0.47	**0.54**	**0.58**	0.46
beauty	0.21	**0.61**	0.04	**0.43**	0.42	**0.47**
gaming	0.13	**0.58**	0.05	**0.47**	**0.40**	0.38
auto	0.52	**0.58**	0.07	**0.47**	**0.47**	0.39
alcohol	0.26	**0.57**	0.07	**0.40**	0.41	**0.42**
education	0.1	**0.55**	0.01	**0.30**	**0.36**	0.34
electronics	0.1	**0.39**	0.02	**0.29**	**0.38**	0.28
finance	0.01	**0.31**	0.01	**0.24**	0.15	**0.16**
telecom	0	**0.23**	0	**0.21**	0.14	**0.16**
beverage	0.01	**0.17**	0	**0.19**	**0.34**	0.32

Table 4: Accuracy result for all classifiers and two datasets

5.1 Methodology

We only focus on English tweets and we use the *bag of words* approach to process these tweets. After filtering stop words, we consider all words of a tweet as features when feeding them to a classifier. Each word and similarly each tweet is assigned a unique ID. For each tweet, we count the number of occurrences of each word so we would have a $W \times D$ matrix where W is the number of distinct words and D is the number of documents (here each tweet is a document). For analyzing single label tweets whose label and category agree, the number of distinct vocabularies is 88,373 and the number of documents (tweets) is 36,559. Therefor, the size of the matrix is very large; however it is also very sparse (i.e. most values in matrix are zeros) and only non-zero values are stored. The only filtering that is implemented here is removing stop words.

Next, we use *tf-idf* – stands for *term frequency inverse document frequency* – weighting scheme [19] to produce a weight for each word. This weight is highest when the word w occurs many times within a small number of documents and vise versa. The *tf-idf* matrix then is fed to two well known classifiers in the area of text mining for building the model; *(i)* Support Vector Machine (SVM) and *(ii)* Naive Bayes (NB). Other classifiers such as Linear Regression, Ridge Classifier, and Nearest Centroid are also implemented, but since their results are not better than SVM we just report their accuracy here and do not go into their details. In the next subsection we cover briefly why we focus on these classifiers.

All classifiers are implemented in Python using SciKit library [16]. We run the classifier on three different cases as follows:

Case 1: considering single label tweets whose label and category agree.

Case 2: considering all single label tweets leveraging only labels and ignoring categories.

Case 3: considering all tweets.

Note that the quality and reliability of specified topics for tweets decreases from Case 1 to Case 3. This allows us to study the effect of training set on classification accuracy which will be discussed in Section 5.

In all these cases, we employ *leave-one-out* cross validation in which we use tweets of 172 accounts for training and the tweets of the remaining one account for testing. Therefor, we repeat this process 173 rounds for each case.

The main motivation for leave-one-out testing (instead of using random tweets) is to assess whether training a classifier by $n-1$ accounts per category leads to a good classification of tweets on the single test account. This shows whether the selection of testing accounts have impact on the classification accuracy.

5.2 Classifiers

Classification and regression are supervised learning techniques to create models for prediction. Regression is when we predict quantitative outputs, and classification is when we predict qualitative outputs [9]. By using a threshold, regression turns into classification, so in this text we use the terms classification and regression interchangeably.

Classifiers are grouped into two categories: Generative and Discriminative. A generative model is a full probabilistic model of all variables, whereas a discriminative model provides a model only for the target variable(s) conditional on the observed variables.

Generative Classifiers: The way generative classifiers work is to model how the data is generated. Then based on generation assumptions, find the class which is most likely to generate the test data. These classifiers explicitly model the actual distribution of each class. One popular classifier in this category is Naive Bayes. This classifier applies Bayes Theorem to distinct between different classes. For the text data, usually word count is considered as a feature, and it is called *naive* because it assumes that the value of a particular feature is unrelated to the presence or absence of any other features.

Discriminative Classifiers: Discriminative algorithms allow to classify points without providing a model of how the points are actually generated. In short, discriminative classifiers try to model the decision boundary between the classes. Support Vector Machine is a typical discriminative classifier. It constructs a set of hyperplanes in space and tries to find a separator between samples, That are called support vectors. SVM does not try to understand the basic information of the individual classes as Naive Bayes does. Ridge Classifier, Nearest Centroid, and Linear Regression are other popular discriminative classifiers that have shown an acceptable performance in text data, which is why we implement them here in this project.

A. Jordan in [12], which is a widely cited study on the subject of discriminative vs. generative classifiers, compares Naive Bayes with Linear Regression. This study shows that discriminative models generally outperform generative models in classification tasks in terms of accuracy but fall behind from generative classifiers in terms of convergence rate.

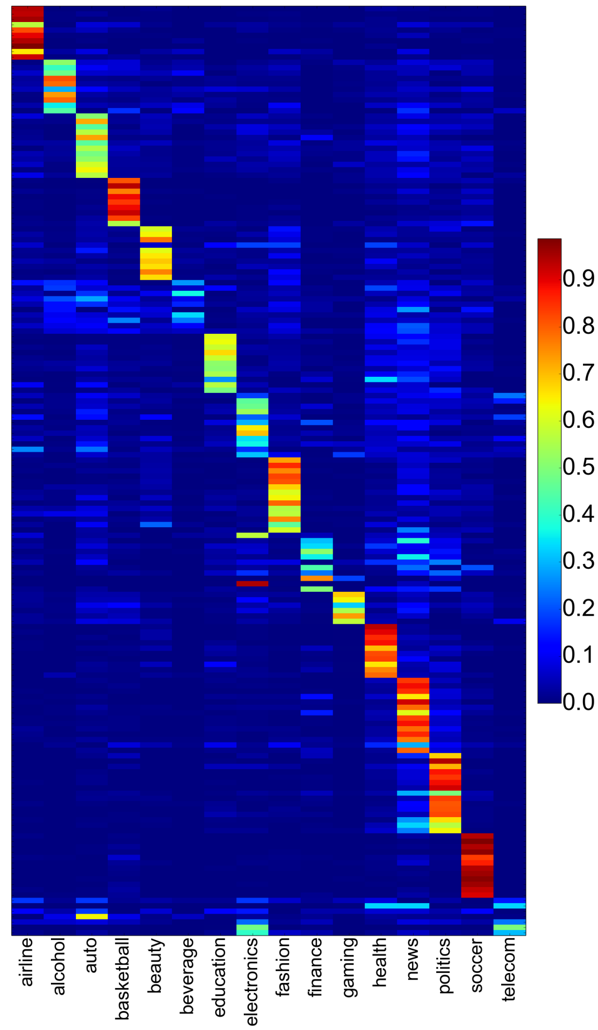


Figure 7: Account based accuracy heat map for support vector machine case 1

5.3 Per Category Analysis

We first examine the accuracy of classifiers at the per category level. Using leave-one-out cross validation, we measure the accuracy of each classifier as its average value across all accounts in that category.

Table 4 presents the per category accuracy for Naive Bayes and Support Vector Machine for all three cases. This table reveals that In all cases, certain categories show higher accuracy. There are categories with higher

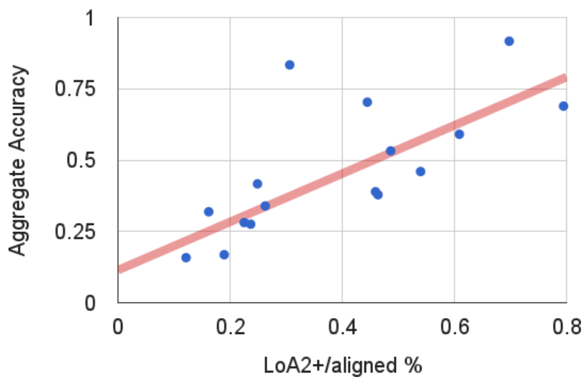


Figure 8: Scatter plot of aggregate accuracy versus LoA2+/aligned for all categories

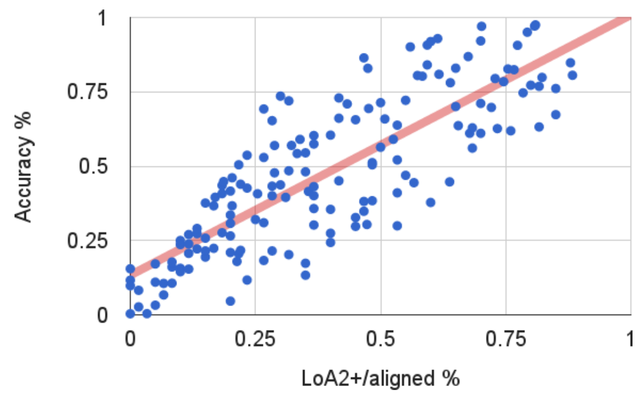


Figure 10: Scatter plot of aggregate accuracy versus LoA2+/aligned for all categories

number of LoA2+/aligned tweets such as basketball and soccer. Furthermore, accuracy for Case 1 is higher than Case 2 and Case 2 is higher than Case 3 which means better training, results in more reliable classification. Another general trend in this table is that SVM outperforms NB in Case 1 and Case 2 but in Case 3 NB surpasses SVM which can be explained by the size of dataset. Since Naive Bayes is a generative classifier it is trained better with larger dataset.

The most interesting point that we learn is that there is a relationship between accuracy and LoA2+/aligned metric that we defined in Section 4. This relationship is depicted in Figure 8. This figure is a scatter plot of aggregate accuracy versus LoA2+/aligned for all categories. As this figure reveals higher number of LoA2+/aligned is equivalent to higher accuracy and vice versa which is consistent with our hypothesis. We selected LoA2+/aligned because it is the most informative feature according to our decision tree.

5.4 Per Account Analysis

In this section, we focus on the accuracy of classifiers in each scenario for individual accounts. Toward this end, we plot the accuracy of SVM classifier in a heat map where X axis presents the accounts list (accounts are grouped based on their category) and Y axis shows the category. Each cell (i, j) shows how often account j's tweets are classified as i. The bluer the cell the less accuracy and vise versa. Figure 7 shows account based accuracy heat map for SVM running on Case 1 dataset. Generally we expect each account is classified as its expected category and the diagonal red band reveals this fact, although there exist misclassification that we explain shortly.

Using the heat map, we can also visualize overlap that we discussed in Section 4.1. Overlap between news/politics and news/finance is clearly visible that confirms our decision tree classification result that is based on LoAi/x features. We also understand from lighter vertical band above news category (13th column) that news has overlap with almost all categories.

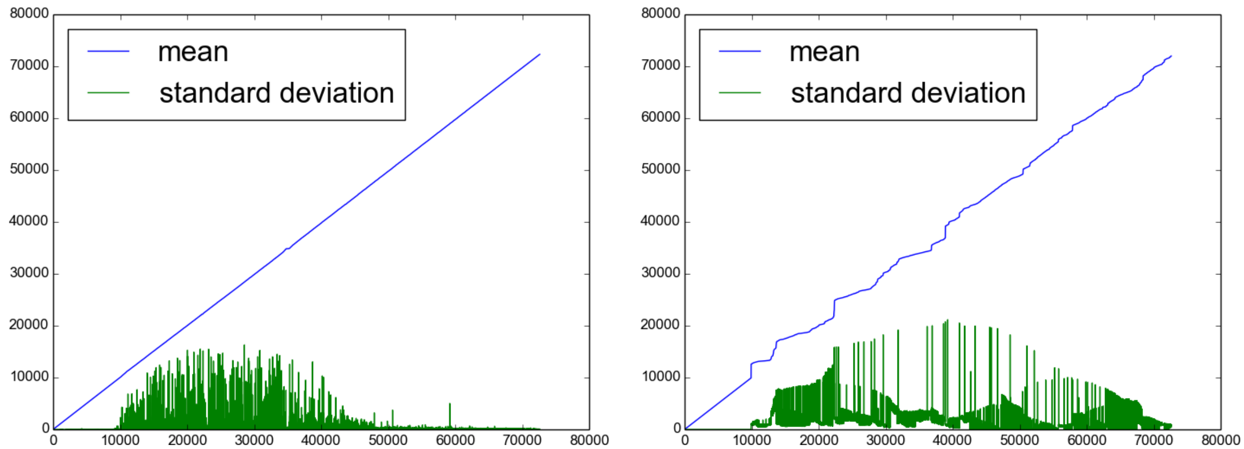
Another interesting point here is that telecom and beverage are not classified precisely, and if we zoom in we observe that some of the low accuracy accounts are those that were aimless which approves our hypothesis in labeling section. A good example here is account VerizonWireless, which is expected to be a telecom account while it is classified as both telecom and electronics. This is consistent with our previous findings in feature classification where electronics and telecom were classified in the same leaves although very inaccurately and also in overlap graph in presented in Figure 3.

Figure 10 plots the scatter plot between accuracy and LoA2+/aligned for all accounts which is even more revealing than Figure 8 in visualizing the relationship between accuracy and LoA2+/aligned.

Now that we can assign a topic to each Twitter account, we examine which keywords play the main role in inferring that topic and figure out if they are distinctive enough to separate one category from another. This analysis is done in the next section. For the next section we just consider Case 1.

6 Extracting Keywords

The purpose of this section is to determine the main key words that classifiers identify as distinguishing category among these collection of categories. For this analysis in addition to removing stop words we also remove



(a) Support Vector Machine

(b) Naive Bayes

Figure 9: Average and standard deviation for all 70k values across all rounds

URLs so that we do not see http or https as an important keyword. After filtering we have roughly 70k keywords that may have different weights/ranks in different rounds. Therefor, first we examine the stability of keyword ranks among 70K individual keywords. In other words we are seeking to answer the following question: How consistent is the rank/weight of keywords in different rounds? For this purpose we sort all keywords in all 173 rounds and keep their ranks so each keyword has 173 ranks. Then we remove the top 35 and bottom 35 (to remove outliers). Then we compute the average and standard deviation of remaining 100 values (ranks) and plot those values for all 70k keywords.

Figure 9 illustrates this stability. It shows both average and standard deviation and apparently for the first 10k keywords the standard deviation is negligible and the average value is pretty stable, and overall SVM is much more stable than NB, which can be explained by the nature of these two classifiers because NB is a generative classifier and can not capture dependency as opposed to discriminative classifiers (*e.g.*, SVM) that learn the boundary between classes instead of learning each class and determining as to which class each tweet belongs to. Consequently in each round Naive Bayes learns the whole data, so it produces more variable weights and consequently more variable ranks.

As a result of above the exercise we can show the keywords in a word cloud so we could visualize the words that a classifier considers important. Thus in each round of leave-one-out cross validation we sort all keywords based on their weight in a list (note that weight range is different for different classifiers since they use different algorithms to calculate weight vector hence we work with ranks instead of weight) and pick the first 200 key-

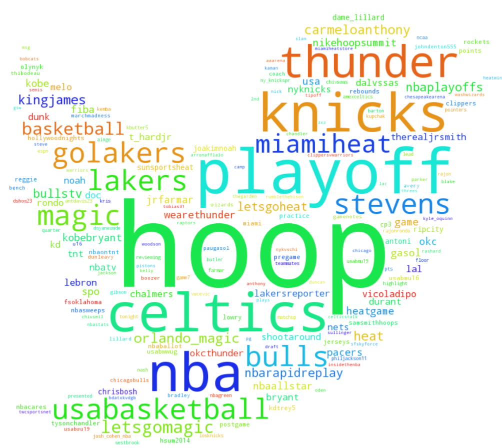


Figure 11: Top 200 keywords for category basketball – the classifier is SVM

words for each category. Then we plot a word cloud per category per classifier to visualize the keywords. A sample of these word clouds is illustrated in Figure 11 (you may find the rest of them online). In this figure, size is related to weight (but not color and centrality).

7 Topic Inference Through Topic Modeling

The number of topics (T) is an input for the topic modeling algorithm, and the result of this algorithm is highly dependent on this variable. In our experiment we set $T = 16$. Accordingly after running this algorithm it returns a list of 16 topics (*i.e.*, t_0 to t_{15}) and a list of keywords associated to each topic and a mapping between

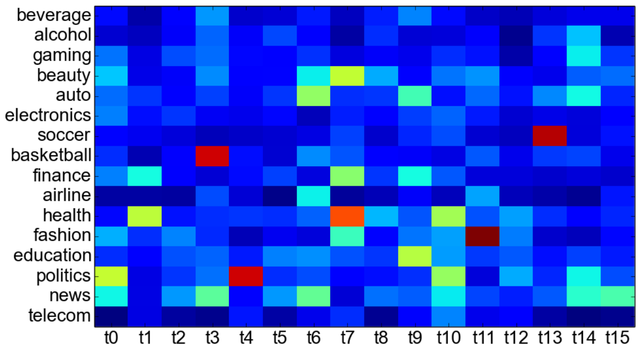


Figure 12: heat map between topic modeling result and account category

documents (*i.e.*, tweet) and topics. To present the result of our experiment we do the following exercise: Tweets of each category can be mapped to several t_i. We count how often each category is mapped to each t_i and plot the result in a heat map. Figure 12 illustrates this heat map.

As it is seen, there are certain topics that are modeled successfully, but not all of them. Despite its incompleteness, this heat map is consistent with Figure 7 in which basketball, soccer, fashion, health, and politics had relatively high accuracies.

8 Discussion

So far we have analyzed tweets of major accounts using two methods; first we characterized tweets and extracted features (*i.e.*, LoAi/x) and performed classification using those features. Then we feed tweets to support vector machine to obtain the accuracy. As a result of these two analysis we can think of an approach to build a valuable training set for certain applications. The approach is as follows:

- To find topic of tweets we need a labeled dataset to train the classifier.
- We measure LoAi/x features for a particular account and compare them with our result.
- If according to our division it is a purposeful account then all tweets of that account could be used for training.

9 Conclusion

We conducted this study in two parts, in part one we characterized tweets based on their labels and introduced a metric called LoAi/x and following is the summary of our findings:

- A majority of tweets of certain categories have an aligned topic.
- Misaligned tweets appear to be caused by multi-topic tweets that suggests pairwise relevance of topics.
- Fraction of tweets with various level of alignment offer valuable features to identify a category.
- These features also seem to reveal the way that entities in each category use Twitter.

In second part we performed text based classification and we found interesting connection between results of part one and part two:

- Certain categories/accounts exhibit higher accuracy in all cases. (*e.g.*, soccer, basketball) these categories/accounts have a relatively higher fraction of aligned tweets (LoA2+/aligned).
- Accuracy of classification depends on the quality and the size of training dataset. More reliable training set results in higher accuracy.
- SVM outperforms NB except when we have larger data set with lower quality/reliability.

References

[1] Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 114–122, 2011.

[2] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1266–1275, 2014.

[3] David M Blei and Jon D McAuliffe. Supervised topic models. In *NIPS*, volume 7, pages 121–128, 2007.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[5] Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):22, 2011.

[6] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 27–38, 2013.

[7] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 515–526, 2013.

[8] JA Hartigan and MA Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, pages 100–108, 1979.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. 2001.

[10] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[11] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg, 1998.

[12] A Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. 2002.

[13] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 170–178, San Francisco, CA, USA, 1997.

[14] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg!, 2011.

[15] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, pages 265–272, 2011.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[17] Matthew Purver, Thomas L Griffiths, Konrad P Körding, and Joshua B Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 17–24, 2006.

[18] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[19] Karen Sparck Jones. Document retrieval systems. chapter A Statistical Interpretation of Term Specificity and Its Application in Retrieval, pages 132–142. London, UK, UK, 1988.

[20] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[21] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 448–456, 2011.

[22] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 937–946, 2009.

[23] Zhe Zhao and Qiaozhu Mei. Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1545–1556, 2013.

Appendix 1

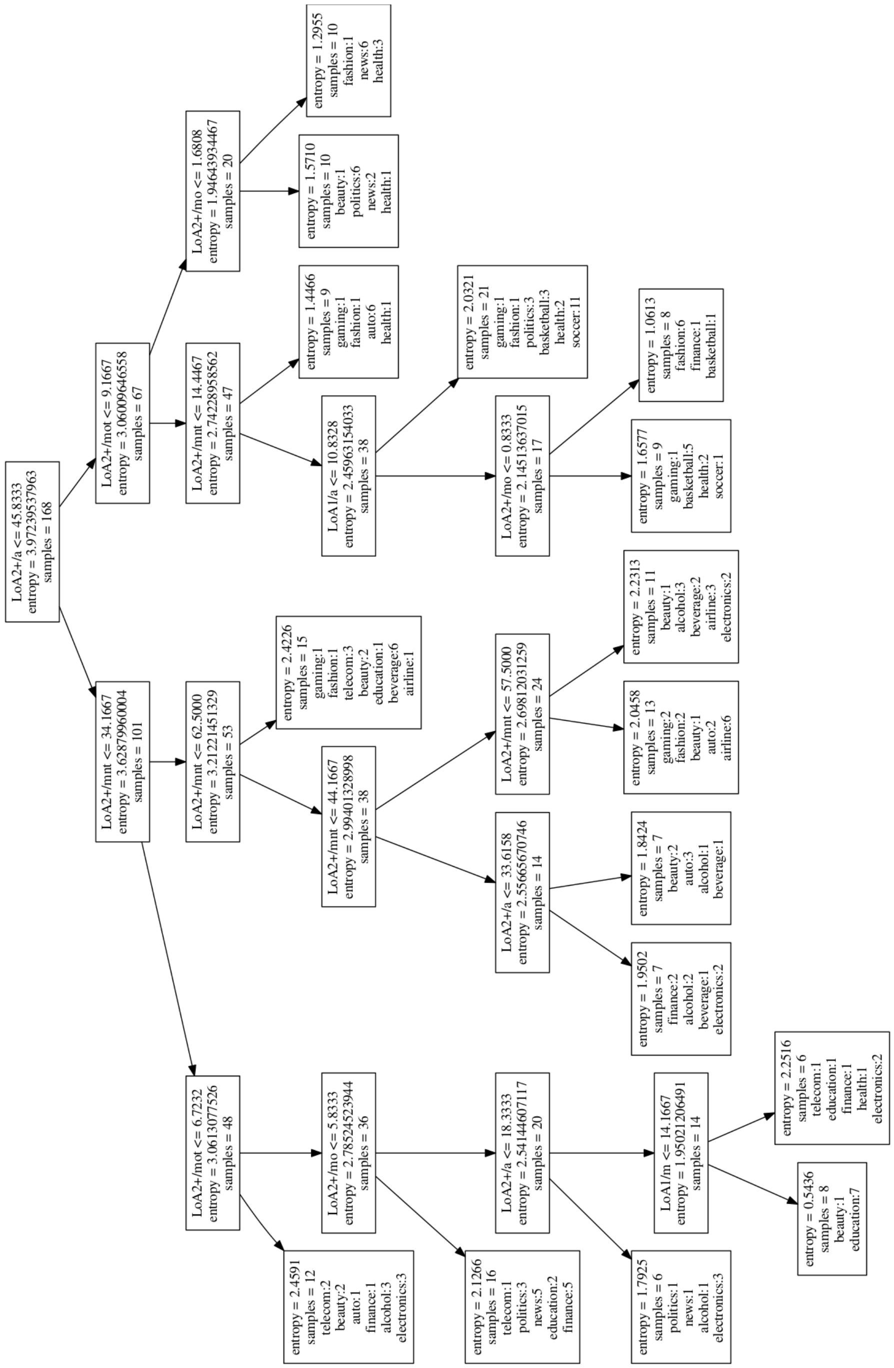


Figure 13: The decision tree based on LoAi/x features

Appendix 2

Topic	Accounts associated with the topic	number of tweets per account
finance #tw: 31,776	Bloomberg	3,233
	BofA_Community	3,198
	Citi	3,215
	NASDAQ	3,243
	Visa	3,215
	Sequoia_Capital	2,772
health #acc: 10 #tw: 27,726	WebMD	3,205
	MayoClinic	3,233
	EverydayHealth	3,229
	ClevelandClinic	3,238
	HopkinsMedicine	3,205
	DoveMed	1,532
	pfizer	1,720
	JNJNews	3,231
	MedicalNews	3,238
	NIHClinicalCntr	1,895
soccer #acc: 12 #tw: 38,522	Arsenal	3,200
	FIFAcom	3,238
	UEFAcom	3,202
	premierleague	3,201
	chelseafc	3,204
	FCBarcelona	3,203
	EuropaLeague	3,222
	ChampionsLeague	3,198
	LFC	3,208
	ManUtd	3,223
	MCFC	3,212
	SpursOfficial	3,211
telecom #acc: 7 #tw: 22,583	Skype	3,252
	VerizonWireless	3,205
	ATT	3,239
	cspan	3,235
	TMobile	3,207
	sprint	3,209
	VZWnews	3,236

Topic	Accounts associated with the topic	number of tweets per account
politics #acc: 15 #tw: 36,923	BarackObama	3,210
	algore	1,304
	SenJohnMcCain	3,235
	billclinton	180
	newtgingrich	3,213
	MittRomney	1,400
	GOP	3,231
	FreedomWorks	3,239
	dccc	3,223
	HouseDemocrats	3,219
	LibDems	3,215
	StateDept	3,209
	OpenGov	623
	TheJusticeDept	1,215
	ObamaNews	3,207
gaming #acc: 6 #tw: 19,383	PlayStation	3,220
	Xbox	3,232
	NintendoAmerica	3,237
	ASTROGaming	3,237
	elgatogaming	3,222
	ScufGaming	3,235
news #acc: 14 #tw: 45,044	cnnbrk	3,204
	BBCBreaking	3,223
	BreakingNews	3,232
	Reuters	3,203
	AP	3,218
	ABC	3,213
	CBSNews	3,241
	nprnews	3,205
	NBCNews	3,203
	BloombergNews	3,242
	CNN	3,198
	PBS	3,212
	CNBC	3,218
	FoxNews	3,232

Table 5: List of all topics with their associated accounts and the number of tweets per topic and per account

Topic	Accounts associated with the topic	number of tweets per account
airline #acc: 10 #tw: 32,229	JetBlue	3,248
	SouthwestAir	3,231
	AmericanAir	3,208
	Delta	3,210
	VirginAmerica	3,244
	USAirways	3,202
	united	3,240
	British_Airways	3,206
	AirCanada	3,214
	VirginAtlantic	3,226
alcohol #acc: 10 #tw: 28,339	TopBrassVodka	3,233
	newbelgium	3,230
	dogfishbeer	3,236
	SierraNevada	3,227
	DeschutesBeer	3,237
	budlight	1,394
	MillerLite	2,156
	Budweiser	2,234
	CoorsLight	3,206
	Skinnygirl	3,186
auto #acc: 12 #tw: 38,589	Audi	3,220
	Lexus	3,228
	Ford	3,216
	chevrolet	3,245
	NissanUSA	3,233
	MBUSA	3,193
	Jeep	3,204
	Toyota	3,226
	JaguarUSA	3,177
	Dodge	3,199
	VW	3,207
	GM	3,241
basketball #acc: 9 #tw: 28,850	NBA	3,200
	usabasketball	3,176
	Lakers	3,206
	chicagobulls	3,205
	MiamiHEAT	3,227
	celtics	3,201
	Orlando_Magic	3,195
	nyknicks	3,242
	okcthunder	3,198
beauty #acc: 10 #tw: 32,211	COVERGIRL	3,214
	Clinique_US	3,246
	revlon	3,203
	LancomeUSA	3,197
	Dove	3,234
	LushLtd	3,236
	tartecosmetics	3,213
	DegreeWomen	3,210
	AvonInsider	3,232
	OlayUS	3,226

Topic	Accounts associated with the topic	number of tweets per account
beverage #acc: 10 #tw: 32,969	pepsi	3,202
	CocaCola	3,234
	redbull	3,221
	mtn_dew	3,237
	drpepper	3,225
	Sprite	3,212
	vitaminwater	3,976
	Tropicana	3,231
	Snapple	3,203
	Lipton	3,228
education #acc: 11 #tw: 33,773	Harvard	3,201
	UOPX	3,210
	Stanford	3,203
	UniofOxford	1,611
	Yale	3228
	Cambridge_Uni	3,221
	TAMU	3,224
	Princeton	3,195
	OhioState	3,229
	UTAustin	3,223
	umich	3,228
electronics #acc: 12 #tw: 37,522	SamsungMobileUS	3,210
	BlackBerry	3,209
	intel	3,203
	Sony	3,204
	nokia	3,203
	htc	3,201
	HP	3,244
	Cisco	3,204
	nvidia	2,926
	Dell	3,206
	lenovo	3,227
	IBM	2,485
fashion #acc: 14 #tw: 34,837	Dior	1,005
	CHANEL	810
	dolcegabbana	3,225
	VictoriasSecret	3,234
	hm	3,198
	Burberry	3,247
	YSL	178
	CalvinKlein	2,746
	armani	3,201
	Versace	3,012
	gucci	2,500
	RalphLauren	1,998
	TommyHilfiger	3,235
	VANS_66	3,248
finance #acc: 10	kickstarter	3,240
	WorldBank	3,203
	AmericanExpress	3,216
	CNNMoney	3,219

Table 6: List of all topics with their associated accounts and the number of tweets per topic and per account